HealTech - A System for Predicting Patient Hospitalization Risk and Wound Progression in Old Patients

Subba Reddy Oota[1,2*], Vijay Rowtula[1,2*], Shahid Mohammed[1], Jeffrey Galitz[1], Minghsun Liu[1], Manish Gupta[2†]
[1]Woundtech Innovative Healthcare Solutions, [2]IIIT-Hyderabad, India
{soota, vrowtula, shmohammed, jgalitz, mliu}@woundtech.net, manish.gupta@iiit.ac.in

Abstract

How bad is my wound? How fast will the wound heal? Do I need to get hospitalized? Questions like these are critical for wound assessment, but challenging to answer. Given a wound image and patient attributes, our goal is to build models for two wound assessment tasks: (1) predicting if the patient needs hospitalization for the wound to heal, and (2) estimating wound progression, i.e., weeks to heal. The problem is challenging because wound progression and hospitalization risk depend on multiple factors that need to be inferred automatically from the given wound image. There exists no work which performs a rigorous study of wound assessment tasks considering multiple wound attributes inferred using a large dataset of wound images. We present HealTech, a two-stage wound assessment solution. The first stage predicts various wound attributes (like ulcer type, location, stage, etc.) from wound images, using deep neural networks. The second stage predicts (1) whether the wound would heal (using conventional in-house treatment) or not (needs hospitalization), and (2) the number of weeks to heal, using an evolutionary algorithm based stacked Light Gradient Boosted Machines (LGBM) model. On a large dataset of 125711 wound images, HealTech achieves a recall of 83 and a precision of 92 for wounds with the risk of hospitalization. For wounds that can be healed without hospitalization, precision and recall are as high as 99. Our wound progression model provides a mean absolute error of 3.3 weeks.

1. Introduction

Nearly 15% of Medicare beneficiaries (8.2 million) had at least one type of wound or infection. The total Medicare spending estimates for all wound types (in the US) ranged from $28.1 to $96.8 billion [1]. Including infection costs, the most expensive estimates were for surgical wounds, followed by diabetic foot ulcers. Hence, any steps in reducing the cost incurred per wound would significantly impact the overall healthcare spending.

Most wounds (∼90% in our dataset) do not need hospitalization. Determining whether you need to visit a doctor or not, and weeks to heal depends on the following factors[1]: (1) how big is the wound (area, depth, has ragged edges, has debris), (2) how bad is the bleeding, (3) wound location (around a joint such as elbow or knee, face, hand, genitals, mouth, near the eye), and (4) is it getting infected (fever, red streaks, oozing pus, etc.). While these are some basic rules, often, it is challenging to apply these objectively and determine the hospitalization risk and weeks to heal.

Incorrect decision to visit a hospital or not has its repercussions. Unnecessary hospitalization leads to discomfort for the patient, costs, and possible hospital-acquired infections. Delay in hospitalization may lead to increased treatment costs, less effective treatment, infections, and in general, worsening of the wound. Traditionally, wounds are assessed manually by clinicians who, in turn, document various observations. Some of the challenges for the clinician diagnosis include: (i) frequent assessments of a patient, (ii) the entry of wound attributes in the database, (iii) applying the right diagnosis and (iv) inter-observer discordance. Automating the wound assessments can help overcome some of the limitations and aid the clinicians to make informed decisions.

In this paper, our first goal is to build a model to predict the patient's risk of hospitalization. Further, for wounds that show a higher likelihood of healing without hospitalization, our second goal is to predict the number of weeks to heal. To achieve the two goals, our proposed system, HealTech, mainly depends on wound images. It also uses patient attributes like age, BMI, etc., since they are complementary and cannot be predicted using wound images. We argue that the development of such a system can help in early detection of the complexities in the wound, which might affect the healing process and also reduce the time

*The first two authors made equal contribution.
†The author is also a Principal Applied Scientist at Microsoft.

[1] https://www.webmd.com/first-aid/does-this-cut-need-stitches

Figure 1. Examples of wound images labeled with various attributes used in our dataset.

spent by a clinician to diagnose the wound. The problem is challenging because wound progression and hospitalization risk can depend on multiple factors like wound/ulcer type, wound location, wound size, etc. Other challenges include combining multiple wound image related factors with patient demography parameters, variations across images in terms of lighting conditions, skin color variations, etc. and use them for the wound assessment tasks.

Motivated by the immense success of convolutional neural networks (CNNs) across various image analytics tasks, we investigate their application for prediction tasks related to wound images. Figure 1 shows a sample of wound images labeled with a variety of wound attributes. The main contributions of this paper are as follows.

- We formulate the problem as a two-stage deep learning based method for hospitalization risk and wound progression prediction.
- In the first stage, we analyze multiple factors to be obtained from wound images for wound characterization. We build classifiers to estimate such factors.
- In the second stage, we develop a heal/hospitalization (or hospitalization risk) classifier and a wound progression (or weeks to heal prediction) model using an evolutionary algorithm-based stacked Light Gradient Boosted Machines (LGBM), which uses the wound factors predicted from first stage, as well as a list of five patient features.
- We experiment with a large dataset of 125711 images, unlike previous research works that experiment with much smaller datasets.

