IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 30, 2020, accepted July 20, 2020, date of publication August 6, 2020, date of current version October 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3014175

Semantic Segmentation of Smartphone Wound Images: Comparative Analysis of AHRF and CNN-Based Approaches

AMEYA WAGH[1], SHUBHAM JAIN[2], APRATIM MUKHERJEE[3], EMMANUEL AGU[4], (Member, IEEE), PEDER C. PEDERSEN[4], (Senior Member, IEEE), DIANE STRONG[4], BENGISU TULU[4], (Member, IEEE), CLIFFORD LINDSAY[5], AND ZIYANG LIU[4]

[1]TORC Robotics, Blacksburg, VA 24060, USA
[2]Nvidia Corporation, Santa Clara, CA 95051, USA
[2]Computer Science Department, Manipal Institute of Technology, Manipal 576104, India
[4]Computer Science Department, Worcester Polytechnic Institute, Worcester, MA 01609, USA
[5]Radiology Department, University of Massachusetts Medical School, Worcester, MA 01655, USA

Corresponding author: Emmanuel Agu (emmanuel@wpi.edu)

This work was supported in part by the National Institutes for Health/National Institute of Biomedical Imaging and Bioengineering (NIH/NIBIB) under Grant 1R01EB025801-01.

ABSTRACT Smartphone wound image analysis has recently emerged as a viable way to assess healing progress and provide actionable feedback to patients and caregivers between hospital appointments. Segmentation is a key image analysis step, after which attributes of the wound segment (e.g. wound area and tissue composition) can be analyzed. The Associated Hierarchical Random Field (AHRF) formulates the image segmentation problem as a graph optimization problem. Handcrafted features are extracted, which are then classified using machine learning classifiers. More recently deep learning approaches have emerged and demonstrated superior performance for a wide range of image analysis tasks. FCN, U-Net and DeepLabV3 are Convolutional Neural Networks used for semantic segmentation. While in separate experiments each of these methods have shown promising results, no prior work has comprehensively and systematically compared the approaches on the same large wound image dataset, or more generally compared deep learning vs non-deep learning wound image segmentation approaches. In this paper, we compare the segmentation performance of AHRF and CNN approaches (FCN, U-Net, DeepLabV3) using various metrics including segmentation accuracy (dice score), inference time, amount of training data required and performance on diverse wound sizes and tissue types. Improvements possible using various image pre- and post-processing techniques are also explored. As access to adequate medical images/data is a common constraint, we explore the sensitivity of the approaches to the size of the wound dataset. We found that for small datasets (<300 images), AHRF is more accurate than U-Net but not as accurate as FCN and DeepLabV3. AHRF is also over 1000x slower. For larger datasets (>300 images), AHRF saturates quickly, and all CNN approaches (FCN, U-Net and DeepLabV3) are significantly more accurate than AHRF.

INDEX TERMS Wound image analysis, semantic segmentation, chronic wounds, U-Net, FCN, DeepLabV3, associative hierarchical random fields, convolutional neural network, contrast limited adaptive histogram equalization.

I. INTRODUCTION

Diabetes Mellitus is a serious medical condition that affected 30.3 million people in 2017 [1]. About 15% of diabetes patients have chronic wounds in the US, which has a treatment cost of about $25 billion annually [2]. The majority of diabetic wounds are located in the lower extremities, may take years to heal, can re-occur and can adversely affect the physical and mental health of the patient if not treated by experts regularly.

Chronic wound care requires regular checkups by wound nurses who debride the wound, inspect its healing progress

The associate editor coordinating the review of this manuscript and approving it for publication was Ruqiang Yan.

and recommend visits to wound experts when necessary. Accurate and timely care decisions are crucial for proper wound healing and delays in visiting a wound specialist could result in limb amputation. To reduce delays in care decisions, wound nurses often send remote wound images to experts for decisions on the best treatment options. Since 2011, our group has been researching and developing the Smartphone Wound Analysis and Decision-Support (SmartWAnDS) system, which can intelligently recommend wound care decisions by analyzing images of a patient's wound and information in their Electronic Health Records (EHR), providing a second opinion for nurses working in remote locations. We envision that SmartWAnDS will standardize the quality of wound care even when the care is provided by nurses without wound expertise and reduce the workload of wound experts. We envision that SmartWAnDS could recommend when patients need visits to wound experts, provide healing scores or suggest minor changes in treatment. The SmartWAnDS system will be available as a smartphone app that can analyze wound images captured using the phone's camera, and the patient's EHR.

The visual characteristics of a wound that are useful in evaluating its health include its size, infection level, granulation tissue amount, necrotic tissue amount, slough and wound depth [3]–[5]. However, prior clinical studies have found a wound size to be the most important measure of its health [6]. For instance, the change in the size of a chronic wound in a 4-week period is an accurate predictor of whether the wound will heal or not [6]. Consequently, the segmentation step is an important step in most wound image analysis pipelines. The goal of our wound segmentation task is to label each pixel of a wound image into one of three semantic categories - wound, skin and background (also called semantic segmentation). Image segmentation has traditionally been performed using methods such as the Conditional Random Fields (CRF) and its variants such as the Associative Hierarchical Random Fields (AHRF). However, following the unprecedented success of Convolutional Neural Networks (CNNs) for image classification in 2012 (AlexNet) [7], CNNs have been found to outperform traditional methods for several computer vision tasks such as image classification [7], segmentation [8] and object detection [9].

Fully Connected Networks (FCN) [10], U-Net [11] and DeepLabV3 [8] are deep learning-based segmentation networks that have outperformed traditional image segmentation methods when given enough data. Wound image analysis has also recently started using deep learning for wound image classification and segmentation as seen in DeepWound [12] and DFUNet [13]. However, to the best of our knowledge, no systematic comparison between a deep learning approach and traditional (non-deep learning-based, graphical or CRF-based) techniques for wound image segmentation has been performed.

In this paper, we present a systematic and comprehensive comparison between Associative Hierarchical Random Fields (AHRF) and three deep learning based models (Fully Convolutional Networks (FCN), U-Net and DeepLabV3) for the task of wound image segmentation. We compare these approaches using a diverse set of performance metrics including segmentation accuracy (dice coefficient), sensitivity to the amount of training data utilized and model inference time. As real-world images and data of actual patients are often difficult to obtain in many medical applications, it is important to compare the performance of these methods with respect to the size of the training datasets. Deep learning methods are well known to be data intensive. We found that when the number of training images is small (<300), AHRF (traditional) has a higher accuracy (dice coefficient) than U-Net but is still not as accurate as FCN and DeepLabV3 which were pre-trained on a subset of the COCO [14] dataset. As the number of training images increases, AHRF begins to saturate and the accuracy gap between AHRF and U-Net shrinks with U-Net eventually becoming more accurate than AHRF. FCN and DeepLabV3 consistently outperformed both U-Net and AHRF for all training set sizes. As we envision that our SmartWAnDS wound assessment system will eventually be deployed on a smartphone, we also examined the computational requirements of each method, inference time, and the need to communicate with a remote server.

The rest of this paper is organized as follows. Section II provides a brief background on the techniques used in this paper followed by the related work in image segmentation in Section III. The methodology used in this paper and a description of the wound image dataset utilized for training is located in Section IV. Sections V and VI present our results and a discussion of our major experiments and analyses of our findings. Finally, in Section VII, we conclude and suggest some directions for future work.

II. BACKGROUND

We compared *semantic segmentation of wound images* using Associative Hierarchical Random Fields (AHRFs) and Convolutional Neural Networks (CNNs) for assigning a label of skin, wound or background to each pixel of an input image. Some background on both approaches are now presented.

A. ASSOCIATIVE HIERARCHICAL RANDOM FIELDS (AHRFs)

Conditional Random Fields (CRFs) model data probalistically and have been found to be effective for various machine learning prediction tasks. AHRFs [15], a variant of CRFs leverage contextual data by considering other pixels in the neighbourhood of the target pixel to be classified, which works better than considering each pixel's label in isolation. AHRFs model the conditional probability that a given pixel should be assigned a certain label, by considering the pixel itself as well as other pixels in its neighbourhood. An energy function consisting of unary, pairwise and higher order potentials is minimized to find the most optimal semantic labels for a given image. The unary potential takes features extracted from the target pixel as input and outputs a probability score for each target class. Pairwise potential ensures that

nearby pixels that have similar features are assigned the same label. Higher order potentials are constructed such that pixels belonging to the same superpixels or cliques have the same label. Graph solving techniques are then used to minimise the energy and determine optimal labeling. Details about AHRF including the energy function minimized are presented in the Methodology section as Equation 1.

B. CONVOLUTIONAL NEURAL NETWORKS (CNNs)

CNNs have been found quite effective for many computer vision tasks in recent years. They act as trainable image filters which can be used to convolve over images sequentially to measure responses or activations of the input image, creating feature maps. These feature maps are then stacked together, passed through non-linear functions, and further convolved with more filters. This convolution process has been found to be effective at extracting visual features or patterns in images that can be useful for tasks such as classification, segmentation, and super resolution. In this paper, we compare three CNN-based architectures for semantic segmentation: FCNs, DeeplabV3 and U-Net, which we now review briefly.

