BURNED: Towards Efficient and Accurate Burn Prognosis Using Deep Learning

Orion Despo, Serena Yeung, Jeffrey Jopling, Brian Pridgen, Cliff Sheckter, Sara Silberstein, Li Fei-Fei, Arnold Milstein

Stanford, Stanford, CERC, CERC, CERC, CERC, Stanford, CERC

Abstract

Early excision of burns saves lives. Like most treatments, this relies on having accurate early-stage burn depth diagnosis. Unfortunately, even considering the experts, there exists a lack of accuracy among the burn community in terms of early diagnosis. Using state of the art deep learning principles, this project aims to improve upon this. First, we created the largest dataset to date of segmented and labeled burns. Second, with minimal training, we show that we can accurately discriminate burnt skin from the rest of an image, paving the way for the calculation of clinically important metrics, such as the percent of the body which is burned (TBSA). Furthermore, we show there is large potential to do even better. We then show that provided we make the necessary data and model adjustments, we can extend this to accurately segment and classify different burn depths. In total, this project brings us one step closer to scaling expert level care to millions of burn victims worldwide.[1]

1. Introduction

Machine learning is transforming a variety of disciplines. By identifying complex patterns from vast amounts of data, it is able to provide much more accurate insights at unprecedented scale. Simultaneously, we see U.S. healthcare starting to place much greater premiums on efficiency and accessibility. As such, we see the field starting to embrace machine learning techniques, shifting it towards prevention methodologies and acting to un-silo medical expertise [R15]. Notably, we've also seen an uptick in applying computer vision principles. For example, Stanford researchers have recently shown deep learning can be used to obtain expert-level identification of skin cancer from images [A17].

Along the lines of early detection, early excision of burns has been shown to improve patient outcomes and reduce costs [N70] [N89] [SMC06]. However, with increasing emphasis on early excision comes an increased need for early diagnosis of burn depth. Clinical diagnosis by visual inspection and physical exam remains the most common means of assessing burn depth in the United States [R14]. This visual assessment attempts to identify the burn depth and the percent of the body that is burned (TBSA). Unfortunately, accuracy of burn depth diagnosis by burn specialists remains at only 40-70% in the first several days following a burn [AAO01] [H09] [M93] [M84], which is the period in which accurate diagnosis is needed in order to proceed with early excision. Moreover, those who are not burn specialists are much less accurate at burn depth assessment [B13]. Given that burn patients are usually diagnosed by non-specialized doctors [B13] and misdiagnosis can lead either to unnecessary excision or delayed excision (resulting in scarring, chronic health complications, etc.), simply by virtue of geographic location, many people are not experiencing the care they need.

Current technologies to sidestep visual inspection, such as laser Doppler Imaging, have low adoption and are also difficult to use and costly [R14]. Ubiquitous smartphones with cameras would overcome the challenges of cost, availability, and difficulty of use. Furthermore, remote assessment of low quality digital images in burns and wounds have been shown to have a high degree of inter-rater reliability [L99] [X06] [K07]. However, remote assessment is hard to scale and diagnosis is still left to the whims of a single (or handful) or doctors.

Given this, there appears to be an exciting opportunity to apply machine learning principles. With this study, we propose to develop state-of-the-art deep learning-based algorithms using digital images for automatic assessment and diagnosis of burn depth. Particularly, given an RGB 2d input image, predict the burn severity on a per-pixel basis, thus categorizing a burns severity and spatial outline. This system could be used by non-specialists to aid in early diagnosis and triage of burns or by specialists to aid in management decisions to promote earlier excision. Ultimately, this will bring us closer to scaling expert level care to millions of burn victims worldwide.

[1] Only Orion is enrolled in CS231N

Figure 1: Given an input image, we want to predict each burn's severity and spatial outline. Each shaded color represents a different burn depth.

