Utilizing Mask R-CNN for Detection and Segmentation of Oral Diseases

Rajaram Anantharaman
Computer Science/Biomedical Informatics
University of Missouri - Kansas City
Kansas City, Missouri 64110
ranantharaman@mail.umkc.edu

Matthew Velazquez
Computer Science/Biomedical Info.
University of Missouri - KC
Kansas City, Missouri 64110
mv3md@mail.umkc.edu

Yugyung Lee
Computer Science
University of Missouri - KC
Kansas City, Missouri 64110
Email: leeyu@umkc.edu

***Abstract*— In this paper, we demonstrate the application of Mask-RCNN, the state-of-the-art convolutional neural network algorithm for object detection and segmentation to the oral pathology domain. Mask-RCNN was originally developed for object detection, and object instance segmentation of natural images. With this experiment, we show that Mask-RCNN can also be used in a very specialized area such as oral pathology. While the number of oral diseases are numerous and varied in the form of Thrush, Leukoplakia, Lichenplanus, etc., we limited our scope to the detection and instance segmentation of two of the most commonly occurring conditions, herpes labialis (commonly referred to as "cold sore") and aphthous ulcer (commonly referred to as "canker sore"). This paper aims at detecting and segmenting cold sores and canker sores only. As always, no computer based detection system can be 100% reliable. An accurate diagnosis by a trained health care professional is necessary since several conditions of the mouth including oral cancer may mimic canker sores.**

***Index Terms*— Mask-RCNN, Object Segmentation, Object Detection, Oral disease**

I. INTRODUCTION

Medical imaging is performed in various modalities, such as X-ray, Magnetic Resonance Imaging (MRI), Computed Topography (CT), microscopy, endoscopy, ultrasound, positron emission tomography (PET), and many more. These images have become one of the primary means of diagnosis, clinical studies, and treatment planning. Computer aided analysis of these images have become increasingly common. As such, reliable algorithms are required for the delineation of anatomical structures and other regions of interest (ROI) in these images [1].

Deep learning algorithms have been widely used for analyzing medical images such as with image classification, object detection, segmentation [2], [3], [4]. Over the past few years, various segmentation techniques with differing accuracy and complexity have been developed and reported for transfer learning to the segmentation task [5]. In brain MRI analysis, image segmentation is commonly used for measuring and visualizing the brains anatomical structures, analyzing brain changes, delineating pathological regions, or for surgical planning and image-guided interventions [6]. Mansoor et al. provide several different approaches for segmenting lungs with pathologic conditions on chest CT images [7]. Cascaded convolutional neural network (CNN) was designed with multiple layers of anisotropic and dilated convolution filters for automatic segmentation for brain tumor [8]. Segmentation and feature extraction was proposed for dental x-ray images by using a level-set method for teeth segmentation [9]. CNN with transfer learning was used for classification of dental diseases [10].

However, all these medical imaging modalities use well-controlled clinical environments with specialized equipment. With the proliferation of mobile phones, there is an opportunity to use readily available visible light images (e.g., taken with smart-phone and standard digital camera) in the preliminary diagnosis of certain conditions. In this paper, we report a deep learning algorithm based on Mask-RCNN, trained with annotations from dental professionals, that provides pixel-wise cold sore and canker sore segmentations of color-augmented visible light images.

The reason why this work is important is because Cold sores, caused by Herpes simplex virus type 1 (HSV-1), is a highly contagious infection which is common and endemic throughout the world. Most HSV-1 infections are acquired during childhood and infection is lifelong. While the vast majority of HSV-1 infections are oral herpes (infections in or around the mouth), a proportion of HSV-1 infections are genital herpes (infections in the genital or anal area). In 2012, an estimated 3.7 billion people under the age of 50, or 67% of the population had HSV-1 infection. Estimated prevalence of the infection was highest in Africa (87%) and lowest in the Americas (40-50%). Symptoms of oral herpes include painful blisters or open sores called ulcers in or around the mouth. HSV-1 is mainly transmitted by oral-to-oral contact causing oral herpes infection via contact with the HSV-1 virus in sores, saliva, and surfaces in or around the mouth [11]. We feel that providing an automated way to detect and isolate cold sores reliably is necessary. In a previous work, Anantharaman et al. presented a Convolutional Neural Network based classifier to classify mouth sores [12]. Our work extends that work to include detection and segmentation as well.