2. Related Work

Wound image analysis: Chronic wound diagnosis, monitoring, and healing process of a wound is an ongoing research area in the field of medical image analysis. While several methods have been proposed in the literature to classify the wound tissue or segmentation of related skin lesions, these experiments fail to provide a robust tool for process automation [2–4]. The classical image processing techniques such as color descriptors and texture detectors have been used to extract features from wound images to classify the skin patches as normal or abnormal to automatically monitor the healing process in [5–10].

Deep learning for medical imaging: Motivated by the immense success of deep learning techniques in general vision, speech as well as text problems, there has been a lot of focus on applying deep learning for medical imaging recently [11, 12]. Specifically, problem areas include image quality analysis [13, 14], image segmentation [15–17], image/exam classification [18, 19], object/lesion classification [20], registration (i.e. spatial alignment) of medical images [21], image enhancement, and image reconstruction [22]. While several researchers used deep learning models for segmenting and classifying the different ulcer images in [23–25], the corpus size is minimal (e.g., Medetec wound images dataset [26] which has only 607 images) leading to relatively brittle systems. Few papers focused on wound assessment [14, 27, 28], but the authors experimented on a limited set of wound attributes and ulcer types, diagnosis was limited to wound type classification, and dataset sizes were 100x smaller than our dataset.

3. HealTech System Architecture

Figure 2 illustrates the architecture of the proposed two-stage system. In the first stage, we use CNNs to predict 14

Wound attribute	# Instances	# Classes	Top classes
Wound/Ulcer Type	125711	5	Diabetic Ulcer (19773), Pressure Ulcer (47541), Surgical Wound (12238), Trauma Wound (13667), Venous Ulcer (32492)
Wound Location	91177	6	Lower Leg (31775), Sacral (20501), Foot (12753), Heel (11226), Ankle (10375), Great Toe (4547)
Wound Stage (3 sub types)	17918 41055 52530	3 3 2	Diabetic: Grade-2 (13248), Grade-1 (2401), Grade-3 (2269) Pressure: Grade-3 (20343), Grade-4 (10581), Unstageable (10131) Remaining Ulcers: Full Thickness (47849), Partial Thickness (4681)
Wound Margin	113519	7	Attached to Wound Base (45595), Flat and Intact (23735), Well Defined (12917), Undefined (10784), Not Attached to Wound Base (7172), Flattened (6741), Thickened (6575)
Wound Area	122417	NA	Numerical values
Wound Volume	99719	NA	Numerical values
Joint Necrosis Exposed	125661	2	No (101275), Yes (24386)
Ligament Necrosis Exposed	125666	2	No (101282), Yes (24384)
Adipose Necrosis Exposed	125607	2	No (101195), Yes (24412)
Muscle Necrosis Exposed	125598	2	No (101261), Yes (24337)
Exudate	90792	2	Class-1 (56124), Class-2 (34668)
Red Granulation	88818	4	Class-1 (54229), Class-2 (12120), Class-3 (11355), Class-4 (11114)
Bone Necrosis Exposed	125641	2	No (101261), Yes (24380)
Adherent Yellow Slough	35803	5	Class-1 (13770), Class-2 (8814), Class-3 (8223), Class-4 (3923), Class-5 (1073)

Table 1. Statistics for the 14 wound attributes in our dataset

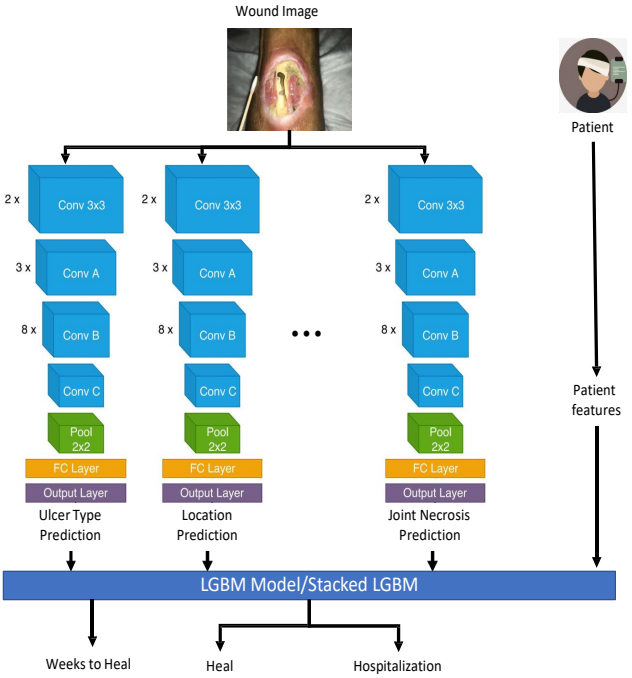

Figure 2. Model architecture with two stages of wound assessment. The first stage predicts features from wound images. The second stage uses wound image features along with patient features to predict hospitalization risk and weeks to heal.

different wound parameters. In the second stage, we use these predictions along with patient features to predict (1) whether hospitalization would be needed for the wound, and (2) weeks to heal (or wound progression).

