Convolutional Neural Network for Combined Classification of Fluorescent Biomarkers and Expert Annotations using White Light Images

Gregory Yauney[1], Keith Angelino[1], David A. Edlund[2], and Pratik Shah[1]*

[1] *MIT Media Lab*
Massachusetts Institute of Technology
Cambridge, MA, USA
{*gyauney, kla11, pratiks*}*@mit.edu*

[2] *Hampden Dental Care*
Lakewood, CO, USA
davidedlund777@gmail.com

***Abstract*—Fluorescent biomarkers are important indicators of disease, but imaging them can require specialized and often-expensive devices. Periodontal and dental diseases resulting from microbial plaque biofilms, if diagnosed early with biomarker images and expert knowledge, can be treated to prevent occurrences of serious systemic illnesses. We report two convolutional neural network classifiers trained with dentist annotations of disease signatures and fluorescent porphyrin biomarker images to identify dental plaque in white light images as a per-pixel binary classification task. The classifiers were trained and tested with millions of image patches from two datasets collected from 27 consenting adults using handheld intraoral cameras. The areas under the receiver operating characteristic curves for the test sets were calculated to be 0.7694 and 0.8720. Once trained, the classifiers predict the location of plaque in white light images without requiring specialized biomarker imaging devices or expert intervention. This generalized approach can be useful in other domains where diagnostic biomarker predicting can augment expert knowledge using standard white light images.**

***Keywords*-biomarkers, convolutional neural networks, deep learning, medical devices, periodontal disease, segmentation**

I. Introduction

Identification and analysis of disease imaging biomarkers have provided clinical experts with faster and more efficient diagnostic methods. Computationally automating these methods can offer affordable and scalable technology-enabled health screenings. Fluorescent biomarkers play an important role in the screening of oral cancer and retinal diseases as well as conditions like periodontal disease and dental plaque [1], [2]. Biomarker images have a high degree of accuracy, but capturing them requires specialized and often expensive hardware, annotations, and analyses by experts, resulting in substantial diagnosis delays and thus are seldom used for patient diagnoses. There is also a lack of standardization for the various approaches used by the individual software accompanying biomarker imaging systems, causing many to be considered suboptimal by physician experts.

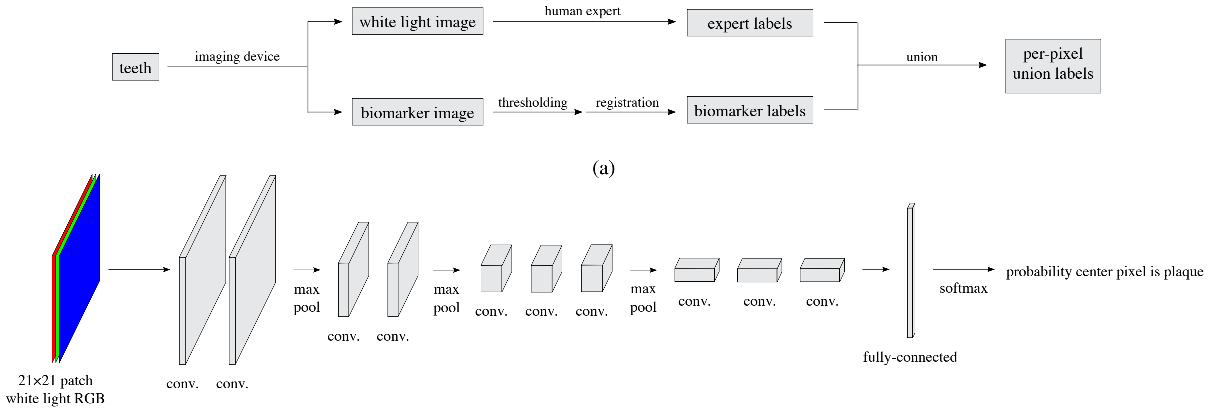
Machine learning and computer vision have automated many aspects of human visual perception. Convolutional neural networks (CNNs) have been used to great effect in numerous computer vision tasks since their introduction for wide-scale image classification [3], largely due to their incorporation of spatial information and their invariance to translation [4]. CNNs are now the predominant technique for analyzing medical images [5]. The specific task of segmentation, which classifies each pixel of an image with membership from among a group of classes, has been studied in medical contexts as diverse as cardiac image segmentation [5] and whole-slide cancer histopathology [6]. Medical segmentation tasks are typically formulated to predict expert labels on specific imaging modalities, such as locating tumors or identifying distinct parts of organs, in either two-dimensional or three-dimensional images [5]. While CNNs have been increasingly used for both medical and non-medical semantic segmentation [7], [8], they have not yet been brought to bear on the specific segmentation problem of predicting the locations of biomarkers in white light images captured by mobile phones and other low-cost imaging devices.