This paper is organized as follows. Section II lists some of the work that is being done with disease detection and segmentation within the oral health domain. Section III describes our technical approach to data collection, ground truth generation, data preparation, training algorithm, and training method. Section IV provides a detailed description of our approach to recording and comparing the results obtained. Section V outlines the limitations of this work and direction for future work. Finally, Section VI presents our conclusion.

978-1-5386-5488-0/18/$31.00 ©2018 IEEE

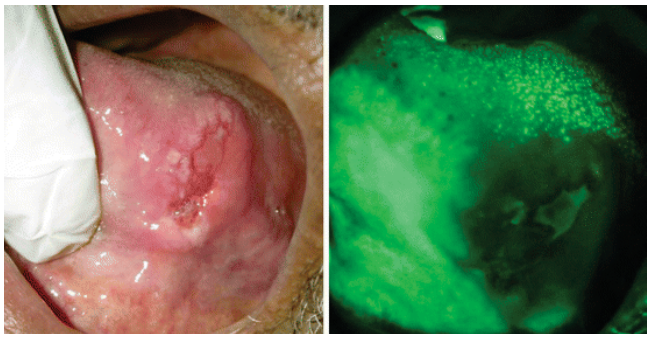

Fig. 1. Comparison of oral cavity images under visible light and fluorescent light. Fluorescent images are taken with the aid of specialized equipment and include features under the skin surface. On the contrary, visible light images (e.g., taken with smartphones) do not include these features. This work focuses on the analysis of visible light images. Image Source: [13].

II. RELATED WORK

A. Segmentation of Gingival diseases

In a recently published work, Rana et al. describe an automated system that performs pixel-wise segmentation of the inflamed gingiva to detect gingivitis and periodontal disease using fluorescence images acquired by an intraoral camera [14]. To the best of our knowledge, this remains the only attempt other than ours to create an automated solution to detect and segment oral diseases using images and computer vision. For their experiment, they collected Intraoral fluorescent images, as seen in Figure 1, from 150 consenting adults, aged 18-90 years old. These images were analyzed by dentists for gingivitis and then used to train a machine learning classifier. The trained classifier accepts an intraoral image of gums and teeth and provides a localized and automated detection of gingival inflammation and periodontal disease on a per-pixel basis.

B. Periodontal disease detection

In a related work from two decades ago, Juan et al. used computer vision techniques and incorporated an off-the-shelf camera to automatically predict gingival probe depth using training data with ground truth measurements [16]. Rana et al. claim that this work made depth predictions of gingival inflammation with reasonable accuracy but lacked other key parameters such as inflammation and bleeding indices, hypervascularization and papillary margin quality [14].

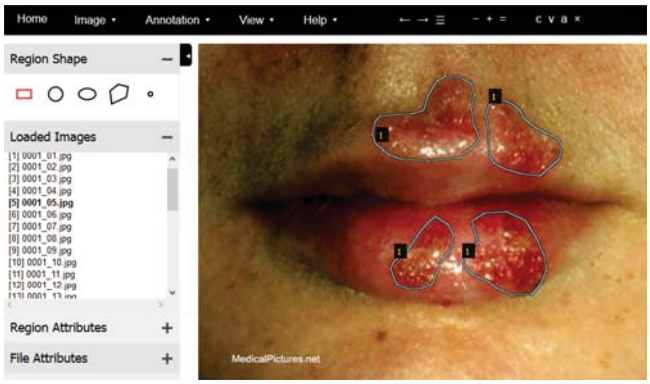

Fig. 2. User Interface of the VGG Image Annotator tool [15]

C. Melanoma detection and segmentation

Although not in the oral health domain, Do et al. have recently proposed a similar study for melanoma detection and segmentation [17]. Their work focuses on accessible detection of malignant melanoma (MM) using mobile image analysis. MM is a type of skin cancer arising from the pigment cells of the epidermis. Their work was unique in that they were using visible light images captured from smartphones for automatic melanoma detection. They go on to mention that most previous works focused on dermoscopic images that are captured in the well-controlled clinical environments with specialized equipment. Our proposed work parallels this approach within our proposed domain.

D. Automatic Nucleus Segmentation using Mask-RCNN

Johnson [18] demonstrated that the Mask-RCNN model, while primarily designed for object detection, object localization, and instance segmentation of natural images, can also be used for the task of segmentation of nuclei in widely varying microscopy images. They were able to achieve decent results with very little modification of Mask-RCNN. Our proposed work takes a similar approach in utilizing Mask-RCNN for an equally challenging dataset that contains widely varying cold sore and canker sore images.