3.1. Stage 1: Wound Feature Prediction from Image

The 14 attributes are described in detail in Table 1. While multiple deep CNN model architectures are available like VGGNet, Inception, Xception, ResNet, DenseNet, etc., we chose to use Xception architecture [29] based on cross-validation accuracy and model training speed. In practice, a deep CNN model like Xception trained from scratch with randomly initialized weights using a few tens of thousands of data points may overfit. Hence, we use ImageNet pre-trained Xception models.

Transfer learning: We employed pre-trained Xception architecture weights [29] trained on more than a million images from the ImageNet database. A pre-trained network is 71 layers deep and can classify images into 1000 object categories. We hope that this pre-training helps the network learn rich feature representations for a wide range of images. The network has an image input size of 299 pixels x 299 pixels. The pre-trained model is fine-tuned with each of the 14 wound attributes as targets.

We tried multiple methods to train the Xception architecture for better prediction accuracy. We first trained individual models for each of the 14 attributes in two ways: (i) fine-tuning all the layers of the model and (ii) gradual unfreezing, i.e., fine-tuning the last ten layers (freezing the rest) and thereafter increasing the layers for fine-tuning by a factor of 10 in every iteration. In each iteration, we fine-tuned the model with a very low learning rate, such that the current training does not drastically change the feature rep-

resentations learned from Imagenet dataset. We observed that the complete fine-tuning model gave the best results. We trained the models using Adam optimizer [30] and categorical cross-entropy loss. We used a batch size of 128 images, and early stopping to halt the training when the model accuracy did not improve for a few epochs.

Multi-task learning (MTL): Each of the 14 CNN models is fine-tuned using task-specific labeled training data in single-task setting initially. We performed the pairwise correlation analysis between these 14 wound variables and then used the multi-task learning setup for joint learning of highly correlated variables. We jointly learned the following combinations: (1) wound type, wound location, wound stage, (2) wound type, wound margin, (3) wound type, joint necrosis exposed, ligament necrosis exposed, adipose necrosis exposed, muscle necrosis exposed, bone necrosis exposed, (4) wound type, red granulation, adherent yellow slough, (5) wound type, exudate. However, for the following attributes, we found better results using a single-task setup rather than MTL and hence use the single task classifier results: wound area, wound volume, wound type, and wound location. Thus, in the MTL set up, we learn a total of 9 Xception CNN models. As shown in Table 2, we also tried a single unified MTL model but accuracy of certain tasks was very low possibly due to class imbalance and lack of strong correlations across some task pairs.

3.2. Stage 2: Heal/Hospitalization Classification and Weeks to Heal Prediction

The 14 predictions from the first stage are combined with 5 patient features (age, BMI (body mass index), ethnicity, pulse rate, location), which cannot be predicted using wound images. The 19 features are used to train (1) a binary hospitalization risk (or heal/hospitalization) LGBM (Light Gradient Boosted Machines) classifier [31], and (2) weeks-to-heal LGBM regression model. LGBM is a gradient boosting framework that uses a tree-based learning algorithm to grow trees leaf-wise rather than level-wise. To identify the best way of encoding the CNN-predicted wound attributes, we tried three methods: (i) we used feature representation from the penultimate layer of each of Xception models (14×2048 dimensions), decreased dimensions for each attribute to 5 (14×5 features) using Principal Component Analysis (PCA), and then used them along with patient attributes (5 features) to train the model. (ii) We combined predictions from Xception models, and their probabilities (14×2 features) along with patient attributes to train the model. (iii) Only the 14 predictions from Xception models were combined with 5 remaining variables to train the model. We observed that the performance of the model trained with Xception model predictions and their probabilities, i.e., the (ii) method, is better than the rest of the methods. Hence, we performed all experiments using the (ii) method with (14×2+5=) 33 features.

Imbalanced dataset handling: The heal/hospitalization data has class imbalance issue, as detailed in Section 4. We observe that the majority class ("heal") has ∼10x more instances compared to the "hospitalization" class. To circumvent this problem, we tried methods such as data augmentation, SMOTE (Synthetic Minority Oversampling Technique) [32] based over-sampling, and under-sampling techniques.

Evolutionary algorithm with Stacked LGBM: Traditionally boosting using LGBM has benefited many prediction tasks. However, in the hope of better accuracy, we experiment with a 2-level stacking model, inspired by the architecture of [33], where individual learners are LGBMs themselves. Unfortunately, stacked LGBMs have too many hyper-parameters to tune manually. Motivated by the recent success of Evolutionary algorithms [34], we used genetic algorithms to discover a promising hyper-parameter configuration. Specifically, our genetic algorithm with stacked LGBM used the following features as chromosomes for hyper-parameter tuning: (i) number of leaves (ii) maximum depth (iii) learning rate (iv) boosting type (v) minimum child samples (vi) maximum bin (vii) the number of iterations. We use the same parameter configuration for each LGBM in the stacked-LGBM model.