1) FULLY CONVOLUTIONAL NETWORK (FCN)

As they have generally performed well for per-pixel tasks, Long *et al* first proposed using FCNs trained end-to-end for semantic segmentation. FCN utilizes a skip architecture that combines semantic information from a deep, coarse layer with appearance information from a shallow, fine layer to produce accurate and detailed segmentations. FCNs have only locally connected layers, such as convolutions, pooling and upsampling, avoiding any densely connected layer. It also uses skip connections from it's pooling layers to fully recover fine-grained spatial information which is lost during downsampling.

2) U-NET

U-Net [11] is an encoder-decoder architecture that uses CNNs. Encoder-decoder networks, as the name suggests have two parts - an encoder and a decoder. The encoder is responsible for projecting the input feature vectors into a low dimensional space in which similar features lie close together. The decoder network takes features from this low dimensional space as input and attempts to recreate the original input features. Thus, the output of the encoder or conversely input of the decoder is called the bottleneck region where a low dimensional representation is present. Encoder-decoder networks have been found to be effective for various tasks such as image denoising, language translation and image segmentation.

3) DeepLabV3

DeepLabV3 [8] utilizes atrous convolutions along with spatial pyramid pooling which enlarges the field of view of filters to incorporate larger context and controls the resolution of features extracted. Employing atrous convolutions in either cascade or in parallel captures multi-scale context due to the use of multiple atrous rates. DeepLabV3 uses a backbone network such as a ResNet [16] as its main feature extractor except that the last block is modified to use atrous convolutions with different dilation rates.

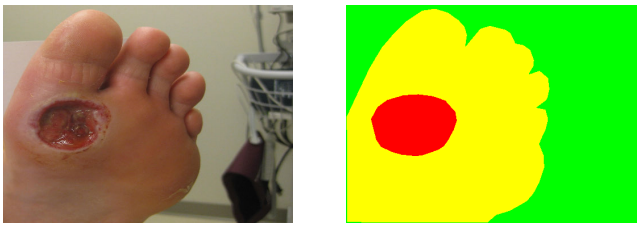

FIGURE 1. Wound image (left), pixel-wise segmentation mask for wound, skin and background (right).

III. RELATED WORK

A. PROBABILISTIC TECHNIQUES FOR WOUND IMAGE ANALYSIS

Prior to the rise in the popularity of deep learning, wound analysis mostly utilized probabilistic techniques such as color space manipulation [17], [18], machine learning classifiers using hand-crafted features [19], clustering techniques [20] and edge detection [21]. These probabilistic approaches generally have the advantage of not being very data intensive as they use hand-crafted features and shallow machine learning models. However, they fail to generalize well to new images captured in varied lighting conditions, skin and wound types. For the purpose of comparison with deep learning, in this paper, we use Associative Hierarchical Random Fields (AHRF) [15] as a probabilistic solution for image segmentation. AHRF uses region growing for connecting pixels that have similar visual features and also uses a combination of handcrafted and learned features for semantic segmentation of an image.

B. CNN-BASED IMAGE SEGMENTATION TECHNIQUES

Researchers have applied CNNs to biomedical applications such as wound segmentation using transfer learning [22], using lightweight mobile deep learning architectures (MobileNet) for wound segmentation [23], region proposal-based Faster R-CNN model for wound localization [24], and the inception module based CNN for classification of skin into healthy and abnormal [13]. These methods all try to segment wound pixels but do not distinguish the skin region from the background in the image. Li *et al.* [25] proposed a method to segment out skin pixels using heuristics for thresholding and region growing as a first step, and then passed forward the cropped image with detected skin to the MobileNet CNN architecture for wound segmentation.

The downside to using neural networks is that they require large datasets to train from scratch which is not always available in applications that use medical or clinical data. This problem can be alleviated by using techniques such as data augmentation to increase variations in the existing data and transfer learning, which uses models that have been

TABLE 1. Statistics of dataset.

Dataset	Avg R	Avg G	Avg B	Std R	Std G	Std B
Dataset 1	0.535	0.533	0.529	0.144	0.142	0.141
Dataset 2	0.459	0.462	0.463	0.153	0.154	0.155
Dataset 3	0.472	0.472	0.473	0.172	0.172	0.173

Mean and Standard Deviation of normalized images in R,G,B channel

TABLE 2. Statistics of dataset.

Dataset	Avg R	Avg G	Avg B	Std R	Std G	Std B	Avg %	Std %
D1-Wound	0.475	0.273	0.232	0.080	0.099	0.089	7.69	7.64
D2-Wound	0.518	0.315	0.286	0.104	0.120	0.112	11.03	10.15
D3-Wound	0.515	0.310	0.260	0.106	0.118	0.109	6.63	7.95
D1-Skin	0.489	0.367	0.308	0.137	0.125	0.121	56.43	13.13
D2-Skin	0.565	0.414	0.392	0.143	0.135	0.133	52.55	15.22
D3-Skin	0.577	0.429	0.363	0.133	0.128	0.126	47.74	17.69

Mean and Standard Deviation of normalized images in R,G,B channel cropped with wound and skin masks

previously trained for similar vision tasks. The deep learning segmentation methods utilized in this paper were organized in two different ways. U-Net had separate classifiers for wound and skin while FCN and DeepLabV3 had just one classifier for both skin and wound. This enabled us compare whether the arrangement of classifiers affected the models performance.

IV. METHODOLOGY

A. DATASETS OF WOUND IMAGES

We gathered 3 different datasets as described below, which include diabetic foot ulcers, arterial, venous, pressure ulcers and surgical wounds. Many of the images exhibit typical wound attributes such as granulation, necrosis and slough. A wound annotation app (shown in Fig-2) was specifically created to expedite pixel-level annotations of wound and skin segments within the given images. The wound annotation app implemented the deep extreme cut algorithm [26], providing consistent wound annotation. Specifically, we did not rely on human labelers, which obviated the need for evaluating inter-rater reliability.

- *Dataset 1* consists of 114 wound images captured with controlled lighting conditions. A wound imaging box was created [27] that simulated a consistent, homogeneous lighting environment. The segmentation masks consist of pixel-level labels where the red color corresponds to the wound segment, yellow corresponds to the skin segment and background is indicated by a green-colored mask 1.
- *Dataset 2* was gathered by scraping publicly available wound images from the internet. It consists of 202 images collected by scraping and 114 images from dataset 1, which yields a total of 316 images. This dataset has images with varying lighting conditions but the wounds were mostly captured from a relatively perpendicular angle.
- *Dataset 3* is the largest dataset with 1442 images in total, which was acquired from the vascular surgery department of the University of Massachusetts Medical Center. This dataset has images with large variations in lighting, viewing angles, wound types and skin texture.

Table-1 shows the mean and standard deviation of the normalized values in the R,G,B channels. It can be observed that the standard deviation of the RGB values is less in dataset 1 as the images were captured using a wound box with controlled lighting and imaging distance, and increases for dataset 3. Table-2 shows the image statistics of only wound and only skin pixels, obtained by cropping the image with the ground truth mask. The standard deviations are quite high for both wound and skin showing significant variations in our datasets. Table-2 also shows the average percentage of wound and skin pixels within a wound image and their corresponding standard deviation. It can be seen that the average wound percentage is less than 10 % whereas skin covers almost 50 % creating class imbalance.

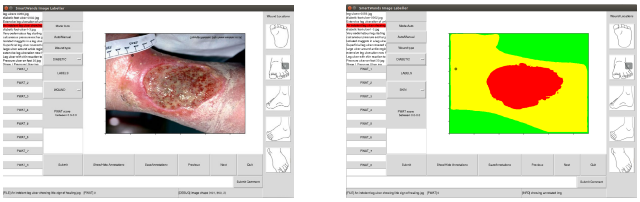

FIGURE 2. Annotation app with wound image view (left), pre-view of the mask after annotating the wound image (right).

B. WOUND IMAGE PRE-PROCESSING

In order to make our algorithms more robust to lighting variations and noisy imaging conditions, several pre-processing techniques were explored. Most of these techniques involved manipulating the images' histograms in some form. The histogram is the probability distribution of pixel intensity values within an image, ranging from 0 to 255. After experimenting with the impact of many techniques on semantic segmentation accuracy such as image sharpening, histogram normalization, contrast enhancement, vignetting, gamma correction, reflectance, histogram matching and Contrast Limited Adaptive Histogram Equalisation (CLAHE), we found that CLAHE was consistently the most effective pre-processing technique.

Contrast Limited Adaptive Histogram Equalization (CLAHE): CLAHE [28] is an image pre-processing technique based on adaptive histogram equalisation [29] which contextually equalizes the histogram of local image regions. Thus, the pixel's intensity is transformed proportional to its rank of intensity among its neighbours defined by a kernel size. This technique was found to significantly enhance both the signal and noise components of an imag, which was not desired. CLAHE ensures that noise enhancement is reduced by using a contrast limiting factor called clip limit. This user defined limit is used as a maximum allowable local contrast enhancement factor. A grid search over the kernel size and clip limit was performed to obtain a kernel size of (24, 24) and clip limit of 3.0 as the most optimal hyperparameters for

FIGURE 3. An example image (left) with the Contrast Limited Adaptive Histogram Equalization Image (right).

our dataset. An example of CLAHE pre-processing with our hyperparameters is shown in Fig-3.