E. Segmentation of corneal endothelium images

Fabijanska [19] proposes to perform cell segmentation using a U-Net-based convolutional neural network. The network is trained to discriminate pixels located at the borders between cells. The edge probability map outputted by the network is next binarized and skeletonized in order to obtain one-pixel wide edges. The author tested her solution on a dataset consisting of 30 corneal endothelial images presenting cells of different sizes, achieving an AUROC level of 0.92. The resulting DICE was on average equal to 0.86.

III. TECHNICAL APPROACH

A. Data collection

Since there is no publicly available oral disease dataset, we had to collect images from publicly available sources. we collected a set of 40 cold and canker sore images from Google images to construct our sparse sore detection dataset. Sore images are selected with different locations, sizes, shapes, colors and various lighting conditions for model generalization. The data consists of cold sore and canker sore images and corresponding annotations for each individual sore detected by an oral pathologist in each image. The sores in the images are derived from a wide range of people of varying races and sex. Furthermore, the images have been captured in variety of lighting conditions and appear in a variety of shapes, intensity, and location. This presents a significant additional challenge as convolutional neural

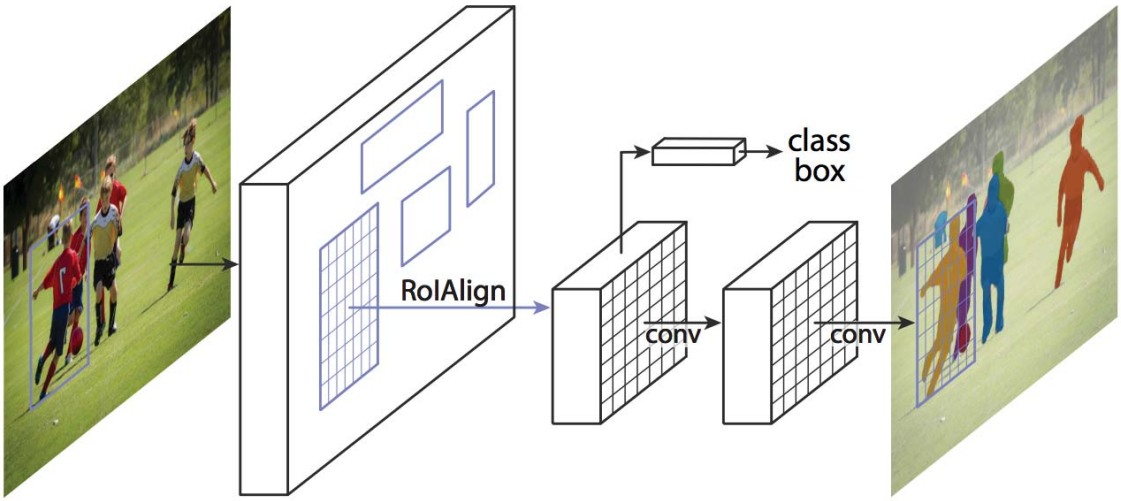

Fig. 3. MaskR-CNN Framework. Image from [20].

networks can be expected to perform best, in general, when the input data is as uniform and standardized as possible. This includes standardization in terms of color, contrast, scale, and class balance. Of these 40 images in our sparse dataset, 30 images were used for training and 10 images were used for validating the model. An additional 10 random images were collected from Google images for testing.

B. Ground truth generation

A copy of VIA (VGG Image Annotator) [15] was provided to display images to an oral pathologist for annotations. An example of this utility can be seen in Figure 2. VIA is a single HTML file that can be downloaded and opened in a browser. No installation is necessary. The oral pathologist provided bounding polygons around regions of cold sores and canker sores along with a region attribute value of either 1 or 2 for each region. The VIA tool saves the annotations in a JSON file, and each mask is a set of polygon points. The polygons provided by the oral pathologist were representative, not exhaustive, resulting in some of the cold sore and canker sore pixels lying outside the bounding polygons, which could have affected the model accuracy. The pixels inside the bounding polygons corresponding to cold sore and canker sore were given a value of 1 and 2; rest of the pixels were regarded as background given a value of 0. This operation resulted in 40 pairs of images and corresponding sore segmentations.

C. Data Preparation

The image dataset was divided into training and validation data: 30 images (0.75) and 10 images (0.25) respectively. Images were resized to a size of 640 pixels wide and 480 pixels tall. The resolution was set to 96 dpi and the bit depth was set to 24. The distribution of the dataset across the sore classes were identical. No augmentation was applied to the training images. The validation dataset was randomly selected to preserve the proportion of images with each sore class.