4. Experiment Results

4.1. Dataset

Our wound image dataset is an accumulation of 4 years (Jun 2015–Mar 2019) of patients' wound care data captured by a Wound care organization. The images have been captured using typical smartphone cameras of various brands. Data collection was carefully done by following the survival model conditions [35] to ensure that we cover the "Patient Demographics details", "Procedures", "Medications", and "Laboratory/Diagnosis of Wound condition". Our dataset contains 125711 images obtained from 11632 patients, corresponding to five wound types: "Diabetic Ulcer", "Pressure Ulcer", "Surgical Wound", "Trauma Wound", and "Venous Ulcer". Along with the images, the dataset also contains 14 wound-related attributes and 5 patient attributes, all of which are gathered by clinicians during in-house patient treatments. The clinician's wound assessment is further reviewed by a team of care monitors led by medical directors to ensure that the clinician's diagnosis is correct. Among the 14 predicted wound variables, as detailed in Table 1, six are binary, two are numeric, and remaining are categorical features. The average age of patients is 73.63 years and the average BMI is 29.66. The wound care organization is HIPAA compliant (Health Insurance Portability and Accountability Act) and got approval from its Institutional Review Board (IRB) before using this data for analysis. We

Wound attribute	Baseline (similar to [5])		Single Task CNN		Single Multi-Task CNN		Multi-Task CNN	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Wound/Ulcer Type	0.46	0.48	**0.80**	**0.81**	0.64	0.70	0.66	0.80
Wound Location	0.49	0.48	**0.87**	**0.87**	0.66	0.58	0.66	0.65
Wound Stage	0.58	0.54	0.63	0.65	0.50	0.55	**0.78**	**0.71**
Wound Margin	0.46	0.46	0.54	0.58	0.29	0.35	**0.60**	**0.63**
Joint Necrosis Exposed	0.55	0.55	0.79	0.83	0.67	0.71	**0.77**	**0.84**
Ligament Necrosis Exposed	0.71	0.61	0.77	0.70	0.67	0.69	**0.74**	**0.81**
Adipose Necrosis Exposed	0.64	0.64	0.69	0.83	0.66	0.69	**0.71**	**0.83**
Muscle Necrosis Exposed	0.66	0.64	0.69	0.83	0.66	0.69	**0.72**	**0.83**
Exudate	0.60	0.60	0.65	0.68	0.59	0.61	**0.71**	**0.73**
Red Granulation	0.44	0.44	0.65	0.68	0.44	0.61	**0.66**	**0.68**
Bone Necrosis Exposed	0.67	0.74	0.76	0.74	0.69	0.72	**0.74**	**0.81**
Adherent Yellow Slough	0.42	0.44	0.62	0.64	0.51	0.54	**0.65**	**0.67**

Table 2. Stage 1 Results: Accuracy for wound attribute prediction using Xception CNN classifiers. The precision and recall for feature engineering based baseline model, single-task, single multi-task and combination of nine multi-task experiments are listed in the table.

have released a small subset of the dataset and code here[2].

We pre-processed the dataset to handle issues like label noise, occlusion, illumination, imbalanced data and deformation. Pre-processing details are mentioned in Section 1 of the supplementary material.

4.2. Results

To reduce clinician workload, our first goal is to predict the 14 wound-related attributes automatically from wound images using Xception models. Our second goal is to use this set of (14 predicted+5 patient) features to build the final heal/hospitalization and weeks to heal prediction models. We use 70:10:20 split for the train:validation:test for all our experiments. Experiments were done on a machine with 4 Tesla V100-SXM2-32GB GPUs.

Baseline Method: Our baseline method is based on traditional image processing features (similar to [5]). For each wound image, we considered the color histogram, canny edge detector features, Hough transform features, and Histogram of Gradients (HOG) features to train the model. Using these features, we build an LGBM classifier for prediction of the 14 image attributes.

Oracle Method: For stage 2, we used the ground truth values for the 14 image features and the 5 patient features. We refer to this as the oracle method. In the remaining, we present comparisons of stage 1 and 2 results obtained using the baseline, oracle, and the proposed HealTech system.

4.2.1 Stage 1 Results

As part of stage 1, we predict the 14 wound attributes using baseline as well as HealTech methods. Table 2 shows the 5-fold cross-validation precision and recall values for each of the 14 wound attributes obtained using baseline features, single-task as well as multi-task Xception CNN models. We can observe that the proposed HealTech based single-task and multi-task CNN results are better compared to the baseline approach. Further, the following variables see significant gains in accuracy when multi-task learning is used: Stage, Margin, Ligament Necrosis Exposed, Adipose Necrosis Exposed, Muscle Necrosis Exposed, Exudate, Red Granulation, Bone Necrosis Exposed. We also tried a single unified MTL CNN model but this usually performs worse than the combination of 9 multi-task CNNs.