C. ASSOCIATIVE HIERARCHICAL RANDOM FIELD (AHRF)

Image segmentation using AHRF, a variant of CRF, consists of two parts: 1) calculating the energy value for an image given its pixel-wise labels, which considers both local features and similar neighboring pixels, 2) a graph solving approach, which tries to determine the optimal assignment of labels to an image such that its energy function is minimized. The mathematical formulation of AHRF is explained below. A high-level workflow of AHRF is also shown in Fig-4

FORMULATION

Let us first define the following variables -

$X = \{X_1, X_2, \ldots X_n\}$ are the variables to be labelled

$L =$ set of labels from which X_i are labeled

$y_i =$ individual label given to X_i such that $y_i \in L$

$M =$ number of paired training instances of the form $\{x^{(i)}, y^{(i)}\}_{i=1}^{M}$

$V = \{1, 2, \ldots n\}$ set of valid vertices or indices of X

$N =$ defined by sets $N_i \forall i \in V$ where N_i denotes the set of all neighbors of X_i

$C =$ set of all cliques c where a clique X_c is a set of variables X that are similar and codependent such as super-pixels $y_c =$ labelling given to each clique c

Using the variables defined above, an AHRF formulation consists of an energy function E which is written as the sum of unary, pairwise and clique-wise potential as shown in equation 1 below.

$$E(y) = \underbrace{\sum_{i \in V} \phi_i(y_i, \theta_u)}_{\text{Unary Potential}} + \underbrace{\sum_{i \in V, j \in N_i} \phi_{ij}^{p}(y_i, y_j, \theta_p)}_{\text{Pairwise potential}} + \underbrace{\sum_{c \in C} \phi_c^h(y_c)}_{\text{Higher order potential}} \quad (1)$$

In the above formulation, θ_u and θ_p are a set of parameters that are learned from the training paired samples $\{x^{(i)}, y^{(i)}\}_{i=1}^{M}$ with the objective of maximizing the conditional distribution $P(y|X)$. The higher order potential is described in equation 2 below.

$$\phi_c^h(y_c) = \min_{i \in L}(\gamma_c^{\max}, \gamma_c^l + \sum_{i \in c} w_i k_c^l \Delta(y_i \neq l)) \quad (2)$$

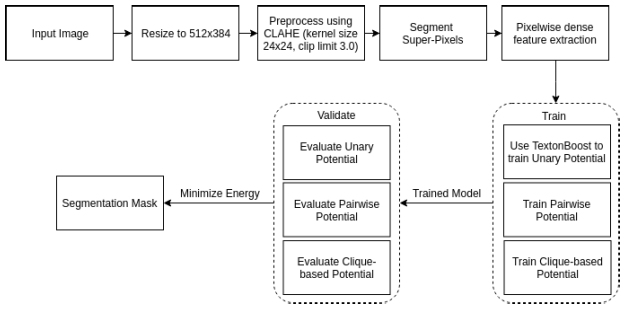

FIGURE 4. AHRF implementation and workflow.

where w_i is the weight of the variable x_i and each variable of a clique is penalized with a cost $w_i k_c^l$ if it has not taken the value of the dominant label of that clique. The value of penalty is truncated at $\gamma_c^{\max}$. This formulation also supports higher order super-pixel based potentials across multiple scales of the image since it allows for cliques to take a free label in the case of multiple dominant labels and also considers relationships between cliques to increase contextual awareness. We have used mean shift segmentation to generate super-pixels. Several different features have been used to calculate the AHRF potentials including textonBoost features on RGB and LAB colorspace, local binary patterns, Histogram of Oriented Gradients (HOG), SIFT features and color distribution features. Given the potential terms and parameters, the optimal labeling can be found by minimizing the overall energy using graph-cut based move making algorithms such as alpha expansion or alpha-beta-swap algorithm.

D. SEMANTIC SEGMENTATION ARCHITECTURES USING CNNs

1) FULLY CONVOLUTIONAL NETWORKS (FCNS)

FCNs differ from the classic CNNs used for image classification tasks. The CNN pipeline for image classification usually has a structure with several convolution layers followed by fully connected layers and outputs one predicted label per image. On the other hand, Long *et al* describe a Fully Connected Network (FCN) as one that uses only convolutions, pooling and activation functions and computes a nonlinear filter [10]. It achieved state-of-the-art segmentation on PASCAL VOC 2012 [30], NYUDv2 and SIFT Flow in 2015.

Classification networks can be converted into FCNs by eliminating the final classifier layers and appending a 1×1 convolution layer with a channel dimension equal to the number of classes to be predicted. This also allows the network to accept arbitrary sized images as input. This modification performs well on segmentation tasks but the output is coarse, which is remedied by adding skips that combine outputs from the lower layers with finer strides to generate the final prediction. This refines the output as local information from the lower layers makes the model pay attention to the global structure. Upsampling is required to fuse these outputs, which is done by deconvolution layers.

Network Structure: We utilized ResNet101 [16] as the backbone of this network. The model consists of four layers

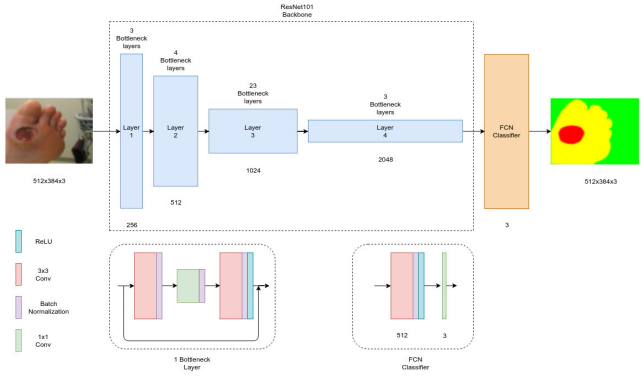

FIGURE 5. Architecture of FCN.

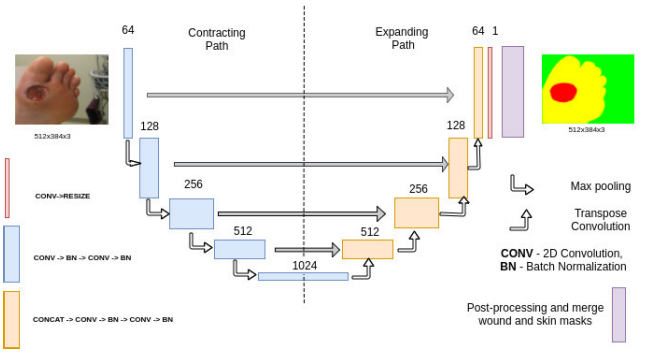

FIGURE 6. Architecture of U-Net.

followed by a classifier that segments the pixels into their respective classes. The four layers contain 3, 4, 23 and 3 bottleneck units respectively where each bottleneck consists of four convolution layers that are followed by a batch normalization step. The ReLU activation function is used after each bottleneck.

The third convolution layer in the bottleneck is a 3×3 convolutional operation while the rest are 1×1 convolutions. After the second layer, the bottleneck layers have an added dilation factor in the 3×3 convolutions for improving performance. The classifier consists of a 3×3 convolution followed by batch normalization and ReLU with dropout steps, ending with a 1×1 convolution with a channel dimension equal to the number of output classes.

2) U-NET

U-Net is a Convolutional Neural Network (CNN) encoder-decoder segmentation architecture proposed by Ronneberger *et al.* [11]. It won the ISBI cell tracking challenge in 2015 and has since been found to perform well on diverse applications of segmentation to medical images. U-Net moves and analyzes a sliding window over a large image, which enables the network to learn contextual information about the image. In our wound segmentation task, this is useful as the network needs to learn the context of skin and discover wound segments inside it. Based on fully convolutional neural networks, U-Net takes advantage of high resolution features from the convolution layers to learn the optimal up-sampling of the image.

Network Structure: The contracting path consists of 5 down convolution blocks. Each block consists of 3×3 convolution operation with ReLU activation and a 2×2 maxpooling. The U-Net architecture was slightly modified by adding batch-normalization layer after the convolution layer in order to normalize the activations. A dropout layer was also added at the end of each block to prevent over-fitting.

In the expanding path, the transpose convolution operation is utilized for upsampling. The convolution operation is the sum of the dot product of all the values in the kernel and the patch of the image. Transpose convolution does exactly the opposite by taking in single values from the feature map and multiplying them by all values of the learned kernel. This helps in fine-grained up-sampling of the feature map. To facilitate the up-sampling operation, features from the convolution layers are concatenated to the feature map obtained from the last layer. As the contracting and expanding paths are symmetric, a U-shape is formed (as seen in Fig-6), from which the architecture gets its name.