Dental imaging is a particular instance of biomarker imaging that can supplement expert knowledge. A dentist or dental hygienist usually examines teeth and can assign individual tooth scores, as with the Quigley-Hein index, to quantify the amount of plaque on a tooth in aggregate. Such an expert can also provide localized annotations on a white light image based on their expert knowledge of what plaque looks like [9]. Both approaches are time-consuming, subjective, and often only identify plaque visible to the human eye. Fluorescent dyes that bind to plaque are also applied to teeth and imaged or inspected live by a dentist or dental hygienist, with the same limitation in sensitivity and specificity [9]. However, illumination with violet and blue light can excite the porphyrins produced by bacterial plaque to emit light with wavelengths of approximately 650 nm, allowing for the capture of fluorescent images of plaque biomarkers with high sensitivity and specificity [2],

* Corresponding author.

2471-7819/17/31.00 ©2017 IEEE
DOI 10.1109/BIBE.2017.00058

IEEE computer society



(a)

(b)

Figure 1: General labeling and classification pipelines. (a) The process to construct union labels from human experts and a biomarker imaging device for every pixel in a white light image. (b) The convolutional neural network architecture for both classifiers that learns a distribution of union labels over white light image patches. It takes as input a 21×21 pixel white light image patch and outputs a probability between 0 and 1 that the patch's center pixel would be labeled as plaque. This is run for every patch in an input image to determine a probability of plaque for each pixel. Conv.: convolutional layer. Max pool: maximum pooling layer.

[10], [11]. A recent publication from our group reported the construction and clinical validation of a low-cost and open-source porphyrin imaging device and an associated imaging processing algorithm [2]. Commercial intraoral cameras equipped with light-emitting diodes are available to capture porphyrin signatures and are considered highly sensitive, but they lack accurate, clinically-validated image processing algorithms and also do not identify non-fluorescent plaque identified by human experts [11].

In this report, we take up the task of automated and device-independent prediction of porphyrin and plaque signatures from standard white light intraoral images of teeth. Datasets of white light and corresponding fluorescent images showing porphyrin and plaque signatures on teeth were captured using a commercial intraoral camera, ACTEON Soprocare (ACTEON North America, Mount Laurel, New Jersey, USA), referred to as the commercial device (CD), and our own clinically-validated research device (RD). Expert raters also labeled plaque signatures using the white light images captured by both devices. Fig. 1 shows our general approach for labeling and classification. Our fully-trained and validated CNNs, after learning from both fluorescent biomarker images as well as expert labels, accept standard white light intraoral images as inputs and predict the location of plaque pixels with high sensitivity and specificity without requiring device or expert intervention.

II. Related Work

A. Plaque segmentation in images

Image processing algorithms have been previously devised by others to segment dental plaque in white light images but were not comprehensive because plaque is not always easily detectable in a white light image alone. Kang *et al.* segmented plaque in white light images by separately using (A) fuzzy c-means clustering with an objective function accounting for spatial proximity [12] and (B) cellular neural networks to interactively choose a threshold for histogram thresholding [13]. Both approaches produced good qualitative results, but relied on plaque having non-fluorescent pigmentation distinct from tooth surfaces and did not capture fluorescent signatures associated with biomarkers or expert annotations.

Segmentation algorithms have found more success with fluorescent biomarker images, where the plaque is often a distinct color, allowing for increased sensitivity and specificity with image processing techniques like color histogram thresholding [2], [9], [14], intensity thresholding [15], and superpixel graph-cut [16]. Segmentation of each type of image alone is limited, even in the ideal scenario of perfect segmentation, to the signatures of plaque captured by an imaging modality and also lack expert annotations.