Table 3 shows model accuracy for different ulcer subtypes. Of the five ulcer types, prediction accuracy is best for pressure ulcers and worst for trauma wounds. The classifier was confused between the (diabetic, pressure and surgical), and (trauma and venous) categories. As per the clinicians' diagnosis, a surgical wound at a later stage can lead to a diabetic ulcer. Similarly, a trauma wound can lead to a venous ulcer. We attribute the lower performance of our model on some of the ulcer types to the visual similarity of these ulcer types. Table 4 describes the detailed accuracy of various ulcer locations. Accuracy is best for sacral and worst for great toe. The classifier was confused between the (great toe, heel and foot), and (ankle, heel and foot) pairs. Table 5 describes the wound stage prediction results for each individual stage. The best results are for full-thickness while the worst is for unstageable. The term unstageable is used when clinicians are confused about the correct wound stage. Thus, the model behaves as expected. The classifier was most confused between the full thickness, partial thickness and Stage-3 classes. To measure the performance of continuous-valued output variables "Area", and "Volume", we use Mean Absolute Error (MAE) as metric. Here, we applied the scaling on the target values, since values are of different range. We found MAEs for area and volume as 1.16 and 1.28, respectively. The minimum area of the wound is 0.0, the avg area is 7.74, and the max area

[2] https://drive.google.com/file/d/1m2yStm6qlc2s_9zBNh_mUtgb5JYRBLOY/view?usp=sharing

Ulcer Type	P	R	F1-score
Diabetic Ulcer	0.79	0.84	0.81
Pressure Ulcer	0.87	0.89	0.88
Surgical Wound	0.76	0.65	0.70
Trauma Wound	0.65	0.56	0.61
Venous Ulcer	0.82	0.89	0.85

Table 3. Ulcer Type Results

Ulcer Location	P	R	F1-score
Lower Leg	0.88	0.90	0.89
Sacral	0.99	0.98	0.98
Foot	0.83	0.83	0.83
Heel	0.84	0.88	0.86
Ankle	0.73	0.77	0.75
Great Toe	0.67	0.55	0.61

Table 4. Wound Location

Stage	P	R	F1-score
Full Thickness	0.83	0.90	0.86
Stage-2	0.68	0.66	0.67
Stage-3	0.70	0.69	0.69
Stage-4	0.64	0.66	0.65
Unstageable	0.76	0.61	0.67
Partial Thickness	0.62	0.60	0.61

Table 5. Wound Stage

	1 Xception-CNN						14 Xception-CNNs					
Sampling Method→	None		Over		Under		None		Over		Under	
Class↓	P	R	P	R	P	R	P	R	P	R	P	R
Heal	0.73	0.97	0.82	0.81	0.85	0.67	0.80	0.93	**0.84**	**0.83**	0.86	0.70
Hospitalization	0.64	0.13	0.49	0.51	0.41	0.66	0.64	0.37	**0.53**	**0.54**	0.45	0.69

Table 6. Stage 2 Results: Heal/Hospitalization accuracy comparison for various sampling methods using 1/14 Xception-CNNs.

		Baseline						Oracle						Xception CNN (HealTech)					
Sampling Method→		None		Over		Under		None		Over		Under		None		Over		Under	
Feature set↓	Class↓	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
All	Heal	0.98	0.98	0.98	0.98	0.98	0.94	0.98	0.99	0.98	0.99	0.99	0.94	0.98	0.99	**0.99**	**1**	1	0.94
All	Hospitalization	0.80	0.73	0.90	0.76	0.51	0.87	0.88	0.75	0.87	0.79	0.52	0.87	0.89	0.77	**0.92**	**0.8**	0.54	0.9
Image	Heal	0.97	0.97	0.98	0.98	0.97	0.85	0.98	0.98	0.98	0.98	0.98	0.87	0.98	0.98	**0.98**	**0.98**	0.98	0.87
Image	Hospitalization	0.82	0.67	0.73	0.67	0.35	0.80	0.8	0.66	0.76	0.75	0.33	0.82	0.79	0.68	**0.77**	**0.76**	0.35	0.83

Table 7. Stage 2 Results: Heal/Hospitalization accuracy comparison for LGBM method using different sampling methods with Baseline/Oracle/Xception-CNN (HealTech).

		Baseline				Oracle				Xception CNN (HealTech)			
Sampling Method→		Over		Under		Over		Under		Over		Under	
Feature set↓	Class↓	P	R	P	R	P	R	P	R	P	R	P	R
All	Heal	0.98	0.99	0.99	0.93	0.98	0.99	0.99	0.94	**0.99**	**0.99**	1	0.94
All	Hospitalization	0.86	0.78	0.49	0.86	0.89	0.79	0.52	0.91	**0.92**	**0.83**	0.55	0.95
Image	Heal	0.97	0.98	0.98	0.78	0.98	0.99	0.99	0.79	**0.98**	**0.99**	0.99	0.8
Image	Hospitalization	0.77	0.65	0.23	0.84	0.78	0.74	0.24	0.86	**0.78**	**0.76**	0.24	0.85

Table 8. Stage 2 Results: Heal/Hospitalization accuracy comparison for GAT Stacked LGBM method using different sampling methods with Baseline/Oracle/Xception-CNN (HealTech).

of the wound is 14.25 on a log scale. Similarly, the minimum volume of the wound is 0.0, avg volume is 8.95, and the max area of the wound is 16.86 on a log scale.