3) DeepLabV3

DeepLabV3 is a convolutional neural network, which uses atrous convolutions in either a cascaded or parallel fashion along with atrous spatial pyramid pooling, enabling the network to capture multi-scale context by using different atrous rates. The performance of DeepLabV3 matched that of other state-of-art models on the PASCAL VOC 2012 segmentation benchmark in 2017. In an ordinary convolutional neural network, pooling and striding cause a reduction in the resolution of the feature maps. Usually, deconvolutional layers are used to upsample and recover spatial resolution. Instead, DeepLabV3 uses atrous convolutions [31] that are essentially convolutions *with holes*, to effectively enlarge the field of view of filters to improve context assimilation without increasing the number of operations and filter parameters.

Atrous Spatial Pyramid Pooling (ASPP) is the main reason for DeepLabV3's impressive performance. It consists of four parallel atrous convolutions with different rates that are then applied to the feature map. The atrous convolutions in the pyramid are all followed by batch normalization. Global context is also incorporated into the model by applying global average pooling on the final feature map of the network followed by 1×1 convolution and batch normalization steps. This output is then upsampled bi-linearly to the desired spatial dimension.

Network Structure: This network also uses ResNet101 as its backbone. The first few layers of this model have a structure similar to the FCN with four layers that have 3, 4, 23 and 3 bottleneck units respectively. The classifier that follows starts off with a 1×1 convolution with batch normalization and a ReLU activation function and this output is fed into the ASPP. The convolution operations in the pyramid are 3×3 with different dilation rates. This is followed by

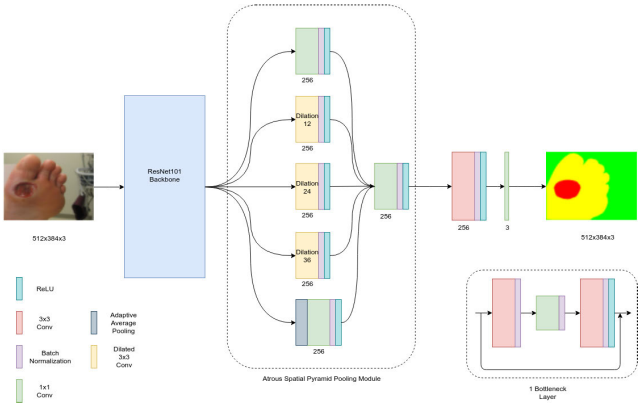

FIGURE 7. Architecture of DeepLabV3.

adaptive average pooling for global context and four convolution operations with batch normalization and ReLU activation steps. All convolutions are 1×1 except for the penultimate convolution which is a 3×3 operation.

LOSS FUNCTION

All the networks described above were trained using Binary Cross Entropy (BCE) as the Loss function.

$$BCE = \sum_{i \in N} (-g_i * \log p_i) \tag{3}$$

where p_i is the softmax output given by the network, N is image size, g is the ground truth labels $g \in \{0, 1\}$, p is the predicted label after applying the softmax operation to the output generated by the output layer of the network.

Dice Coefficient Score: is a common metric for determining the performance of image segmentation methods [32]. It quantifies the overlap of a segmented image with ground truth segmentation labels. In this paper, we use the Dice Coefficient as our evaluation metric to compare segmentation results as it incorporates both precision and accuracy. The Dice Coefficient is defined as follows -

$$dice_{coeff} = \frac{2 * \mid p_{bin} \cap g \mid}{\mid p_{bin} \mid + \mid g \mid} \tag{4}$$

where p_{bin} is the binary value of the predicted mask after performing a binary threshold on p_i at 0.5. $p_{bin} \in \{0, 1\}$.

The final loss function is a weighed sum of BCE and $dice_{coeff}$ where k is a manually tuned parameter. The BCE loss helps in increasing the confidence of the network to detect true positives whereas the dice loss penalizes the network for wrong positions of the predicted wound. As both are log losses, they are additive.

$$Loss = BCE - k * \log dice_{coeff} \tag{5}$$

POST-PROCESSING

The segmentation maps predicted by the networks are sometimes discontinuous and often require post-processing. Hence, the outputs are usually post-processed using a Conditional Random Field (CRF) with Gaussian edge potentials for improving segmentation accuracy [33]. A CRF is characterized by a Gibbs distribution and the Gibbs energy of the graph $G = (V, E)$ is defined in 1 without the higher order term.

For our implementation, the unary potential is defined as the negative log of the softmax output of the network. Thus when the output of the network for a given pixel is close to 1, the unary potential for the corresponding graph node is 0, whereas if the output is close to 0, the unary potential goes to infinity. As the unary and pairwise potentials are calculated independently, the labels predicted by the unary potential alone are significantly affected by noise. A pair-wise potential is devised to incorporate the association between neighboring pixels. The pairwise kernel is defined as in equation 6.

$$\phi^p_{i,j}(x_i, x_j) = \mu_p(x_i, x_j) \sum_{m=1}^{K} w^m k^m(f_i, f_j) \tag{6}$$

where μ is the Potts model and $K(f_i, f_j)$ are Gaussian kernels.

$$k^m(f_i, f_j) = \underbrace{w^1 \exp\left(-\frac{\mid p_i - p_j \mid^2}{2\theta_\alpha^2} - \frac{\mid I_i - I_j \mid^2}{2\theta_\beta^2}\right)}_{\text{appearance kernel}} + \underbrace{w^2 \exp\left(-\frac{\mid p_i - p_j \mid^2}{2\theta_\gamma^2}\right)}_{\text{smoothness kernel}} \tag{7}$$

The appearance kernel associates pixels with similar color and penalizes pixels with large differences in color. It considers both pixel intensities in individual image channels I and their positions p. In our case, the image vector I has $[R, G, B]$ pixel values from the input image, and is parameterized by θ_α and θ_β. The smoothness kernel penalizes only based on the nearness of the pixels and is parameterized by θ_γ.

E. TRAINING THE AHRF MODEL

AHRF uses gradient boosting techniques to optimize the unary potential and graph-cut algorithm to optimize the CRF graph. The Contrast Limited Adaptive Histogram Equalization (CLAHE) [29] pre-processing technique was found to increase the dice score of wound segmentation. Optimal parameters of the CLAHE technique were found using grid search on Dataset 1. The parameters for CLAHE implementation of openCV used in our results are kernel size of 24, 24 and clip limit of 3.0. AHRF was trained on a multi-threaded high performance cluster with 20 CPUs and 100 GB memory.The framework parallelizes feature extraction and utilized up to 40 threads.

F. TRAINING THE SEMANTIC SEGMENTATION NETWORKS

All the networks utilize high resolution features from the convolution layers in learning the optimal up-sampling of the image. In our experiments, all images were resized to a standard dimension of 512 x 384 before being input to to the network. As all the images in the dataset were of varying dimensions and aspect ratios, we averaged the dimensions of

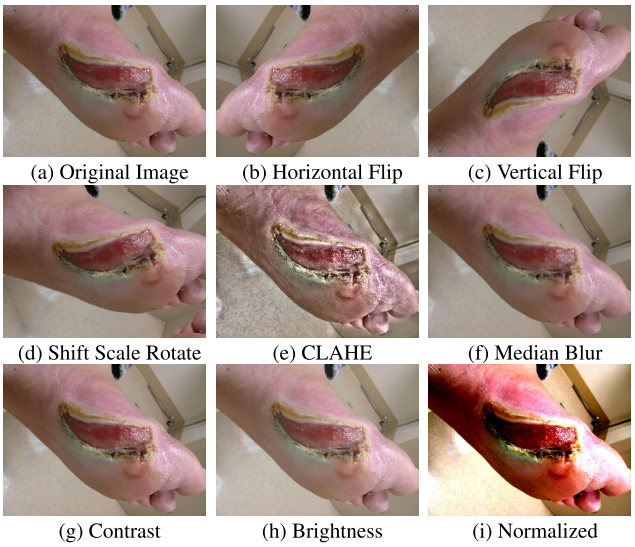

FIGURE 8. Sample image augmentations done online during training.

all images and approximated them to the closest even value required to maintain an aspect ratio of 4:3.

As the number of image samples in our datasets were inadequate for neural networks, a probabilistic data augmentation pipeline was implemented to generate synthetic augmentations using the albumentations library [34]. The augmentations used were geometric in nature including vertical flip, horizontal flip, random rotate, scale and translation. To compensate for various lighting conditions augmentations such as CLAHE, random contrast and blurring were also added to the pipeline. At run time, every augmentation was chosen with a probability p. Only one augmentation from the set CLAHE, random contrast, median blur and random brightness was chosen with a probability $p = 0.5$ and the rest of the augmentations were chosen with $p = 0.5$ each. This ensured that CLAHE and blurring, or contrast and blurring were not performed on the same image. Refer: Fig-8.

The FCN and DeepLabV3 models we utilized were pre-trained on a subset of the COCO train2017 dataset, while U-Net was initialized with weights from the Carvana Image Classification Challenge. The networks were then fine-tuned using images from wound datasets using Stochastic Gradient Descent (SGD). FCN and DeepLabV3 were trained for only 50 epochs as their superior initial weights made them converge quickly. U-Net was trained for more epochs [500-600 epochs] with early stopping. Six-fold validation was used to evaluate the generalization of the networks. The models were implemented in PyTorch [35] and its built-in optimizers were used for the training process.