D. Training Algorithm

The U-Net architecture was developed explicitly with the segmentation of medical images in mind, and used to produce state-of-the-art results on on the ISBI challenge for segmentation of neuronal structures in electron microscopic stacks as well as the ISBI cell tracking challenge 2015 [21]. The Mask-RCNN model was developed in 2017 and extends the Faster-RCNN model for semantic segmentation, object localization, and object instance segmentation of natural images [20]. We were quite surprised and fascinated by how well Mask-RCNN was used to outperform all single-model entries on every task in the 2016 COCO challenge, a large-scale object detection, segmentation, and captioning challenge [22].

Mask-RCNN, relies on region proposals which are generated via a region proposal network. Mask-RCNN follows the Faster-RCNN model by having a feature extractor followed by this region proposal network. This is then supplemented by an operation known as ROI-Pooling consisting of three important modifications to produce standard-sized outputs suitable for input to a classifier [23]. First, Mask-RCNN replaces the somewhat imprecise ROI-Pooling operation used in Faster-RCNN with an operation called ROI-Align that allows very accurate instance segmentation masks to be constructed; and second, Mask-RCNN adds a network head (a small fully convolutional neural network) to produce the desired instance segmentations; as illustrated in Figure 3. Finally, mask and class predictions are decoupled; the mask network head predicts the mask independently from the network head predicting the class. This entails the use of a multi-task loss function [23].

$$L = Lcls + Lbox + Lmask \tag{1}$$

The multi-task loss function of Mask R-CNN combines the loss of classification, localization and segmentation mask.