Figure 4 shows gradient heatmaps of different wound variables as predicted by the HealTech's Xception CNN models. The results showcase that GradMap is correctly highlighting the wounded part but not the surroundings. This encourages us to believe that CNNs are performing predictions based on a good semantic understanding of the wound image.

Experiments on Medetec Dataset [26]: The dataset contains 367 wound images belonging to the five major ulcer types. To evaluate our model performance, we fine-tune our Xception CNN models on the publicly available Medetec dataset. Also, we train several baseline models, as proposed earlier in the literature. Previous models [27, 36] report results on Medetec dataset only on these three binary attributes: necrosis, granulation, and slough level. Hence, we show these comparative results in Table 9. From the Table 9, we observe that our Xception CNN based HealTech model outperforms the state-of-the-art models on the Medetec dataset. Also, note that the baseline models [27, 36] focused on first segmenting the wound patch from the image and then making the prediction. Unlike these methods, HealTech directly applies the Xception models on the Medetec dataset of images thereby avoiding the expensive segmentation pre-processing step.

Class	Bayesian [27]	SVM [27]	RF [36]	HealTech
Necrosis	0.79	0.80	0.82	**0.89**
Slough	0.78	**0.91**	0.85	0.88
Granulation	0.87	0.88	0.87	**0.92**

Table 9. Stage 1 Results: Binary classification Micro-Averaged Accuracy on the Medetec Dataset.

Figure 3. HealTech Results: (i) Top row showcases sample wound images for which our model correctly predicted the heal vs hospitalization class, (ii) bottom row showcases sample wound images along with the actual and predicted weeks to heal.

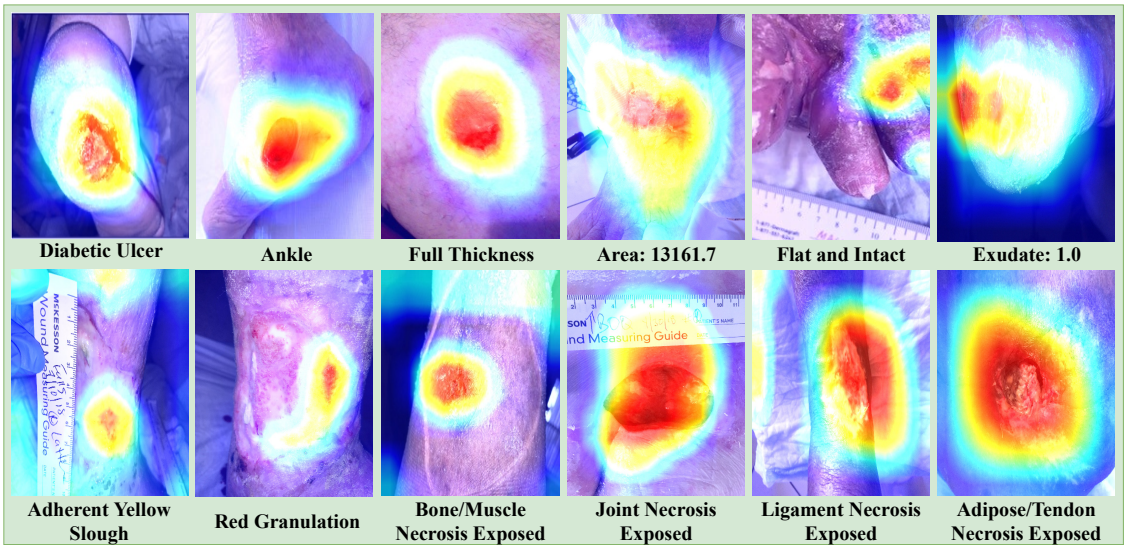

Figure 4. GradMaps of different wound variables predicted by Xception models trained by respective wound variables. We can observe that the model accurately predicted the location of the wound, despite occlusion and other limitations.

4.2.2 Stage 2 Results

Heal/Hospitalization Classification: First, we train an Xception-CNN heal/ hospitalization classifier using the wound image alone. Next, we train 14 different Xception-CNN models to predict each image attribute separately. The second last layer output from each of the 14 models is concatenated and connected to the output softmax layer for heal/ hospitalization prediction. Table 6 shows the results obtained using the one-CNN versus the 14-CNNs models.

Further, we train a heal/ hospitalization classifier using the golden set of 19 features (including 14 wound image attributes), which were provided by clinicians. We refer to this as the oracle classifier. However, to save on labeling bandwidth of the clinician, we predict these 14 wound image attributes automatically, and then build a heal/hospitalization classifier using 5 patient features + 14 predicted image attributes. Overall, we experiment with three methods: baseline, oracle, and Xception-CNN based classifiers (HealTech), across multiple methods (no sampling, under-sampling, and over-sampling) to create our classifier using: (1) LGBM and (2) (Genetic Algorithms Tuned) GAT-Stacked LGBM methods. Of a total of 20376 samples, 18932 examples are of heal category, and 1444 examples belong to the hospitalization class.