FCN and DeepLabV3 were trained on a High Performance Cluster (HPC) with a Tesla K40 and 2 Intel Xeons and took one day to train all folds. On the other hand, U-Net was trained on an i7 CPU with 32GB memory and a GTX1080Ti GPU and took 5 days to train. Two separate networks were trained for U-Net - one for classifying between wound vs non-wound pixels, and the other for classifying skin vs non-skin pixels. The masks of these two networks are combined at the end to generate a final segmentation mask. All inferences were run on the GTX1080Ti. As the Gaussian edge-based CRF model used for post-processing could not be optimized during back propagation of the network, the $\theta_\alpha, \theta_\beta, \theta_\gamma$ parameters were optimized separately using grid search.

TABLE 3. Results for Dataset 1 (95 Train, 19 Validation).

Model	MAE	Haus-dorff	Dice Wound	Dice Skin	Inference time
CLAHE + AHRF	0.0904	11.553	0.750	0.9060	3-5 min
U-Net	0.1880	13.983	0.490	0.7950	40 msec
U-Net + CRF	0.1552	13.165	0.520	0.8900	2 sec
CLAHE + U-Net	0.1523	13.085	0.532	0.8901	50 msec
CLAHE + U-Net + CRF	0.1231	12.365	0.591	0.8903	2 sec
FCN	**0.0681**	**10.791**	**0.7822**	**0.9410**	41 msec
CLAHE + FCN	0.0777	11.136	0.7667	0.9378	51 msec
DeepLabV3	0.0783	11.164	0.7625	0.9352	56 msec
CLAHE + DeepLabV3	0.0811	11.220	0.7595	0.9320	66 msec

G. EVALUATION

All semantic segmentation methods were evaluated using k-fold cross validation over the entire dataset with $k = 6$. Performance of the model on test set is measured by using the Dice Coefficient Score.

V. RESULTS AND DISCUSSIONS

1) COMPARING SEGMENTATION INFERENCE TIME

AHRF is a graph optimization method and takes about 3-5 minutes to infer segmentation masks for a single image of size 512×384 on all three datasets (see column 4 of Tables - 3 to 5). Although the graph optimization step is faster, the feature extraction and evaluation steps makes inference in AHRF significantly slow. Consequently, it would be challenging to implement AHRF on mobile devices. CNNs on the other hand utilize a series of matrix multiplications and additions amenable for implementation on GPUs, which most smartphones are equipped with. FCN, U-Net and DeepLabV3 had an average inference time of approximately 41, 50 and 56 milliseconds on all three datasets.

2) COMPARING SEGMENTATION ACCURACY

Dataset 1: As observed in Table-3, AHRF is significantly more accurate than U-Net by a difference of 0.159 dice score on the wound segments in dataset 1. Both Pre-processing (CLAHE) and Post-processing (CRF) improve the performance of the segmentation of U-Net. However, even with these pre- and post- processing techniques, U-Net is not as accurate as AHRF. On the other hand, FCN and DeepLabV3 both outperform AHRF even with less data, which can be attributed to the models being trained on a subset of COCO train2017 and then fine-tuned to our dataset. FCN outperforms DeepLabV3 by 0.0197 in dice score which is because FCN is a lighter model and hence,

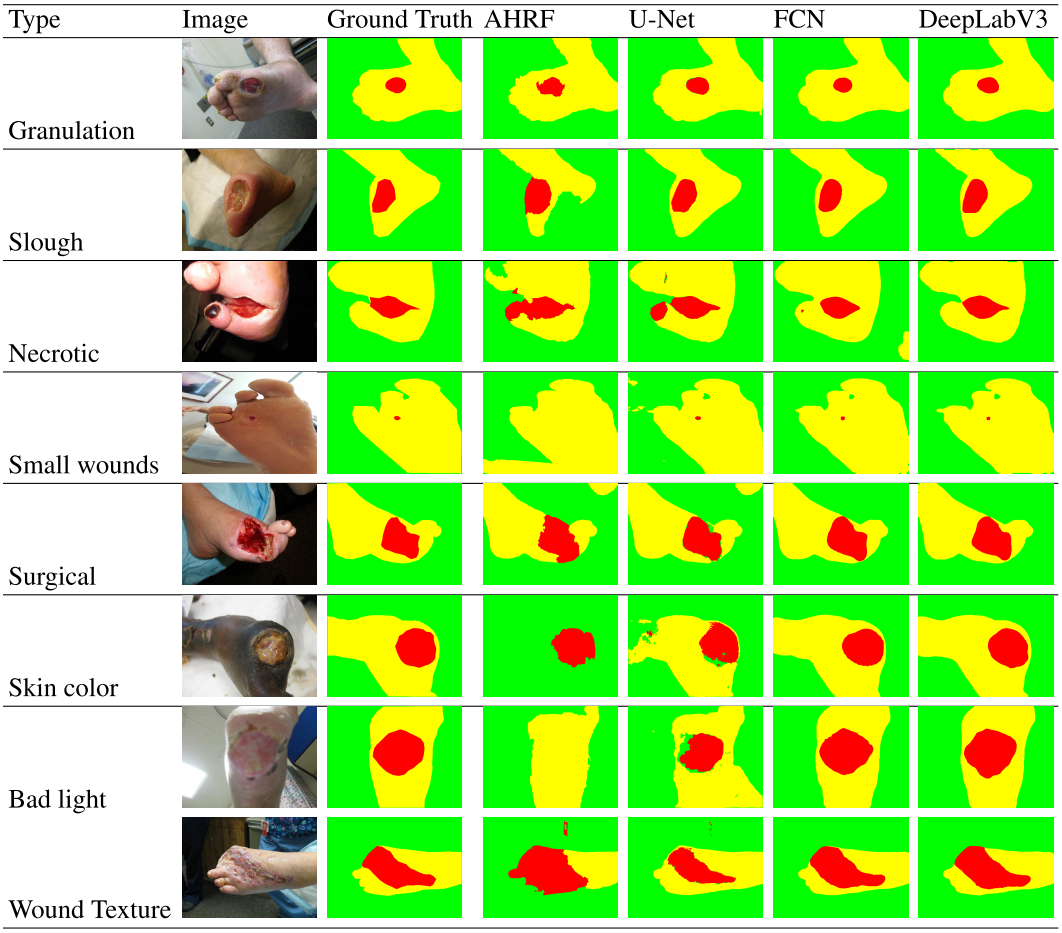

FIGURE 9. Performance of AHRF, U-Net, FCN and DeepLabV3 trained on Dataset 3 for segmenting a variety of images with different colors, textures and lighting conditions of skin, wound and background.

fits the data distribution slightly better than DeepLabV3. FCN and DeepLabV3 outperforms AHRF by dice scores of 0.0322 and 0.0125 respectively.

Dataset 2: As the dataset size increases, the networks generalize to the distinct features and textures that define a wound. As seen in Table-4, U-Net has a slightly higher dice score than AHRF (more accurate). Pre-processing U-Net using CLAHE improved its accuracy but the improvement observed is less than that obtained for Dataset 1 but it underperforms FCN and DeepLabV3 again, with a difference of 0.124 and 0.122 in dice score respectively. Ultimately, as the size of the training data increases, U-Net's dependence on pre- and post- processing decreases as it learns better features. The performance of FCN and DeepLabV3 is not affected by pre- and post- processing due to their pre-trained weights and model architectures (DeepLabV3). FCN and DeepLabV3 outperformed AHRF by dice scores of 0.135 and 0.133 respectively.

Dataset 3: The third dataset containing 1442 images is roughly four times the size of dataset 2. Even though it has more variance (see Table -1, Table-2), the CNNs generalize to all types of wounds and generate segmentation masks close to the ground truth, whereas the performance of AHRF decreases slightly. As observed in Fig-13 - sample 1, AHRF tends to get confused for the same image as the variations in the dataset increases, making it less robust. The CNNs also generate better segmentation masks for smaller wounds as seen in Fig-9-sample 4. U-Net has a significantly higher dice score than AHRF with a margin of 0.106 dice co-efficient and does not require any pre/post-processing. AHRF is observed to over-segment and often performs poorly on edges and wounds with difficult textures. FCN and DeepLabV3 still outperform U-Net by a dice score of 0.117 and 0.121 respectively, which highlights the impact of using pre-trained

TABLE 4. Results for Dataset 2 (263 Train, 53 Validation).

Model	MAE	Haus-dorff	Dice Wound	Dice Skin	Inference time
CLAHE + AHRF	0.1072	11.969	0.706	0.8865	3-5 min
U-Net	0.1152	12.169	0.665	0.897	40 msec
U-Net + CRF	0.1147	12.156	0.667	0.897	2 sec
CLAHE + U-Net	0.1028	11.861	0.717	0.892	50 msec
CLAHE + U-Net + CRF	0.1023	11.848	0.716	0.895	2 sec
FCN	**0.0645**	**10.907**	**0.8418**	**0.9342**	41 msec
CLAHE + FCN	0.0744	11.356	0.8220	0.9262	51 msec
DeepLabV3	0.0662	10.949	0.8392	0.9330	56 msec
CLAHE + DeepLabV3	0.07201	11.318	0.8268	0.9278	66 msec

TABLE 5. Results for Dataset 3 (1201 Train, 241 Validation).