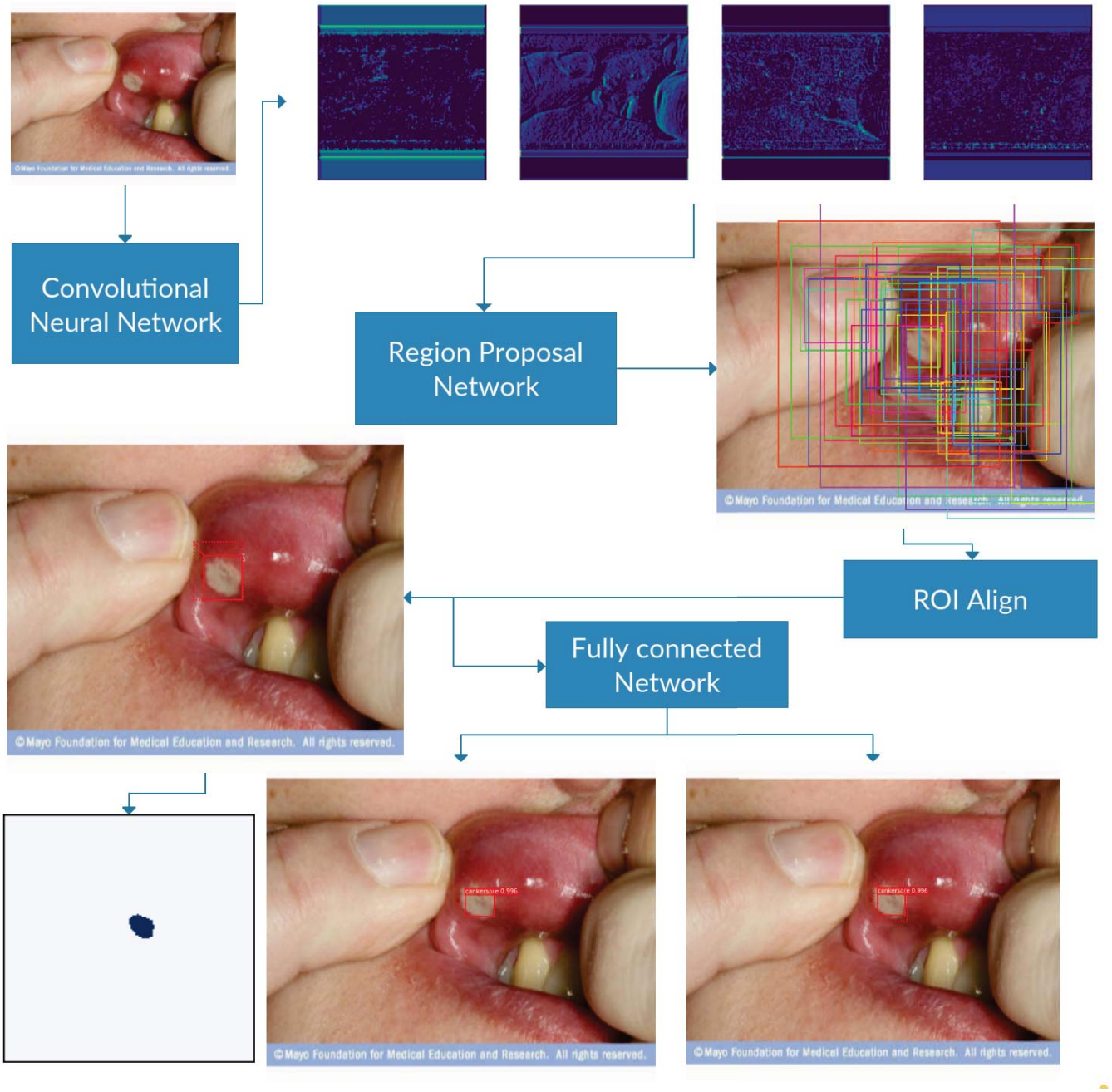

Fig. 4. Mask R-CNN Workflow

To learn more about MaskR-CNN, we encourage readers to refer to [20].

Mask R-CNN consists of several modules. Our workflow, along with the outputs that each module generates for our case is depicted in Figure 4.

1) Backbone: This is a standard convolutional neural network (typically, ResNet50 or ResNet101) that serves as a feature extractor. The early layers detect low level features (edges and corners), and later layers successively detect higher level features (cold sore, canker sore, etc.). Passing through the backbone network, the image is converted from 640x480px x 3 (RGB) to a feature map of shape 32x32x2048. This is depicted in Fig. 4. The input image passes through the convolutional neural network backbone to create the feature map. This feature map becomes the input for the stages that follow. While the backbone described above works great, it can be improved upon. The Feature Pyramid Network (FPN) was introduced by the same authors of Mask R-CNN as an extension that can better represent objects at multiple scales. FPN improves the standard feature extraction pyramid by adding a second pyramid that takes

Fig. 5. Final Predictions and Color Splash

the high level features from the first pyramid and passes them down to lower layers. By doing so, it allows features at every level to have access to both, lower and higher level features. Our implementation of Mask RCNN uses two seperate combinations. 1) ResNet101 + FPN backbone, and 2) ResNet50 + FPN backbone.

2) Region Proposal Network (RPN): The RPN is a lightweight neural network that scans the image in a sliding-window fashion and finds areas that contain objects. The regions that the RPN scans over are called anchors. Which are boxes distributed over the image area, as show in Fig. 4. This is a simplified view, though. In practice, there are about 200K anchors of different sizes and aspect ratios, and they overlap to cover as much of the image as possible. Using the RPN predictions, we pick the top anchors that are likely to contain objects and refine their location and size. If several anchors overlap too much, we keep the one with the highest foreground score and discard the rest (referred to as Non-max Suppression). After that we have the final proposals (regions of interest) that we pass to the next stage.

3) ROI Classifier and Bounding Box Regressor: This stage runs on the regions of interest (ROIs) proposed by the RPN. And just like the RPN, it generates two outputs for each ROI:

- Class: The class of the object in the ROI. Unlike the RPN, which has two classes (FG/BG), this network is deeper and has the capacity to classify regions to specific classes (cold sore, canker sore, etc.). It can also generate a background class, which causes the ROI to be discarded.
- Bounding Box Refinement: Very similar to how its done in the RPN, and its purpose is to further refine the location and size of the bounding box to encapsulate the object.

Classifiers dont handle variable input size very well. They typically require a fixed input size. But, due to the bounding box refinement step in the RPN, the ROI boxes can have different sizes. Thats where ROI Pooling comes into play. ROI pooling refers to cropping a part of a feature map and resizing it to a fixed size. Its similar in principle to cropping part of an image and then resizing it. The authors of Mask R-CNN suggest a method they named ROIAlign, in which they sample the feature map at different points and apply a bilinear interpolation. In our implementation, we used TensorFlows cropandresize function for simplicity and because its close enough for most purposes. The output of the ROI classifier and the Bounding Box Regressor is depicted in the last row of Fig. 4.

4) Segmentation Masks: The mask branch is a convolutional network that takes the positive regions selected by the ROI classifier and generates masks for them. The generated masks are low resolution: 28x28 pixels. But they are soft masks, represented by float numbers, so they hold more details than binary masks. The small mask size helps keep the mask branch light. During training, we scale down the ground-truth masks to 28x28 to compute the loss, and during inferencing we scale up the predicted masks to the size of the ROI bounding box and that gives us the final masks, one per object as shown in the last row of Fig. 4 on the left.