The experiments in Table 7 describe the performance

	Baseline			Oracle			Xception CNN (HealTech)		
Feature set↓	LR	LGBM	GS-LGBM	LR	LGBM	GS-LGBM	LR	LGBM	GS-LGBM
All	4.4	4.1	4.0	4.0	3.4	**3.1**	4.0	3.6	3.3
Image	4.4	4.3	4.1	4.2	3.8	**3.3**	4.3	4.0	3.4

Table 10. Stage 2 Results: Weeks to Heal prediction comparison for LGBM (HealTech) method and the baseline Linear Regression on MAE. LR=Linear Regression. GS-LGBM=GAT Stacked LGBM

of LGBM models, and experiments in Table 8 summarize the performance of GAT Stacked LGBM method in predicting patients heal/ hospitalization risk. We experimented with two different features sets to assess our performance of Baseline, Oracle, and Xception CNN (HealTech) models: (i) All features: which include wound attributes and patient features. (ii) Image features: contains just the wound image attributes. First, we observe that results in Tables 7 and 8 are much better compared to simple 1/14 Xception-CNN models (Table 6). Next, better results are obtained when all the features are utilized for training a GAT Stacked LGBM model rather than just the image features. Surprisingly, as Table 7 shows, the baseline also performs well for the heal class, but HealTech outperforms the baseline significantly on F1 for the hospitalization class. As shown in Table 8, we observed the best precision and recall of 0.92 and 0.83, respectively for the hospitalization class, when trained with over-sampling methods with GAT Stacked LGBM. We found these results to be statistically significantly better compared to the baseline. The under-sampling experiments were conducted to increase the recall of the hospitalization risk patients. We also observe that the under-sampling based results were better when we use GAT Stacked LGBM model, with 0.95 recall for the hospitalization class, although the precision is relatively bad. Overall, we observe that our proposed GAT Stacked LGBM models with over-sampling provide the best results.

Weeks to Heal Prediction: For the weeks to heal prediction our labeled data follows a power law distribution (refer Section 4 of the supplementary material). To perform the weeks to heal prediction, we use the same features as used in the Heal/Hospitalization classification model. However, here the target variable is the weeks to heal for a particular wound image. The results in Table 10 illustrate the performance of the LGBM model in comparison with the baseline linear regression model. Again, we use either the set of all features or just the image features, to assess our weeks to heal model performance. To measure the model performance, we use mean absolute error (MAE) as the metric. The minimum number of weeks to heal is one, the maximum is 30 weeks, and the average is 14 weeks in our dataset. The LGBM model achieves an MAE of 4.0 weeks if we consider only image features, and 3.6 MAE on all features. Our GS-LGBM model is even better with an MAE of 3.3 using all features, and an MAE of 3.4 using image features only. Note that the results obtained using the wound attributes predicted by Xception CNN (HealTech) models are very similar to those obtained using the Oracle method (i.e., ground truth labels for the 14 image attributes).

Feature Importance Analysis: We observed that age and BMI are the most important patient attributes. Wound area was the best predictor for the heal/ hospitalization classifier, which is expected since the wound area intuitively correlates with the seriousness of the wound. Similarly, wound location, area, and type are important image attributes for weeks to heal prediction. More details are in Section 2 of the supplementary material.

Case Studies: Top row of Figure 3 shows sample wound images for which our model correctly predicted the heal vs. hospitalization class, while the bottom row shows sample wound images along with the actual and predicted weeks to heal. The results clearly look very intuitive.

Error Analysis: We analyzed the error cases in detail for both the wound assessment tasks. Among patients where the actual class was hospitalization but predicted class was heal, we found the following discriminative patterns: 65% cases had class-4 red granulation, 81% had class-2 adherent yellow slough. Among patients where the actual class was heal but predicted class was hospitalization, we found the following discriminating patterns: 87% cases had joint necrosis exposed. For the weeks to heal prediction model, we observed that most cases have very small error, while some cases have large errors (refer Section 3 in the supplementary material).

5. Conclusion

We proposed two interesting wound assessment tasks: hospitalization risk prediction and weeks to heal prediction. Our HealTech system operates in two stages. In the first stage, the wound image is analyzed by Xception CNN models to predict 14 critical wound attributes. In the second stage, these attributes are leveraged to perform the two wound assessment predictions. We are the first to perform extensive experiments on a large dataset for wound analysis. Our experiments show that HealTech leads to good accuracy values, and therefore can be practically deployed. We believe that our solution can help clinicians make better decisions regarding whether a patient needs to be sent to an acute wound care facility in order to heal. We hope our work can support clinicians in the diagnosis process while reducing the time spent on assessing each wound.