2. Related Work

There have been previous attempts exploring the development of image recognition platforms for diagnosis of burns. [P14] combines optical coherence tomography and pulse speckle imaging to achieve an ROC of .87 in classifying burn depth of manually created pig skin burns. The three burn depths were superficial, partial thickness, and full thickness. [K05] uses reflective spectrophotometer data combined with a Radial Basis Function neural network to achieve 86% classification accuracy in predicting whether the healing times of burns will be within 14 days or not. Neither of these studies use raw digital images, instead relying on more expensive/complicated technology. Furthermore, they go through lengthy processes to manually specify features, use datasets of just 68 and 41 images respectively, and do not segment.

Closer to the ideal system, a group at the University of Seville uses raw digital images as the inputs to their algorithms. [B05] first segment the burns then feed the proposed burnt skin into a Fuzzy-ARTMAP neural network to classify whether the burn depth is superficial, deep, or full thickness. With a test set of 62 images, they achieve a classification accuracy of 82%. It is important to note that the segmentation and classification were not done end-to-end. In 2013, the same group attempted two different classification tasks using a different feature representation strategy i.e. they tried to replicate a plastic surgeon's classification process using a pyschophysical experiment. In the first, they try to classify burn depth (superficial, deep, full thickness) using a KNN. They achieve a 66.2% success rate. When trying to predict whether a burn needed a graft or not, they achieve 83.8%. In both, they use a test set of 74 images [B13].

In total, we see past work has relied upon complicated feature engineering techniques and datasets of less than 100 images. Both of which are clearly hard to scale/generalize with.

Fortunately, deep learning is a class of machine learning algorithms that uses data to automatically learn classifier features instead of relying on humans to hand-design them, thus decreasing the amount of bias that humans inject into these systems. As a result, deep learning methods have recently shown superhuman performance on a variety of image and pattern recognition tasks, such as classification of over 200 different dog breeds [YYH15] [O15] [K15]. These principles are also starting to be applied in the medical domain. Researchers have recently used deep convolutional nets (particularly GoogleNet) and a training set of over 100,000 images to achieve skin cancer identification accuracy on par with dermatology experts [A17]. Given skin cancer is also commonly diagnosed visually, this represents the potential to apply a similar methodology to burn wounds.

In total, our method will advance upon previous work in three main areas: 1) creation of the largest dataset of labeled and segmented burns to date 2) adaptation of state-of-the-art deep learning techniques that have been shown to out-perform humans 3) end-to-end segmentation and classification i.e. predicting both the spatial outline and burn depth in a single network. Not only is segmentation technically more challenging, but it allows for the calculation of TBSA, which allows for more accurate diagnoses. Thus, end-to-end classification/segmentation has much more clinical utility.

3. Methods

3.1. Data Collection

Figure 2: Examples of images in the novel BURNED dataset.

To accomplish the stated task, a novel dataset called BURNED was created. First, we obtained 749 images from

the Santa Clara Valley Medical Center and manually curated an additional 180 images from Google Search. Next, we adapted an online annotation tool [BC08] to make it easy for plastic surgeons to outline and label various burns on a given image. This required making immense UI modifications to streamline the process for the plastic surgeons as well as adding in a secure login and backend database. Given the data was considered medically risky, we had to deploy the tool on Google Cloud. The collection process consisted of two steps. First, a set of 3 plastic surgeons went through and outlined the burns on each image. Second, each image was randomly assigned 3 plastic surgeon labelers (out of 6). The available labeling choices were: superficial (S), superficial/deep partial thickness, full thickness (FT), and undebrided (U). Though we allowed labelers to differentiate between superficial partial thickness and deep partial thickness, we merged them into partial thickness (PT) for the analysis due to the size of the dataset. Furthermore, since the dataset was still being prepared, the analysis only uses 656 images, which constitutes 1351 masks/burns. Even then, this set constitutes the largest labeled and segmented burn dataset.