B. Convolutional neural networks

A CNN is a type of artificial neural network; it is a sequence of linear and non-linear functions applied to input data in which the linear functions, in the layers closest to the input data, take the form of convolutions instead of arbitrary linear transformations. Convolutional layers are each usually followed by maximum pooling layers, finally followed by fully-connected layers and a softmax function. The output of a CNN is most often interpreted as a probability distribution over possible classes. CNNs are made to learn a distribution

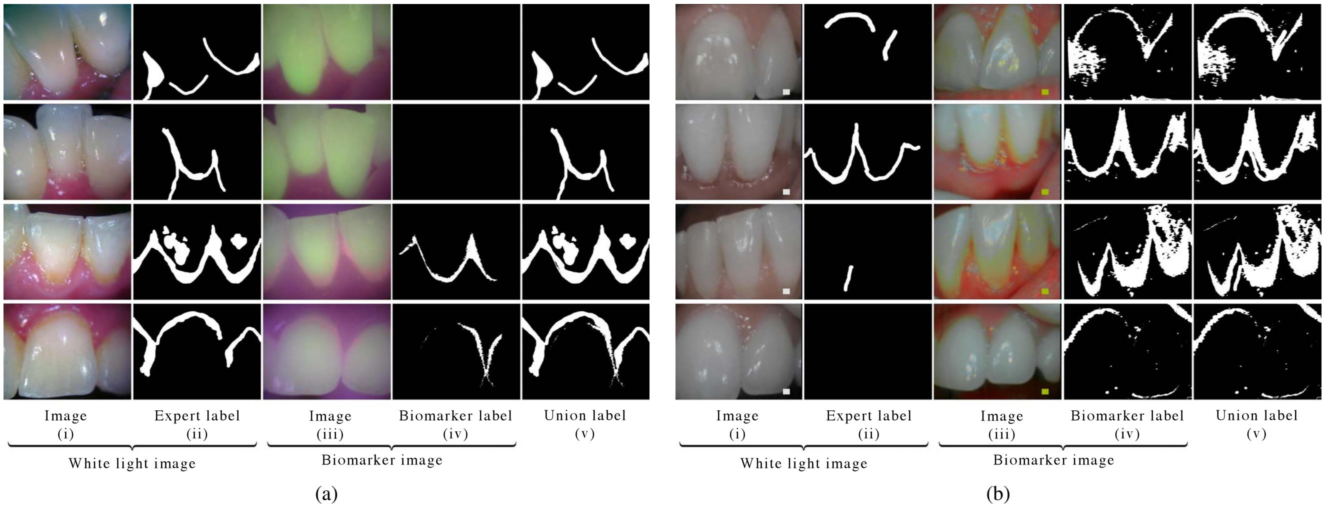


Figure 2: Representative images and labels from the collected datasets. Label images in columns (ii), (iv), and (v) in both subfigures show plaque as white and non-plaque as black. (a) Representative sets of images from the research device dataset. Each row shows images corresponding to one white light image across all grids. Columns, from left to right: (i) white light image, (ii) localized expert annotations for the white light image, (iii) fluorescent biomarker device image of the same location with porphyrins shown in red, (iv) labels extracted from the biomarker image by thresholding and registering to align with the white light image, and (v) the union label created from combining plaque in the expert label and the biomarker label. (b) Representative sets of images from the commercial device dataset. Plaque in (iii) is shown in yellow.

over training data with stochastic gradient descent, with gradient calculation by backpropagation. Convolution allows for sparse weights and maximum pooling for translation invariance [4]. Network architectures, fixed configurations of layers with corresponding parameters, are often reused for multiple tasks once they have been proven to classify accurately. The architecture called VGG, a deep network with uniformly small receptive fields, learns very hierarchical features and is frequently used for object detection [17].

For the task of segmentation, both patch-based classifiers, which take into account a small neighborhood around each pixel, and full-image classifiers have been used to classify all pixels in an image [5]. Fully convolutional networks (FCNs), which use only convolutional and maximum pooling layers to produce a per-pixel heatmap rather than a single distribution over categories, have been increasing in popularity for medical image segmentation since they were initially used for semantic segmentation [5], [7].