E. Training Method

Our version of MaskR-CNN is based on an existing implementation by Matterport Inc. [24] released under an MIT License, and which is itself based on the open-source libraries Keras and Tensorflow. We used both ResNet-50 feature pyramid network model and a ResNet-101 feature pyramid network model as a backbone. We also trained our model with two different sets of the same images. The first set contained 30 training and 10 validation images while the second contained 20 training and 20 validation images. Rather than training the network end-to-end from the start, we initialize the model using weights obtained from pretraining on the MSCOCO dataset [22] and proceed to training only the network heads. In total we train for 20 epochs using stochastic gradient descent with 100 training steps per epoch, momentum of 0.9, and learning rate of 0.001. We used a batch size of 2 on a single NVIDIA 1080 Ti GPU. The only image preprocessing we did was to resize the images to a standard size of 640X480 pixels. We skipped detections with less than 90% confidence. The

Fig. 6. Mask R-CNN results (ResNet-101) for Cold Sores (top) and Canker Sores (bottom).

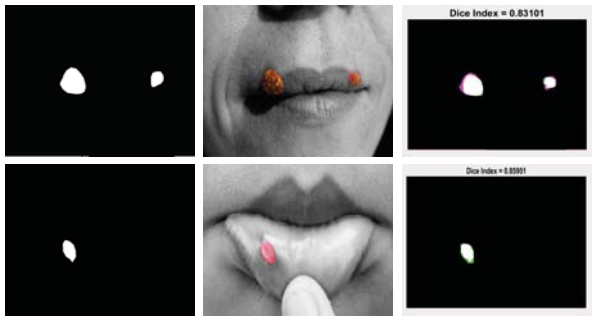

Fig. 7. Mask R-CNN results for Cold Sore (top) and Canker Sore (bottom). Images shown are Ground Truth, Predicted Mask, and Dice Index.

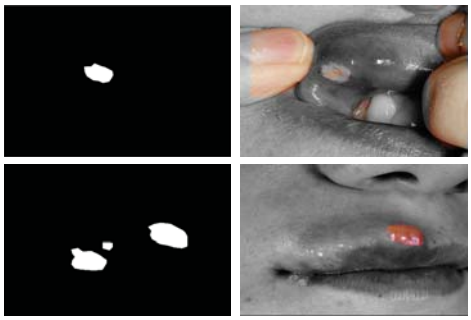

Fig. 8. Ground truth mask and failed prediction on ResNet-50 backbone (top). Ground truth mask and partial prediction on ResNet-101 backbone (bottom).

final predictions and the color splash are shown in Figure 5. Additional predictions generated by our model can be seen in Figure 6.

IV. RESULTS

We compare our models by calculating individual Dice coefficients per predicted image mask, and then combining those results into one overall average per class. The individual Dice indices are derived from the formula:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} \tag{2}$$

In our context, given two masks (X and Y), the Sorensen index equals twice the number of pixels common to both masks divided by the sum of the number of pixels in each mask. This effectively compares the overlap, or lack thereof, between our dental professional validated ground truth mask and the predicted mask from our model. An example of this overlap comparison can be seen in Figure 7.

Our top model, with ResNet-101 as it's backbone, achieved an average testing Dice coefficient score of .774 (Cold) and .714 (Canker). This version was trained on 30 images (15 per class) and exhibited a noticeable increase in overall Dice score compared to the ResNet-101 model trained on only 20 images (10 per class). While the smaller model showed a similar average with Cold sores (.759 vs. .774), the Canker sore score increased significantly (.484 to .714) when using the additional images for training. This is likely related to the different angles and sizes of the training images. Many variations exist with how these oral pictures are taken and that has a large impact when reducing an already small dataset. The results of these training size and class differences are summarized in Table I.

TABLE I
AVERAGE DICE COEFFICIENT SCORES (RESNET-101)

Training Size	Class	Validation	Test
10	Cold	.875	.759
15	Cold	.865	.774
10	Canker	.556	.484
15	Canker	.948	.714

Our other variation on the model, with ResNet-50 as it's backbone, achieved poor results. The majority of test images did not detect any ROI as the additional layers provided by ResNet-101 proved to be very beneficial when dealing with a sparse dataset. For the test images that did detect ROI, we saw a greater frequency of fingers and teeth that would be included with the predicted mask. An example of this failed prediction can be seen in Figure 8. As an outcome of this, our recommendation is to use larger backbone architectures when leveraging a smaller dataset upfront.