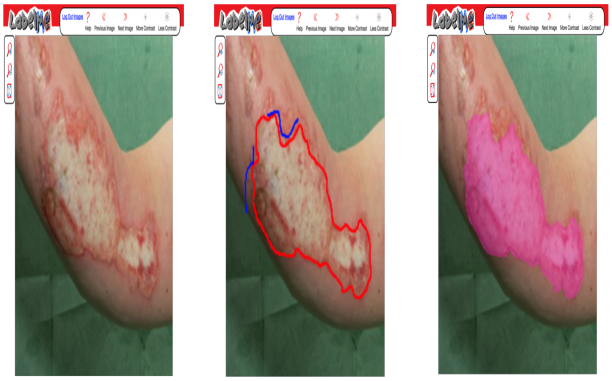

Figure 3: Example of the annotation interface. The second image represents a surgeon outlining the burn. The third image shows the resulting mask.

3.2. Technical Approach

The network architecture is based off of a fully convolutional network [JET15]. There are two key aspects of the model. First, this adapts traditional deep convolutional nets (AlexNet, VGG, GoogleNet), which have a fully connected layer at the end, to accept arbitrary inputs and produce corresponding spatial outputs. This is important as for a given input image of dimension $HeightxWidthx3$, we want to produce an output of shape $HeightxWidthxClasses$ to represent the segmentation mask. Second, nets like VGG are traditionally used to capture and predict coarse, high level information, whereas semantic segmentation requires

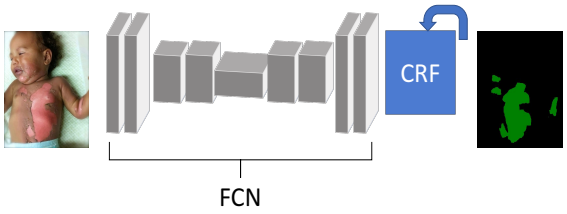

Figure 4: The network architecture for this project. The input is the raw image, which is then passed to an FCN with a CRF layer attached. The resulting output is a per-pixel mask.

simultaneously capturing fine-grained, pixel level information. As such, after the high level information is obtained (such as whether the image contains a superficial burn), FCN-8 (based on VGG-16) upsamples back to the original image dimensions and combines predictions at the final layer with predictions at earlier layers. This enables the model to capture the pixel-level, fine-grained information i.e. which pixels correspond to the superficial burn.

We next attach a conditional random field formulated as an RNN cell onto the FCN [S15]. Through essentially adding smoothness constraints between neighboring pixel labels, this layer acts to further fine-grain some of the outputs of the FCN making predictions more consistent and improving common semantic segmentation metrics. In other words, pixels near and similar to each other should have the same class assignment.

The model was pre-trained on the PASCAL VOC 2012 dataset.

3.3. Metrics

Pixel Accuracy (PA)

$$\frac{1}{n_{cl}} \sum_{i=1}^{n_{cl}} \frac{n_{ii}}{t_i}$$

Mean IOU

$$\frac{1}{n_{cl}} \sum_{i=1}^{n_{cl}} \frac{n_{ii}}{t_i + \sum_{j=1}^{n_{cl}} n_{ji} - n_{ii}}$$

Figure 5: n_{cl} represents the number of classes (not including the background class). t_i represents the total number of pixels for class i. n_{ij} is the number of pixels of class i predicted to be class j. These are common metrics for semantic segmentation [S15].

We use two main metrics to evaluate performance of the system. Pixel Accuracy (PA) measures how much of the burn we recovered in our prediction. Mean intersection-

over-union (IOU) can be thought as an F1 score, penalizing us for over-predicting. The combination of these two methods allow us to quantify whether we over or under predict a certain class. For example, if the pixel accuracy is greater than IOU, this means we are over-predicting a class.

Figure 6: An example of the difference between the two metrics. The left image is the ground truth and the right is predicted. PA is close to 1, but IOU is much less than 1.

4. Results

For all results below, we used a ADAM optimizer with a learning rate of $5e^{-}5$. With SGD as the optimizer, a learning rate several magnitudes smaller had to be used, causing the training process to take substantially larger. If we used a larger learning rate with SGD, we ran into exploding gradient problems. Even though adding in gradient clipping helped alleviate this, the results had much more variance, causing us to choose ADAM. Each image was resized to 250x250 and a batch size of 1 was used. Though [S15] uses 500x500, we chose 250x50 as it resulted in 4x shorter training times with no substantial drop in accuracy. The batch size of 1 was recommended by [S15]. We used a 60/20/20 training, validation, test split and ran each model for only 15 epochs (due to resource constraints). Code was adapted from [Ker16] [Gro17] [Jun17].