Model	MAE	Haus-dorff	Dice Wound	Dice Skin	Inference time
CLAHE + AHRF	0.1235	12.377	0.6287	0.9016	3-5 min
U-Net	0.0832	11.372	0.733	0.9506	40 msec
U-Net + CRF	0.0830	11.369	0.734	0.9506	2 sec
CLAHE + U-Net	0.0875	11.479	0.7200	0.9474	50 msec
CLAHE + U-Net + CRF	0.0878	11.486	0.719	0.9473	2 sec
FCN	0.0464	10.092	0.8518	0.9532	41 msec
CLAHE + FCN	0.0559	10.878	0.8357	0.9518	51 msec
DeepLabV3	**0.0379**	**9.764**	**0.8554**	**0.9617**	56 msec
CLAHE + DeepLabV3	0.0534	10.631	0.8390	0.9578	66 msec

TABLE 6. Common Validation Set - WOUND.

Model	Dataset 1	Dataset 2	Dataset 3
AHRF	0.673	0.687	0.675
U-Net	0.416	0.717	0.784
FCN	**0.7822**	0.8453	0.859
DeepLabV3	0.7625	**0.8537**	**0.876**

Average dice coefficients for common images in all 3 datasets.

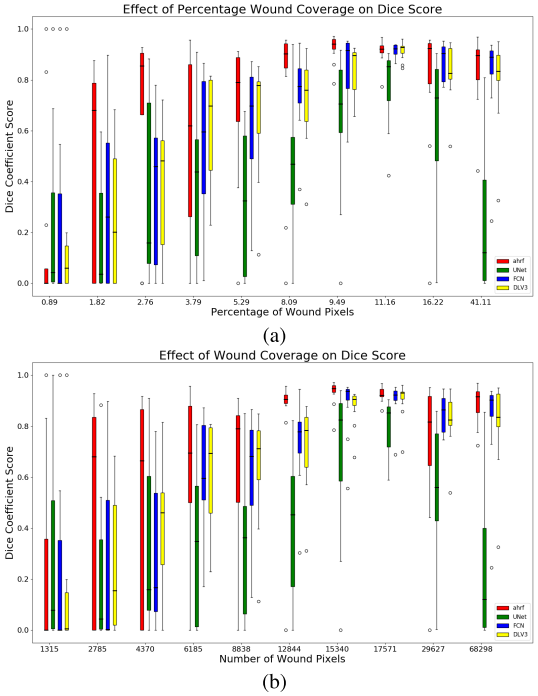

FIGURE 10. Box plots of wound percentage of wound pixels vs dice coefficient for dataset 1.

models. DeepLabV3, a deeper model, outperforms FCN as dataset 3 has significantly more data for it to work with.

a: COMMON VALIDATION DATASET

In order to get a final conclusion on the accuracy of all the segmentation methods, we compared their segmentation accuracy on a common validation set after being trained on datasets 1, 2 and 3 respectively (see Table-6) It can be concluded from this table that the accuracy of deep learning models increases with increase in the data samples while the performance remains same or sometimes worsens for AHRF, a graph based segmentation architecture.

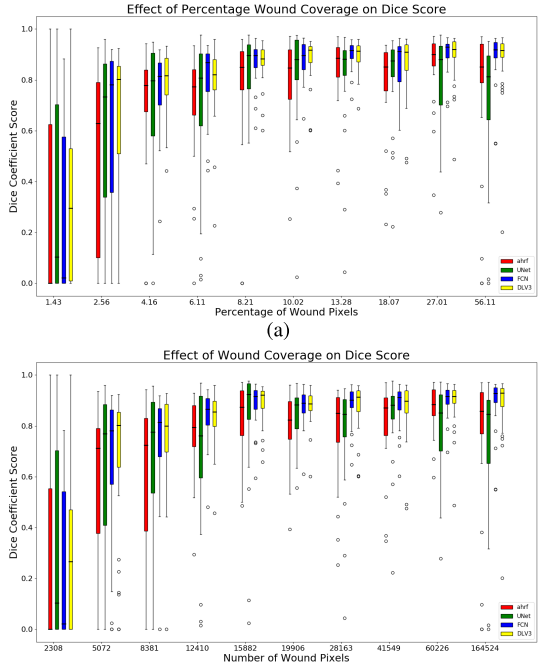

FIGURE 11. Box plots of wound percentage of wound pixels vs dice coefficient for dataset 2.

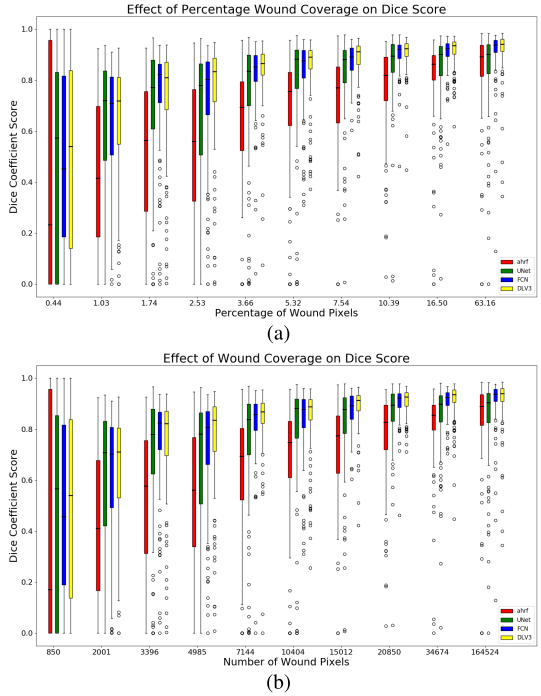

FIGURE 12. Box plots of wound percentage of wound pixels vs dice coefficient for dataset 3.

3) MODEL ROBUSTNESS TO WOUND COLORS IN BACKGROUND

In many wound imaging situations, colors found in many wounds such as red and yellow may appear in the background by accident. Thus, it is important to compare how robust (i.e. does not detect those background colors as part of the wound)