When using the ResNet-101 backbone, these anomalies were much rarer although some masking difficulties were seen when analyzing images with multiple sores. While the majority were masked correctly, these sore variations were the most difficult overall to predict. We believe that increasing the amount of multi-sore training images will help to alleviate this problem in the future.

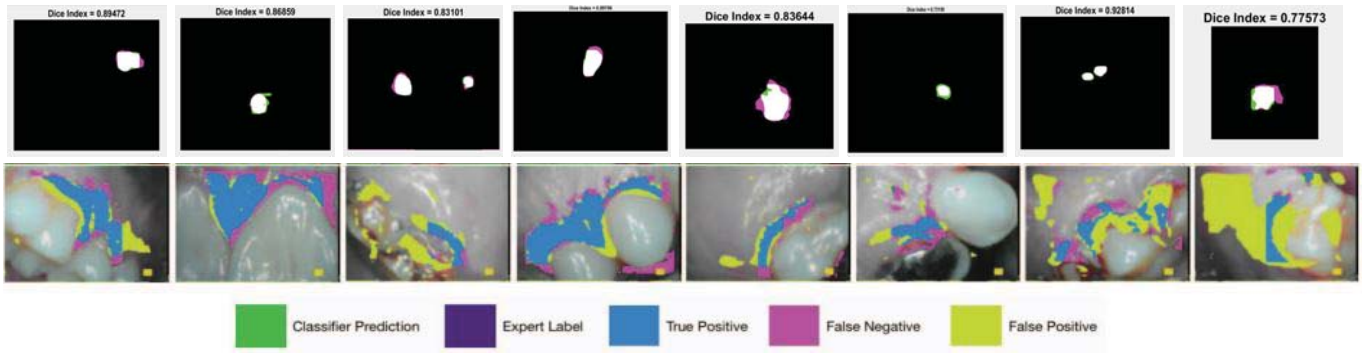

Fig. 9. Segmentation results for inflamed gingiva [6] (bottom) vs. our Canker/Cold sore model (top).

TABLE II
COMPARISON AGAINST RELATED WORK

Segmentation Target	Number of Classes	Training Size	Pixel Accuracy
Inflamed Gingiva [6]	1	258 Images	.621
Canker/Cold Sore	**2**	**30 Images**	**.744**

We compare our Mask R-CNN model to another pixel-wise classifier [6] also in the oral disease domain. As can be seen from Table II, while the segmentation target differs, our model achieves higher pixel accuracy (.744 vs. .621) despite having a smaller training dataset (30 images vs. 258 images). Their work faced similar difficulties regarding the training dataset in that obtaining a greater variety of camera angles and mouth positions would have helped strengthen their results. Sample predictions from both models can be observed in Figure 9.

V. LIMITATIONS/FUTURE WORK

As mentioned previously, our next step would be trying the same sparse dataset against a ResNet-152 architecture. Since we observed significant accuracy gains going from ResNet-50 to ResNet-101, it would be of interest to see if a similar increase in accuracy could be achieved from even more layers.

Currently there are very few, if any, public datasets on canker/cold sores. While this research highlighted the strengths of using Mask R-CNN on small datasets, future research would focus on growing the size of the training data to factor in more image variations. We hypothesize that this would allow for more accurate predictions when given an input image that varies largely from the training data. In this domain, this can be a common occurrence as the angle and distance at which the picture could be taken can vary widely from image to image.

Additionally, future work will include expanding the dataset to encompass more classes. Oral cancer and some gum diseases can be difficult to detect early, but can possibly be simplified by a deep learning implementation. Our model is built in a way that allows for easy expansion of classes so the most difficult piece for that expansion is obtaining the additional training data that would be required.

For our ideal implementation, having a mobile application would fit well within our goals. Being able to capture a live image and have it analyzed for possible oral diseases quickly would significantly benefit rural areas that lack proper dentist coverage. To account for the additional input variations that this would bring, better localization would need to be performed in order to identify the most relevant crop of the overall image. This would help with instances where the image is taken from a distance and other non-relevant features of the individual are on display.

VI. CONCLUSION

For this paper, we studied the potential of using the state of the art in instance segmentation techniques, specifically Mask R-CNN as applied to the oral pathology domain. The paper proposes a simple yet effective system that performs pixel-wise segmentation of visible light images of the oral cavity and successfully segments cold sores and cankers sores. While we achieved promising and encouraging results albeit with a sparse dataset and two classes of diseases, this is just the beginning. Oral diseases are many and varied. A bigger dataset including other classes of oral diseases might result in a more useful system. While segmentation techniques such as Mask R-CNN are being actively researched in other areas of medicine such as melanoma, further research is necessary and warranted in the field of oral pathology.