4.1. Burn/No Burn

We first wanted to see if we could discriminate between burnt skin and the rest of the image. We also used this as an opportunity to experiment with model architectures before we attempted to differentiate between the different burn depths.

Overall, we are able to accomplish our task exceedingly well (given the size of the dataset and minimal training time). First, we see the FCN with conditional random field outperforms the plain FCN (though only slightly). Because of this, we chose to use this architecture for the rest

	PA	IOU
FCN - No CRF	.82	.54
FCN - CRF	.85	.56
FCN - CRF, Aug	.85	.67
Pascals	N/A	.75

Figure 7: We see the breakdown in metrics for the test set. The FCN with CRF layer and data augmentation clearly does the best. The pascals represents the models performance on the PASCAL VOC 2012 dataset, one of the gold standards for semantic segmentation [S15]. Though the datasets comparable, this is meant to show we should not expect to get an IOU of 1.

of the experiments. Second, we are erring on the side of over-predicting burnt skin. Third, adding in data augmentation substantially reduces the amount of over-predicting i.e. pixel accuracy remains the same but IOU improves by 20%. The augmentation strategy consisted of making 4 modified copies of each training image, where each copy was subjected to random crops, flips, and illumination changes. Given this is a relatively minor augmentation strategy, we see there is room to improve the results further simply by making the augmentation strategy more complex or pronounced. For example, [A17] use an augmentation factor of 720 compared to our 4.

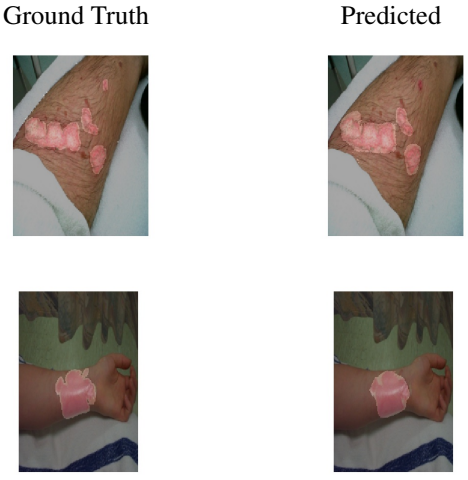

Figure 8: Examples of images where we are able to accurately segment burnt skin.

Looking at the common errors, which see three consistent themes. First, we slightly over-predict the boundary of the burn. In this case, more training and augmentation should help. Furthermore, this is likely an artifact of the human error and imprecise nature of the annotation tool i.e.

we're likely never to get a perfect IOU. Second, many times we seemingly over-predict a burn, but in reality, we are catching burns that were not labeled i.e. we are doing better than the metrics indicate. This is again an artifact of the labeling process. To elaborate, the segmentation aspect of the data collection process is the most tedious and thus acts as the biggest bottleneck (considering the time constraint plastic surgeons face). It is almost impossible to expect the surgeons to be able to segment all burns in a photo, especially the minor ones. Given this, we must change our modeling paradigm to not assume we have completely labeled data i.e. shift to semi-supervised methods. Furthermore, we can use this insight to form a feedback loop in our data collection process. For example, allow the surgeons to agree or disagree with a predicted segmentation, which we then feed back into the training set. Third, the algorithm struggles to differentiate between burnt skin and non-clear (but non burnt skin). This category is somewhat similar to the previous, but involves skin that is freckled or is slightly red due to some other condition (but is not a burn). Adding in various images of non-burnt skin, such as from a dermatology corpus, should help here. In other words, the algorithm simply has not seen enough images of the various types of non-burnt skin. Overall, given the minimal training time and dataset size, this represents a very successful first attempt at segmenting burnt skin with clear pathways forward to further improve our results.