III. Technical Approach

We model the white light segmentation problem as a per-pixel binary classification implemented with a patch-based CNN. Two such CNNs were trained on the white light RGB intensities of porphyrin plaque pixels identified by two different fluorescent biomarker imaging devices and the regions annotated by experts on corresponding white light images. The specialized biomarker imaging devices were incorporated during the training phase of the algorithm while only a white light image was used for classification. The performance of the algorithm was evaluated using two different fluorescent biomarker imaging devices which differ by capturing unique signatures associated with newer (CD) and older and more mature plaque (CD and RD) in differing types of plaque images. Fig. 1 shows the full label extraction and classification pipelines. Fig. 2 shows representative sets of images in the collected datasets.

A. *Classifier model*

We hypothesize that plaque's presence on part of a tooth can be deduced from the information contained in the immediate neighborhood. Hence, the CNN architecture takes as input an $n \times n$ white light image patch and outputs a prediction of whether the patch's center pixel corresponds to plaque or not. We propose using this local patch-based method for classifying plaque because plaque's free-form shape makes bounding boxes a poor model of plaque presence; a per-pixel annotation is required. Training on patches rather than full images additionally allows for much more training data from fewer images.

The network architecture is a truncated version of VGG, as one of our goals was to use the smallest possible model with a great enough capacity to learn the training distribution [17]. We experimentally determined the minimum depth that did not result in underfitting on the training set, which is up to and including the thirteenth layer of the VGG16 architecture, followed by a smaller fully-connected layer of 256 nodes and the final softmax function. The CNNs were trained using adaptive stochastic gradient descent with momentum. The loss function captures the softmax cross-entropy in classification of all patches in the current mini-batch. Gradients are calculated by backpropagation. To help prevent overfitting to the training data, we trained with a dropout probability of 0.5 [18].

There is a trade-off between patch size and the amount of different images required; larger patches contain more contextual information around the center pixel but can therefore each capture less of the variation than a smaller patch would, thereby requiring more training patches. That is, the space of variation is larger for larger patches. After initial optimization experiments, a patch size of 21×21 pixels was chosen. Both classifiers were implemented in TensorFlow [19], and training was performed on an NVIDIA Corporation GM200 GeForce GTX TITAN X. Training hyperparameters were determined through grid search: mini-batch size of 100, learning rate of 1×10^{-6}, 3 epochs.

B. Human subjects datasets

The Massachusetts Institute of Technology's Committee on Humans as Experimental Subjects reviewed and approved protocol 1603518893. The CD is an intraoral probe that illuminates plaque with both 450 nm and white light, and then digitally embellishes the color of newly-formed plaque-affected areas in hues of yellow and orange [11]. The RD captures porphyrin signatures associated with mature plaque biofilms formed by anaerobic bacteria via filtered light images; details of the RD are available in the prior work [2]. The RD is also capable of capturing white light images.

27 adult subjects consented to imaging of incisors and canines. Each subject was imaged sequentially in (1) CD white light mode, (2) CD plaque mode, (3) RD white light mode, and (4) RD plaque mode. In total, 47 pairs of images were captured with the CD, and 49 pairs were captured with the RD. Illumination conditions were kept as constant as possible across subjects in each dataset.

C. Annotations: biomarkers and experts

Datasets from the CD and RD comprise white light images and corresponding fluorescent biomarker images. To ensure that the white light and biomarker images aligned with each other, we used a well-established image registration technique: a perspective transformation that minimizes the mean-squared error between the intensities of the images was applied to the biomarker images [20]. Binary pixel-level classifications of plaque and not-plaque were extracted by histogram thresholding the fluorescent biomarker images using empirically-determined thresholds for each dataset [2]. The devices are not guaranteed to capture all plaque in an image due to the absence of pophyrins in some plaque. Expert dental professionals independently annotated regions showing plaque on the white light images of teeth captured by both devices.

We then constructed union labels to represent the full extent of plaque detected by both the experts and fluorescent biomarker imaging, as shown in Fig. 1a. We believe this is a clinically valid approach because both expert labels and binary images capture independently verified signatures of plaque that are not-entirely mutually inclusive. A small percentage of the per-pixel plaque labels in each union label image were detected by both the expert and device, indicating distinct roles for each labeling method. These final pixel-level annotations contained in the union are the labels used to train and test the classifiers. Fig. 2 shows representative sets of images from both datasets.