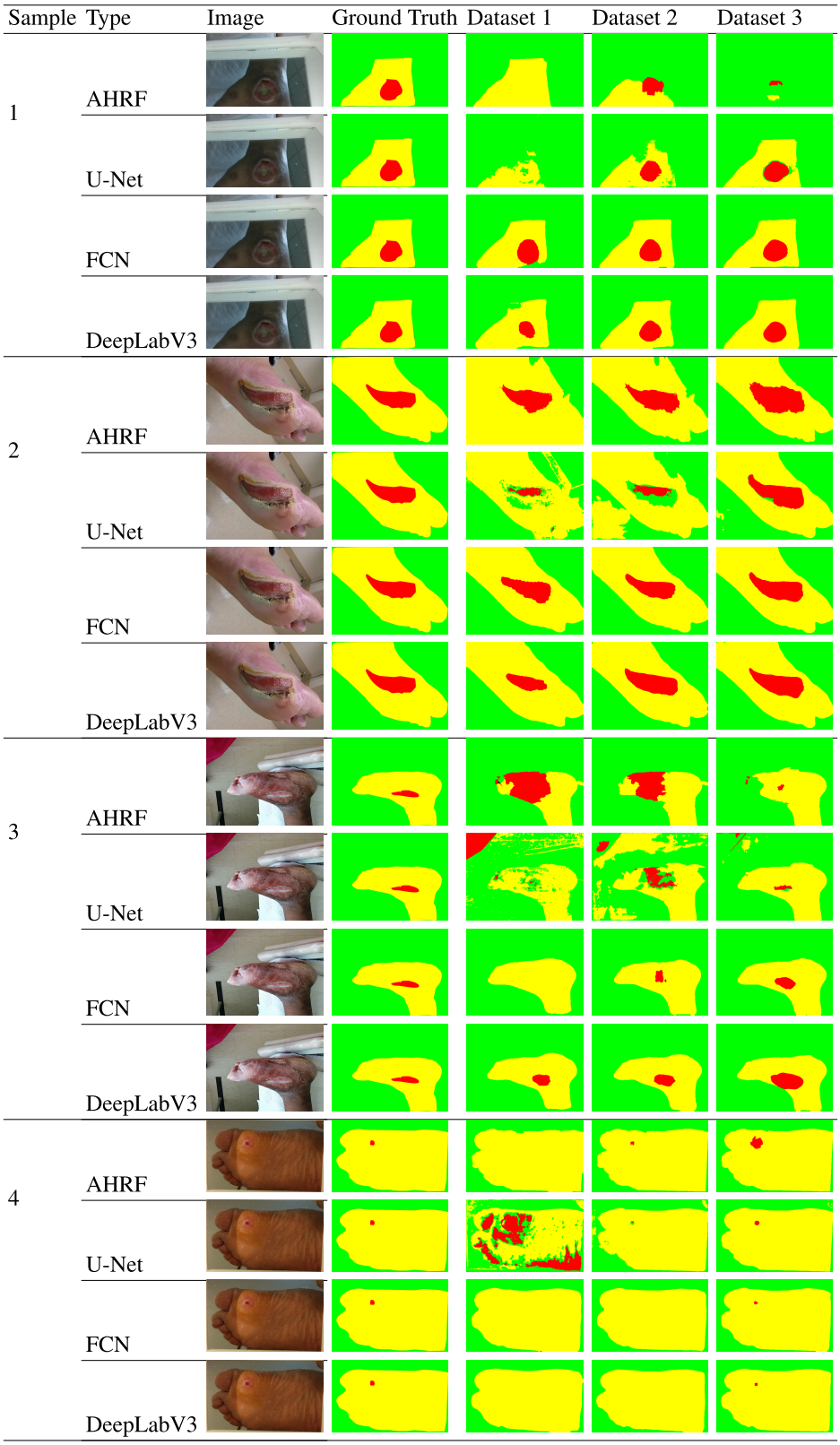

FIGURE 13. Comparison of AHRF, U-Net, FCN, DeepLabV3 and their accuracy trends on all 3 datasets. Sample 1 shows how the accuracy of AHRF improves from Dataset 1 to Dataset 2 but then decreases when more, noisier data is added in Dataset 3. The deep learning networks on the other hand shows consistent improvement as more data is added. The samples demonstrate how skin pixels are segmented more accurately than the wound segment because of the huge class imbalance in data.

the segmentation methods are when such colors appear in the background. Since the networks are pre-trained and are being fine-tuned on the wound segmentation task, the network tries to learn the most prominent features of the wound at first. It can be clearly observed in Fig-13 - sample 3, dataset 1, that U-Net initially (on smaller datasets) tends to classify any red color in the wound image as belonging to the wound segment. This can be justified from Table-2 which shows that the mean value of the Red channel of the wound segment of dataset 1 is higher than the Blue and Green channels. However, as U-Net is trained on more data, it starts to learn and rely on texture information as well. This can be seen in Fig-13 sample 3, where U-Net does not confuse the red cloth in the top left corner with the wound when trained on dataset 3. FCN and DeepLabV3 do not face this issue as they utilize pre-trained weights, alleviating their dependence on just color. AHRF on the other hand uses hand-crafted features, is more robust to wound colors in the background. It requires fewer images to achieve its performance limits and thus does not confuse the red cloth with the wound irrespective of which dataset it has been trained on. This shows that handcrafted features help AHRF understand textures better than U-Net when trained on smaller datasets but due to information contained in their initial weights, FCN and DeepLabV3 already take textures into consideration.

4) EFFECT OF CLASS IMBALANCE

We compared the accuracy of the CNNs and AHRF for wound images with varying sizes of wound and skin segments. It can be observed in Sample 4 of Fig-9 how detection of skin pixels (larger segments) is better than that of the wound segment (smaller) for the networks because of the huge class imbalance in data. This trend is not observed for AHRF because AHRF is trained jointly for all three classes. Hence, the wound classifier can utilize the information learned for skin. For example, areas not classified as skin but surrounded by skin automatically get a higher probability of belonging to the wound class.

5) SENSITIVITY TO THE RELATIVE PROPORTION OF THE WOUND SEGMENT

The sensitivity of segmentation to changes in the proportion of image covered by the wound is studied for all three datasets. Figures 10, 11, and 12 show the accuracy of AHRF and the CNNs as the wound size varies in the form of box plots. We show box plots that includes information on both the mean Dice score as well as its variation across various folds. The width of the box plot shows how stable the reported mean Dice score is across various folds. Dice score variance is shown for percentage of wound pixels in the images. The wound percentage is defined as the ratio of number of wound pixels to the total number of pixels. Due to the connective property of AHRF which results from its clique potential, it fails to work well on images that have small wounds because neighboring skin pixels cause a small wound to also be classified as skin. The deep learning networks do not face this problem and work well on wounds of a small size.

The box-plot in Fig-10 shows that AHRF performs better than U-Net even with large variations in the wound size for dataset 1 while FCN and DeepLabV3 match its performance. The CNNs fail to detect wounds smaller than 10% of the wound image whereas AHRF generates some slight segmentations. The box-plot in Fig-11 and Fig-12 shows increased accuracy with as wound size increases for datasets 2 and 3. The height of the boxes shows variance in the performance of all architectures. It can be observed that images with more than 5% of wound pixels have better results for all the architectures. This result can be used to create a guideline for taking usable wound images or cropping the images in a pre-processing phase by keeping the wound percentage more than 5%. For instance, the photographer can be asked to retake (or zoom in) images in which the wound percentage is less than 5%.

6) SEGMENTATION ACCURACY FOR WOUNDS WITH DIFFERENT WOUND ATTRIBUTES AND SKIN TYPES

As seen in Fig-9, both AHRF and the CNNS have shown good generalizability to various wound tissue types, skin colors and lighting conditions. Granulation, slough and necrotic are different types of wound tissue which occur in wounds, which differ in their color and texture. However, both AHRF and the networks have shown good segmentation results on wounds containing a combination of these tissues. The networks generalize well to darker skin tones and bad lighting conditions as well.

VI. DISCUSSIONS AND CONCLUSION

In this work, a comprehensive systematic analysis of semantic segmentation of smartphone camera captured wound images using AHRF, FCN, U-Net and DeepLabV3 has been performed. All segmentation methods achieve good results which generalize well in wound images with various skin and wound tissue types, and background clutter. However, due to differences in the two approaches (AHRF vs deep learning), some trade-offs have to be considered before deciding on a model for practical purposes.

AHRF had increased segmentation accuracy when input images were pre-processed using CLAHE. CLAHE pre-processing with U-Net showed improvements only for smaller datasets. CRF post-processing also improved the accuracy of U-Net on smaller datasets. Pre- and post- processing did not change the performance of FCN and DeepLabV3.

AHRF is more accurate and generalizes better than U-Net for small datasets (< 300 images) but is outperformed by fine-tuned FCN and DeepLabV3 models pre-trained on PASACL VOC: AHRF has more reliable predictions because it depends on texture features and not just color. Its hand-crafted visual features also enable it to learn wound features with fewer images. U-Net on the other hand, performed moderately well for segmenting skin but not wound pixels on Dataset 1 (smallest dataset). FCN and DeepLabV3 performed well in segmenting both skin and wound pixels across all 3 datasets.

CNNs are more accurate for larger datasets (> 300 images). As the size of dataset increases, the segmentation accuracy of the deep learning networks increase while that of AHRF saturates after a point and sometimes even worsens with the addition of more training data. As FCN, U-Net and DeepLabV3 have many more trainable hyperparameters than AHRF, they are able to absorb and utilize more data and generalize better. They also show better performance on smaller wound sizes as compared to AHRF. This is because AHRF has a region growing property due to its pairwise and clique potentials which causes smaller wounds to sometimes become part of the surrounding skin clique which are wrongly segmented as skin.

CNNs have a considerably faster inference time than AHRF: mainly because AHRF uses many hand-crafted features and clustering techniques, which take time to be computed. In our experiments, AHRF took about 4-5 minutes for segmenting one image while FCN, U-Net and DeepLabV3 could segment the same image in about 40, 50, 60 milliseconds respectively. This makes the networks a more viable option for implementation on mobile devices, where resources are constrained, especially if real-time segmentation is required. The long inference time of AHRF makes it difficult to use even in a client-server scenario, as a network connection would probably timeout before segmentation is complete.

Initial weights of deep learning approaches make a considerable difference: U-Net generally outperforms FCNs, but FCN outperforms U-Net in our experiments by a margin 0.075 for dataset 3 as seen in Table-6. FCN and DeeplabV3 were initialized with pre-trained weights from COCO train2017 while U-Net was initialized with weights from the Carvana Image Classification Challenge. Using these weights for U-Net was better than using random initialization but are still no match for COCO train2017 weights. DeepLabV3 outperformed FCN by a margin of 0.017 for dataset 3 in Table-6.

VII. FUTURE WORK

One possible future direction for this research could be experimenting more with lighting variations and performing an error analysis of the various factors which affect the segmentation performance of AHRF and the deep learning models. Models can be made more robust by using Generative Adversarial Networks (GANs) [36] for synthesizing more training data. More effective ways of image pre-processing such as auto-augmentation [37] can also be used which trains a neural network to decide on the best possible pre-processing step for a given input image. Finally, parallelizing AHRF to make it faster, especially the feature extraction might be a fruitful direction for further research.

ACKNOWLEDGMENT

The authors would like to thank computational resources supported by the Academic & Research Computing group at Worcester Polytechnic Institute for the access to turing high performance cluster acquired through NSF MRI grant DMS-1337943 to WPI. *(Ameya Wagh and Shubham Jain contributed equally to this work.)* This work was partly funded by the National Institutes for Health (NIH) National Institute for Biomedical Imaging and Bioengineering (NIBIB) under Grant 1R01EB025801-01.