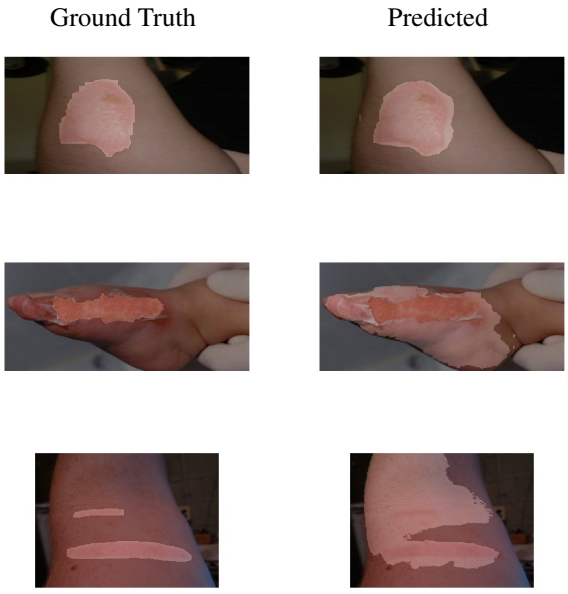

Figure 9: At the top, the algorithm slightly over-predicts the bounds. In the middle, the algorithm seemingly over-predicts, but actually catches less severe burns which were not labeled. At the bottom, the algorithm struggles to differentiate between burnt skin and non-clear skin.

4.2. Multi-Burn

We next attempted to extend our model to predicting the 4 different burn depths. Our initial results are middle of the ground with a pixel accuracy of .60 and IOU of .37.

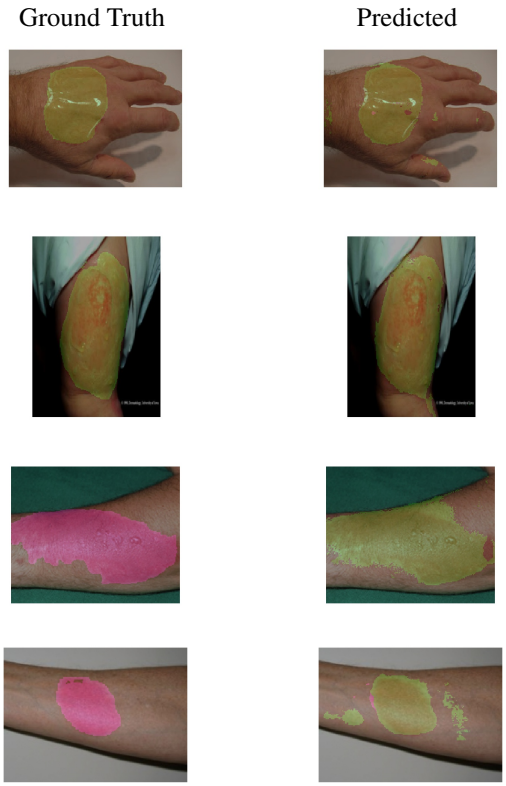

Figure 10: The first two sets of images represent us accurately segmenting and classifying partial thickness burns. The last two images are superficial which we said were partial thickness.

We find that we are only good at predicting partial thickness burns with the rest of the burns having a sub .15 IOU (Figure 11). Furthermore, when looking at the confusion matrix of our pixel predictions, we find that when we mispredict these other burns, it is because we are predicting them to be partial thickness instead or categorizing them as not burnt skin (Figure 12).

Digging deeper, it is evident the main culprit is the class imbalance between the different burn depths (Figure 13). Particularly, partial thickness burns are present in 334 out of the 396 images in the training set. The next largest is full thickness, which appears in only 106 images. Given the drastic imbalance, it is clear the algorithm is simply predicting partial thickness for the most part and has not been able to learn to differentiate the rest of the burn depths.