D. Training data and test data

The pixel dimensions of the RD images were 512×384 while those of the CD images were 640×480. Accounting for the margins, each type of image is represented by 183,393 or 290,625 21×21 patches, respectively. A random sample of half the patches in each training image were used for training to limit overfitting on extremely similar patches, while all patches in test images were used for testing. Patches from a single image were not split among the train and test sets.

Each image was assigned to one of three groups based on the amount of plaque in the union label: low plaque, medium plaque, high plaque. We randomly assigned 70% of images from each group to the training set and the remaining 30% to the test set to ensure that the plaque quantity distribution of the training and test sets were roughly the same. Training only on images with high plaque and testing on images with low plaque, for example, would likely produce many false positives. For the CD dataset, $4,687,980$ patches from 33 images were used for training and $3,977,694$ patches from 14 images were used for testing. For the RD dataset, $3,209,360$ patches from 35 images were used for training and $2,750,895$ patches from 14 images were used for testing. Images were normalized in the RGB colorspace, and white light images in the RD dataset all received the same color balancing to account for variation in illumination. Each feature in the training set was standardized to have a zero mean and unit variance, and the transformations with the same parameters were applied to each feature in the test set.

The distribution of classes in the test data was skewed towards non-plaque by a ratio of approximately 9:1. Oversampling the plaque examples forced the training set to have an equal number of plaque and non-plaque examples [4]. While this negatively impacted precision and recall compared to training with a skew approximating that of the test data, it prevented the classifiers from simply learning the prior skew [21]. This was partially accounted for when calculating test accuracies by choosing a final threshold that minimized errors while weighting false positives more than false negatives.

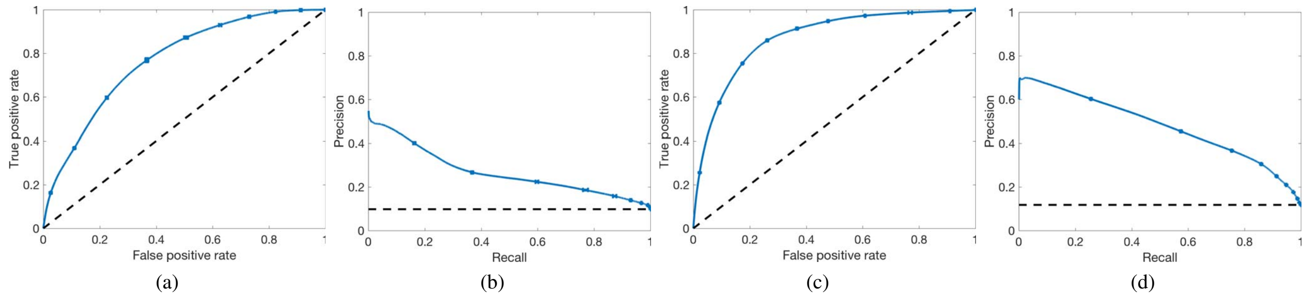


Figure 3: Threshold-averaged receiver operating characteristic (ROC) and precision-recall curves for both test sets. Random chance is shown as a dashed black line in all graphs. Error bars show 95% confidence intervals. (a) ROC curve for research device (RD) test set. Area under the curve (AUC) = 0.7694. (b) Precision-recall curve for RD test set. AUC = 0.2679 (c) ROC curve for commercial device (CD) test set. AUC = 0.8720. (d) Precision-recall curve for CD test set. AUC = 0.4784

IV. Results and Discussion

RD dataset: Fig. 3a shows the receiver operating characteristic curve (ROC) for the test set with an area under the curve (AUC) of 0.7694. Training ROC AUC was 0.8324 (data not shown). Fig. 3b shows the precision-recall curve. Training accuracy and test accuracy were 87.93% and 84.67%, respectively.

CD dataset: Fig. 3c shows the ROC for the test set with an area under the curve of 0.8720. Training ROC AUC was 0.8839 (data not shown). Fig. 3d shows the precision-recall curve. Training accuracy and test accuracy were 80.83% and 87.18%, respectively.