References

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

[2] H. Wannous, S. Treuillet, and Y. Lucas, "Supervised tissue classification from color images for a complete wound assessment tool," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2007.

[3] H. Wannous, S. Treuillet, and Y. Lucas, "Robust tissue classification for reproducible wound assessment in telemedicine environments," *Journal of Electronic Imaging*, 2010.

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

[5] H. Wannous, Y. Lucas, and S. Treuillet, "Enhanced assessment of the wound-healing process by accurate multiview tissue classification," *IEEE transactions on Medical Imaging*, 2010.

[6] M. Kolesnik and A. Fexa, "Multi-dimensional color histograms for segmentation of wounds in images," in *International Conference Image Analysis and Recognition*, Springer, 2005.

[7] M. Kolesnik and A. Fexa, "How robust is the svm wound segmentation?," in *Proceedings of the 7th Nordic Signal Processing Symposium-NORSIG 2006*, IEEE, 2006.

[8] E. S. Papazoglou, L. Zubkov, X. Mao, M. Neidrauer, N. Rannou, and M. S. Weingarten, "Image analysis of chronic wounds for determining the surface area," *Wound repair and regeneration*, 2010.

[9] F. J. Veredas, H. Mesa, and L. Morente, "Efficient detection of wound-bed and peripheral skin with statistical colour models," *Medical & biological engineering & computing*, 2015.

[10] C. P. Loizou, T. Kasparis, O. Mitsi, and M. Polyviou, "Evaluation of wound healing process based on texture analysis," in *2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE)*, pp. 709–714, IEEE, 2012.

[11] H. Greenspan, B. Van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.

[12] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[13] M. Lalonde, L. Gagnon, M.-C. Boucher, *et al.*, "Automatic visual quality assessment in optical fundus images," in *Proceedings of vision interface*, vol. 32, pp. 259–264, Ottawa, 2001.

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

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[16] 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," *Computational intelligence and neuroscience*, 2018.

[17] X. Liu, C. Wang, F. Li, X. Zhao, E. Zhu, and Y. Peng, "A framework of wound segmentation based on deep convolutional networks," in *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, IEEE, 2017.

[18] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, I. Eric, and C. Chang, "Deep learning of feature representation with multiple instance learning for medical image analysis," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 1626–1630, IEEE, 2014.

[19] S. Zahia, D. Sierra-Sosa, B. Garcia-Zapirain, and A. Elmaghraby, "Tissue classification and segmentation of pressure injuries using convolutional neural networks," *Computer methods and programs in biomedicine*, 2018.

[20] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[21] B. D. de Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Išgum, "End-to-end unsupervised deformable image registration with a convolutional neural network," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 204–212, Springer, 2017.

[22] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll, "Learning a variational network for reconstruction of accelerated mri data," *Magnetic resonance in medicine*, vol. 79, no. 6, pp. 3055–3071, 2018.

[23] B. García-Zapirain, M. Elmogy, A. El-Baz, and A. S. Elmaghraby, "Classification of pressure ulcer tissues with 3d convolutional neural network," *Medical & biological engineering & computing*, vol. 56, no. 12, pp. 2245–2258, 2018.

[24] A. Khalil, M. Elmogy, M. Ghazal, C. Burns, and A. El-Baz, "Chronic wound healing assessment system based on different features modalities and non-negative matrix factorization (nmf) feature reduction," *IEEE Access*, vol. 7, pp. 80110–80121, 2019.

[25] M. Elmogy, B. García-Zapirain, C. Burns, A. Elmaghraby, and A. Ei-Baz, "Tissues classification for pressure ulcer images based on 3d convolutional neural network," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 3139–3143, IEEE, 2018.

[26] S. Thomas, "Medetec wound database," 2017.

[27] R. Mukherjee, D. D. Manohar, D. K. Das, A. Achar, A. Mitra, and C. Chakraborty, "Automated tissue classification framework for reproducible chronic wound assessment," *BioMed research international*, 2014.

[28] M. Goyal, N. D. Reeves, A. K. Davison, S. Rajbhandari, J. Spragg, and M. H. Yap, "Dfunet: Convolutional neural networks for diabetic foot ulcer classification," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2018.

[29] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[31] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in neural information processing systems*, 2017.

[32] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[33] Z.-H. Zhou and J. Feng, "Deep forest: Towards an alternative to deep neural networks," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, (Melbourne, Australia), AAAI Press, 2017.

[34] E. Real, S. Moore, A. Selle, S. Saxena, Y. L. Suematsu, J. Tan, Q. V. Le, and A. Kurakin, "Large-scale evolution of image classifiers," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, JMLR. org, 2017.

[35] P. Wang, Y. Li, and C. K. Reddy, "Machine learning for survival analysis: A survey," *ACM Computing Surveys (CSUR)*, 2019.

[36] F. J. Veredas, R. M. Luque-Baena, F. J. Martín-Santos, J. C. Morilla-Herrera, and L. Morente, "Wound image evaluation with machine learning," *Neurocomputing*, 2015.