Given this, we experimented with three different strategies to account for this imbalance. In the first, we upsam-

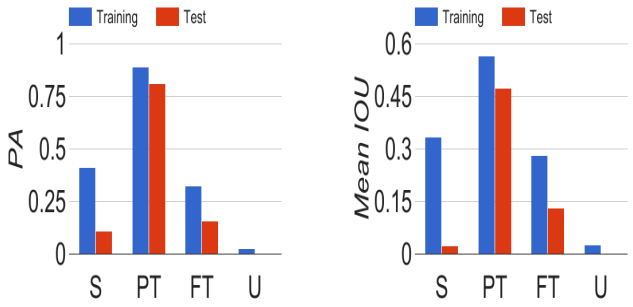

Figure 11: We are only accurate at predicting partial thickness burns.

	Predicted				
	B	S	PT	FT	U
B	650K	105K	**458K**	50K	4K
S	**13K**	11K	**54K**	2K	0
PT	78K	23K	607K	21K	0
FT	**72K**	2K	**88K**	44K	0
U	**35K**	0	**26K**	2K	0

Figure 12: Confusion matrix of results on the test set. Most of our wrong predictions are predicting no burn or over-predicting partial thickness burns.

	S	PT	FT	U
Pixels	660K	**3.8M**	1.1M	313K
Images	98	**564**	163	86

Figure 13: Breakdown of the number of pixels and images corresponding to each burn depth category across the whole dataset used. We see partial thickness burns make up a clear majority.

pled the classes until the amount of pixels and images associated with each class were relatively the same. To do so, we copied and augmented the images that did not contain partial thickness burns. In the training set, this corresponded to 62 images. This resulted in extremely modest improvements (which could be due to chance and natural variability in the learning process alone). Next, we attempted to employ a weighted cross entropy loss. In our formulation, we take the pre-softmax predictions for each pixel and weight the corresponding class index by the ratio of the number of pixels in the background class to the number of pixels with that class. For example, there were 21M pixels associated with the background class in the training set and 418K associated with superficial burns. The weight for the superficial prediction score was then 50. In the model, the weighting is done by creating a 1x1 conv layer (5x5x1x1 with no biases) before the softmax. In one scenario, we fix this layer's weights. In another, we allow the model to learn this layer. Both of these resulted in substantially worse results.

	PA	IOU
FCN - CRF	.60	.37
Upsampled	.57	.39
Fixed Weighted	.33	.19
Learned Weighted	.36	.24

Figure 14: Performance on the test set when we expand to multiple burn depths. Fixed weighted refers to keeping the 1x1 conv layer static while learned refers to allowing this layer to be updated.

5. Discussion

Upsampling produced a decrease in PA and a slight increase in IOU. However, looking at the training versus test metrics, we see that it led to substantially more overfitting for superficial, full thickness, and undebrided burns. Looking back, this makes sense as we only upsampled/augmented images that did not contain partial thickness burns. Since 84% of images contained PT burns, this means that we essentially magnified and sampled from the tails of the natural burn image distribution and expected it to generalize to the whole distribution, which we see does not work (nor does it make sense from a statistical perspective). Instead, we need to find a strategy for upsampling then augmenting those images that contain partial thickness burns. Simply using a normal augmentation procedure on these images does not work, as it leads to the same (or sometimes even worse) class imbalance.

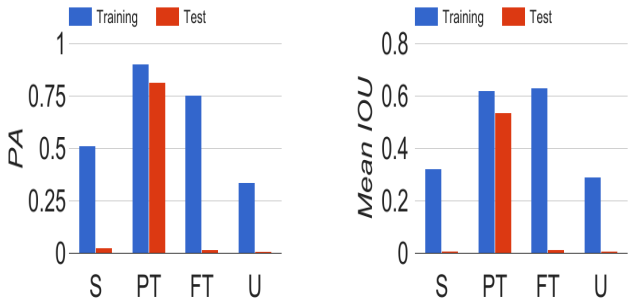

Figure 15: Results on the test set when we upsampled. Compared to Figure 11, we see even more overfitting compared to no upsampling.

Our attempt to account for the class imbalance using a weighted prediction before the softmax led to considerably worse results. This too also makes sense. Consider

this set of unnormalized probabilities for a given pixel: $[1.0, -1.5, 2.0, -.5, -10]$. This represents scores for background class (not a burn), superficial, partial thickness, full thickness, and undebrided respectively. Ideally, the weighting mechanism would close the gap between superficial and partial thickness. However, any weighting (assuming we have positive coefficients) would cause the superficial score to become even more negative, thus making the gap even larger. What we should have done is weighted the normalized probabilities i.e. the scores after the softmax.