ACKNOWLEDGEMENT

We would like to thank Matterport, Inc. and Waleed Abdulla for providing the source code for their implementation of Mask R-CNN on Python 3, Keras, and TensorFlow [24]. This research was possible only because of their contribution to the open source community.

REFERENCES

[1] N. Sharma and L. M. Aggarwal, "Automated medical image segmentation techniques," *Journal of medical physics/Association of Medical Physicists of India*, vol. 35, no. 1, p. 3, 2010.

[2] 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.

[3] M. J. Moghaddam and H. Soltanian-Zadeh, "Medical image segmentation using artificial neural networks," in *Artificial Neural Networks-Methodological Advances and Biomedical Applications*. InTech, 2011.

[4] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. M. Hoffman *et al.*, "Opportunities and obstacles for deep learning in biology and medicine," *Journal of The Royal Society Interface*, vol. 15, no. 141, p. 20170387, 2018.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[6] I. Despotović, B. Goossens, and W. Philips, "Mri segmentation of the human brain: challenges, methods, and applications," *Computational and mathematical methods in medicine*, vol. 2015, 2015.

[7] A. Mansoor, U. Bagci, B. Foster, Z. Xu, G. Z. Papadakis, L. R. Folio, J. K. Udupa, and D. J. Mollura, "Segmentation and image analysis of abnormal lungs at ct: current approaches, challenges, and future trends," *RadioGraphics*, vol. 35, no. 4, pp. 1056–1076, 2015.

[8] G. Wang, W. Li, S. Ourselin, and T. Vercauteren, "Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks," in *International MICCAI Brainlesion Workshop*. Springer, 2017, pp. 178–190.

[9] A. E. Rad, M. S. M. Rahim, and A. Norouzi, "Digital dental x-ray image segmentation and feature extraction," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 11, no. 6, pp. 3109–3114, 2013.

[10] S. A. Prajapati, R. Nagaraj, and S. K. Mitra, "Classification of dental diseases using cnn and transfer learning," *2017 5th International Symposium on Computational and Business Intelligence (ISCBI)*, pp. 70–74, 2017.

[11] A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *Journal of pathology informatics*, vol. 7, 2016.

[12] R. Anantharaman, V. Anantharaman, and Y. Lee, "Oro vision: Deep learning for classifying orofacial diseases," in *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 2017, pp. 39–45.

[13] M. Kuriakose, "Contemporary oral oncology," *Cham*, pp. 355–421, 2017.

[14] A. Rana, G. Yauney, L. C. Wong, O. Gupta, A. Muftu, and P. Shah, "Automated segmentation of gingival diseases from oral images," in *Healthcare Innovations and Point of Care Technologies (HI-POCT), 2017 IEEE*. IEEE, 2017, pp. 144–147.

[15] A. Dutta, A. Gupta, and A. Zisserman, "Vgg image annotator (via)," *URL: http://www. robots. ox. ac. uk/˜ vgg/software/via*, 2016.

[16] M.-C. Juan, M. Alcañiz, C. Monserrat, V. Grau, and C. Knoll, "Computer-aided periodontal disease diagnosis using computer vision," *Computerized medical imaging and graphics*, vol. 23, no. 4, pp. 209–217, 1999.

[17] T.-T. Do, T. Hoang, V. Pomponiu, Y. Zhou, C. Zhao, N.-M. Cheung, D. Koh, A. Tan, and T. Hoon, "Accessible melanoma detection using smartphones and mobile image analysis," *IEEE Transactions on Multimedia*, 2018.

[18] M.-C. Juan, M. Alcañiz, C. Monserrat, V. Grau, and C. Knoll, "Computer-aided periodontal disease diagnosis using computer vision," *Computerized medical imaging and graphics*, vol. 23, no. 4, pp. 209–217, 1999.

[19] A. Fabijańska, "Segmentation of corneal endothelium images using a u-net-based convolutional neural network," *Artificial intelligence in medicine*, vol. 88, pp. 1–13, 2018.

[20] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.

[21] 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*. Springer, 2015, pp. 234–241.

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[23] J. W. Johnson, "Adapting mask-rcnn for automatic nucleus segmentation," *arXiv preprint arXiv:1805.00500*, 2018.

[24] W. Abdulla, "Mask r-cnn for object detection and instance segmentation on keras and tensorflow," https://github.com/matterport/Mask_RCNN, 2017.