REFERENCES

[1] J. Beck, "2017 national standards for diabetes self-management education and support," *Diabetes Educator*, vol. 44, no. 1, pp. 35–50, 2018.
[2] 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 Regeneration*, vol. 17, no. 6, pp. 763–771, Nov. 2009.
[3] P. E. Houghton, C. B. Kincaid, K. E. Campbell, M. G. Woodbury, and D. H. Keast, "Photographic assessment of the appearance of chronic pressure and leg ulcers.," *Ostomy/Wound Manage.*, vol. 46, no. 4, pp. 6–20, 28–30, 2000.
[4] C. A. Murphy, P. Houghton, T. Brandys, G. Rose, and D. Bryant, "The effect of 22.5 kHz low-frequency contact ultrasound debridement (LFCUD) on lower extremity wound healing for a vascular surgery population: A randomised controlled trial," *Int. Wound J.*, vol. 15, no. 3, pp. 460–472, Jun. 2018.
[5] Ø. H. Sundby, I. Irgens, L. Ø. Høiseth, I. Mathiesen, E. Lundgaard, H. Haugland, H. Weedon-Fekjár, J. O. Sundhagen, G. Sandbàk, and J. Hisdal, "Intermittent mild negative pressure applied to the lower limb in patients with spinal cord injury and chronic lower limb ulcers: A crossover pilot study," *Spinal Cord*, vol. 56, no. 4, pp. 372–381, Apr. 2018.
[6] P. Sheehan, P. Jones, A. Caselli, J. M. Giurini, and A. Veves, "Percent change in wound area of diabetic foot ulcers over a 4-Week period is a robust predictor of complete healing in a 12-week prospective trial," *Diabetes Care*, vol. 26, no. 6, pp. 1879–1882, Jun. 2003.
[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.
[8] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587
[9] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351, 2015, pp. 234–241.
[12] 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 *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 1017–1021.
[13] 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 Trans. Emerg. Topics Comput. Intell.*, early access, Sep. 14, 2018, doi: 10.1109/TETCI.2018.2866254.
[14] T.-Y. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, 2014, pp. 740–755.
[15] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr, "Associative hierarchical random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1056–1077, Jun. 2014.
[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*. [Online]. Available: http://arxiv.org/abs/1512.03385
[17] A. A. Perez, A. Gonzaga, and J. M. Alves, "Segmentation and analysis of leg ulcers color images," in *Proc. Int. Workshop Med. Imag. Augmented Reality*, 2001, pp. 262–266.
[18] R. Mukherjee, D. D. Manohar, D. K. Das, and A. Achar, "Automated tissue classification framework for reproducible chronic wound assessment," *BioMed Res. Int.*, vol. 2014, Jul. 2014, Art. no. 851582.
[19] 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 Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2098–2109, Sep. 2017.

[20] M. K. Yadav, D. D. Manohar, G. Mukherjee, and C. Chakraborty, "Segmentation of Chronic Wound Areas by Clustering Techniques Using Selected Color Space," *J. Med. Imag. Health Informat.*, vol. 2, no. 3, pp. 22–29, 2013.
[21] H.-F. Shih, T.-W. Ho, J.-T. Hsu, C.-C. Chang, F. Lai, and J.-M. Wu, "Surgical wound segmentation based on adaptive threshold edge detection and genetic algorithm," *Proc. SPIE*, vol. 225, Feb. 2017, Art. no. 1022517.
[22] M. Goyal, M. H. Yap, N. D. Reeves, S. Rajbhandari, and J. Spragg, "Fully convolutional networks for diabetic foot ulcer segmentation," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2017, pp. 618–623.
[23] X. Liu, C. Wang, F. Li, X. Zhao, E. Zhu, and Y. Peng, "A framework of wound segmentation based on deep convolutional networks," in *Proc. 10th Int. Congr. Image Signal Process.*, Jan. 2018, pp. 1–7.
[24] M. Goyal, N. Reeves, S. Rajbhandari, and M. H. Yap, "Robust methods for real-time diabetic foot ulcer detection and localization on mobile devices," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 4, pp. 1730–1741, Jul. 2018.
[25] F. Li, C. Wang, X. Liu, Y. Peng, and S. Jin, "A composite model of wound segmentation based on traditional methods and deep neural networks," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–12, May 2018.
[26] K.-K. Maninis, S. Caelles, J. Pont-Tuset, and L. Van Gool, "Deep extreme cut: From extreme points to object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 616–625.
[27] L. Wang, P. C. Pedersen, D. M. Strong, B. Tulu, E. Agu, and R. Ignotz, "Smartphone-based wound assessment system for patients with diabetes," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 2, pp. 477–488, Feb. 2015.
[28] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Proc. Graph. Gems*, 1994, p. 6.
[29] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Comput. Vis., Graph., Image Process.*, vol. 39, pp. 355–368, Sep. 1987.
[30] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.
[31] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," 2016, *arXiv:1606.00915*. [Online]. Available: http://arxiv.org/abs/1606.00915
[32] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 2006.
[33] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 109–112.
[34] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, Feb. 2020.
[35] A. Paszke, G. Chanan, Z. Lin, S. Gross, E. Yang, L. Antiga, and Z. Devito, "Automatic differentiation in PyTorch," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, 2017, pp. 1–4.
[36] J. Li, J. Jia, and D. Xu, "Unsupervised representation learning of image-based plant disease with deep convolutional generative adversarial networks," in *Proc. 37th Chin. Control Conf. (CCC)*, Jul. 2018, pp. 1–16.
[37] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," 2018, *arXiv:1805.09501*. [Online]. Available: https://arxiv.org/abs/1805.09501

AMEYA WAGH received the M.S. degree in robotics engineering from the Worcester Polytechnic Institute, Worcester, MA, USA, in 2019. He is currently a Software Engineer with TORC Robotics. His current research interests include computer vision and deep learning.

SHUBHAM JAIN received the B.Tech. degree from IIT Kanpur, India, in 2017, and the M.S. degree in robotics engineering from the Worcester Polytechnic Institute, Worcester, MA, USA, in 2019. He is currently a Computer Vision Engineer with NVIDIA. His current research interests include autonomous driving and related computer vision problems.

APRATIM MUKHERJEE is currently pursuing the bachelor's degree in computer science and engineering with the Manipal Institute of Technology, India. His research interests include robotics and computer vision.

EMMANUEL AGU (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Massachusetts Amherst, Amherst, MA, USA, in 2001. He is currently a Professor with the Computer Science Department, Worcester Polytechnic Institute, Worcester, MA, USA. He has been involved in research in mobile and ubiquitous computing for over 16 years. He is currently involved in mobile health projects to assist patients with diabetes, obesity, and depression.

PEDER C. PEDERSEN (Senior Member, IEEE) received the B.S. degree in electrical engineering from the Aalborg Engineering College, Aalborg, Denmark, in 1971, and the M.E. and Ph.D. degrees in bioengineering from The University of Utah, Salt Lake City, UT, USA, in 1974 and 1976, respectively. In October 1987, he joined the faculty of the Worcester Polytechnic Institute, Worcester, MA, USA, where he is currently an Emeritus Professor at the Electrical and Computer Engineering Department. He was an Associate Professor with the Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA, USA. His research interests include elastography methods for quantitative imaging of the Young's modulus in soft tissues and the development of a low-cost, portable personal ultrasound training simulator with structured curriculum and integrated assessment methods to satisfy the training needs of the widely used point-of-care scanners. Another research effort has been the design of a smartphone-based diabetic wound analysis system, specifically for foot ulcers.

DIANE STRONG received the B.S. degree in mathematics and computer science from the University of South Dakota, Vermillion, SD, USA, in 1974, the M.S. degree in computer and information science from the New Jersey Institute of Technology, Newark, NJ, USA, in 1978, and the Ph.D. degree in information systems from the Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA, USA, in 1989. Since 1995, she has been a Professor with the Worcester Polytechnic Institute, Worcester, MA, USA, where she is currently a Full Professor with the Foisie School of Business and also the Director of the Information Technology Programs. She is also a member of the Faculty Steering Committee of the Healthcare Delivery Institute, WPI. Since 2006, she has been focused on effectively using IT to promote health and support healthcare delivery. Her research has been concerned with an effective use of IT in organizations and by individuals.

BENGISU TULU (Member, IEEE) received the Ph.D. degree in management of information systems and technology from Claremont Graduate University, CA, USA. She is currently an Associate Professor with the Foisie Business School, Worcester Polytechnic Institute, Worcester, MA, USA. She is one of the founding members of the Healthcare Delivery Institute, WPI. Her research interest includes the development and implementation of health information technologies and the impact of these implementations on healthcare organizations and consumers.

CLIFFORD LINDSAY received the B.S. degree in computer science from the University of California at San Diego, La Jolla, CA, USA, in 2001, and the Ph.D. degree in computer science from the Worcester Polytechnic Institute, Worcester, MA, USA, in 2011. He is currently an Assistant Professor with the University of Massachusetts Medical School. He is also involved in applying computer vision and image processing methods to improve the quality of medical images.

ZIYANG LIU is currently pursuing the Ph.D. degree in computer science with the Worcester Polytechnic Institute, MA, USA. His current research interests include computer vision and deep learning.