Overall, we see there is clear potential for creating a system which accurately segments burn depths, but we need to find creative strategies to account for the significant class imbalance.

6. Next Steps

6.1. Dataset

We need to make some modifications to the dataset. First, we need to focus on acquiring images that contain superficial and full thickness burns. We will likely need to expand our Google Search curation, particularly for superficial burns, given that images coming from a medical center are biased towards containing more severe burns. Second, we need to develop a strategy for augmenting the images that contain partial thickness burns instead of only augmenting those that do not contain them. We've already discussed how only augmenting images that do not contain PT burns is the same as magnifying the tails of the distribution, which we do not expect to generalize well. Perhaps blacking out the pixels associated with the PT burns and augmenting the rest of the pixels could help here. Third, we need to add more varied non-burnt skin images to the dataset. A good source could be a dermatology corpus. Finally, as long as we have humans performing the segmentation and labeling, we will be introducing bias into the dataset. Though a significant cost, the use of laser Doppler Imaging (or potentially thermal imaging) would be the gold standard to creating an actual ground truth dataset.

6.2. Modeling

On the modeling side, we need to do proper weighted cross entropy loss instead of the weighted loss we employed i.e. weight the logits after the softmax and not before. We also need to change our paradigm and assume we have incompletely labeled data. In other words, we are no longer in a supervised setting, but rather a semi-supervised setting. Finally, it could be interesting to experiment with a different network architecture, such as Mask R-CNN [K17].

6.3. Metrics

Though pixel accuracy and intersection over union are useful from a computer science perspective, we need to convert them into more clinically valuable metrics, such as specificity and sensitivity. This needs to happen at both the pixel level and the classification level. This is straight forward on the pixel level, but on the classification level, a discussion needs to happen with plastic surgeons to determine what exactly constitutes a recovered burn. Do we base it off of IOU or PA? What threshold of these do we use? This is the reason our results cannot directly be compared to those in the related work section. We could give an IOU threshold of .5 and use that to calculate classification metrics, thus enabling a more direct comparison. However, this decision should be rooted in clinical utility. Any arbitrary threshold that we set now would cause unproductive comparisons.

6.4. EHR

Since the majority of the images come from Valley Medical Center, we have access to the patient information. Specifically, we can identify the age and race of the patient. It would be interesting to analyze how the algorithm performs across different races considering the same burn depth is visually different across varied skin tones. Furthermore, adding in age and race could prove valuable in increasing the predictive ability of our model.

6.5. TBSA

Given the success of discriminating between burnt skin and the rest of the image, it seems highly likely that we can use a similar approach to calculate total body surface area (the percentage of the body that is burned). To do so, we'll first need to segment the body's outline from the rest of the image and then turn the algorithm into a 3 class segmentation problem: background, regular skin, and burnt skin.

7. Conclusion

In conclusion, we've created the world's largest segmented and labeled burn dataset. With minimal training, we are able to accurately segment burnt skin from the rest of the image, with clear potential to do even better. This opens the door to performing clinically valuable calculations, such as TBSA. Furthermore, we've shown that with some dataset and modeling changes, we can extend the algorithm to segmenting multiple burn depths. This project brings us one step closer to scaling expert level care to millions of burn victims worldwide.

References

[N70] Herndon D. N. et al. "A New Concept in the Early Excision and Immediate Grafting of Burns". In: *The Journal of Trauma* 10.12 (1970), pp. 1103–1108.

[M84] Heimbach D. M. et al. “Burn Depth Estimation - Man or Machine”. In: *Journal of Trauma* 24.5 (1984), pp. 373–378.

[N89] Herndon D. N. et al. “A Comparison of Conservative Versus Early Excision: Therapies in Severely Burned Patients”. In: *Annals of Surgery* 209.5 (1989), pp. 547–553.

[M93] Niazi Z. B. M. et al. “New Laser Doppler Scanner, a Valuable Adjunct in Burn Depth Assessment”. In: *Burns* 19 (1993), pp. 485–489.

[L99] Roa L. et al. “Digital Imaging in Remote Diagnosis of Burns”. In: *Burns* 25 (1999), pp. 617–623.

[AAO01] Pape S. A., Skouras C. A., and Bryne P. O. “An Audit of the Use of Laser Doppler Imaging (LDI) in the Assessment of Burns of Intermediate Depth”. In: *Burns* 27 (2001), pp. 233–239.

[B05] Acha B. et al. “Segmentation and Classification of Burn Images By Color and Texture Information”. In: *Journal of Biomedical Optics* 10.3 (2005), p. 034014.

[K05] Yeong E. K. et al. “Prediction of Burn Healing Time Using Artificial Neural Networks and Reflectance Spectrometer”. In: *Burns* 31 (2005), pp. 415–420.

[SMC06] Ong Y. S., Samuel M., and Song C. “Meta-analysis of Early Excision of Burns”. In: *Burns* 32 (2006), pp. 145–150.

[X06] Murphy R. X. et al. “The Reliability of Digital Imaging in the Remote Assessment of Wounds: Defining a Standard”. In: *Annals of Plastic Surgery* 56.4 (2006), pp. 431–436.

[K07] Shokrollahi K. et al. “Mobile Phones for the Assessment of Burns: We have the Technology”. In: *Emergency Medical Journal* 24 (2007), pp. 753–755.

[BC08] Russell B.C. et al. “LabelMe: a database and web-based tool for image annotation”. In: *International Journal of Computer Vision* 77.1-3 (2008), pp. 157–173.

[H09] Hoeksema H. et al. “Accuracy of Early Burn Depth Assessment by Laser Doppler Imaging on Different Days Post Burn”. In: *Burns* 35 (2009), pp. 36–45.

[B13] Acha B. et al. “Burn Depth Analysis Using Multidimensional Scaling Applied to Psychophysical Experimental Data”. In: *Transactions on Medical Imaging* 32.6 (2013), pp. 1111–1120.

[P14] Ganapathy P. et al. “Dual-imaging system for burn depth diagnosis”. In: *Burns* 40 (2014), pp. 67–81.

[R14] Resch T. R. et al. “Estimation of Burn Depth at Burn Centers in the United States: A Survey”. In: *Journal of Burn Care Research* 35.6 (2014), pp. 491–497.

[JET15] Long J., Shelhamer E., and Darrell T. “Fully Convolutional Networks for Semantic Segmentation”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 3431–3440.

[K15] He K. et al. “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”. In: *The IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 1026–1034.

[O15] Russakovsky O. et al. “ImageNet Large Scale Visual Recognition Challenge”. In: *International Journal of Computer Vision* 115.3 (2015), 211fffdfffdfffd252.

[R15] Wachter R. *The Digital Doctor: Hope, Hype, and Harm at the Dawn of Medicine’s Computer Age*. McGraw-Hill Education, 2015.

[S15] Zheng S. et al. “Conditional Random Fields as Recurrent Neural Networks”. In: *The IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 1529–1537.

[YYH15] LeCun Y., Bengio Y., and G. Hinton. “Deep Learning”. In: *Nature* 521 (2015), pp. 436–444.

[Ker16] Martin Kersner. *train-CRF-RNN*. 2016. URL: `https://github.com/martinkersner/train-CRF-RNN` (visited on 04/12/2017).

[A17] Esteva A. et al. “Dermatologist-level Classification of Skin Cancer using Deep Neural Networks”. In: *Nature* 542 (2017), pp. 115–118.

[Gro17] Torr Vision Group. *crfasrnn*. 2017. URL: `https://github.com/torrvision/crfasrnn` (visited on 04/20/2017).

[Jun17] Alexander Jung. *imgaug*. 2017. URL: `https://github.com/aleju/imgaug` (visited on 04/20/2017).

[K17] He K. et al. “Mask R-CNN”. In: *arXiv* (2017).