Both classifiers outperform chance, indicating that differences between plaque and non-plaque patches can be learned. Interpreting the ROC AUCs, the RD classifier has a 0.7694 probability of classifying a plaque example as more likely to be plaque than a non-plaque example, and the CD classifier has a probability for the same task of 0.8720. The CD classifier performed better than the RD classifier, especially with regards to the precision-recall curve. We believe this is due to a greater variation in illumination conditions for the white light images in the RD dataset.

The skew in the test sets can make the ROC curve overstate the test set performance. The precision-recall curves show that both classifiers have difficulty maintaining high precision while having high recall, as is expected when test sets are skewed [21]. When the test datasets were artificially made to have an equal number of positive and negative examples, the area under the RD precision-recall curve increased from 0.2679 (Fig. 3b) to 0.7394 (data not shown) and the area under the CD precision-recall curve increased from 0.4784 (Fig. 3d) to 0.8541 (data not shown). The test ROC AUCs remained relatively unchanged at 0.7574 and 0.8719, respectively (data not shown). The higher areas under the precision-recall curves for the balanced test sets indicate that test skew is responsible for the current precision-recall curves. Collecting more test data would not significantly decrease the skew, as it originates from the general level of plaque in the population.

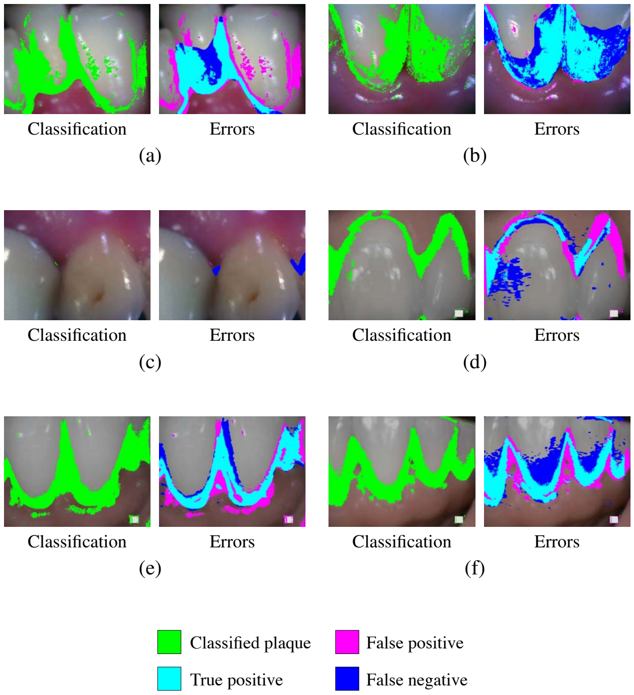


Figure 4: Classifications and errors overlaid on input white light images. Left of each pair: classifications overlaid on the input white light image. Right of each pair: Types of errors with respect to union labels overlaid on the input white light image. (a)-(c) Representative images from research device test set. (d)-(f) Representative images from commercial device test set.

Fig. 4 shows predictions and the type of errors seen from both classifiers. The classifier did not simply learn to classify margins of teeth. If that had been the case, we would have always expected a substantial number of false positives around the margins of teeth with little plaque, which is not

	Research device dataset			**Commercial device dataset**	
	AUC ± STD	Accuracy ± STD		AUC ± STD	Accuracy ± STD
	0.9822 ± 0.0025	96.25% ± 0.26%		0.9246 ± 0.0023	89.41% ± 0.06%
	0.9633 ± 0.0061	93.88% ± 0.24%		0.9625 ± 0.0018	93.22% ± 0.14%
	0.9621 ± 0.0200	94.34% ± 0.79%		0.9526 ± 0.0013	93.11% ± 0.20%
	0.9257 ± 0.0059	83.97% ± 1.60%		0.9028 ± 0.0041	90.08% ± 0.19%
	0.9222 ± 0.0290	93.40% ± 0.95%		0.8929 ± 0.0048	87.58% ± 0.26%
	0.8908 ± 0.0130	89.71% ± 0.77%		0.8780 ± 0.0011	88.44% ± 0.17%
	0.8224 ± 0.0410	87.63% ± 0.72%		0.8690 ± 0.0070	88.20% ± 0.16%
	0.8179 ± 0.0260	82.53% ± 1.30%		0.8653 ± 0.0016	79.61% ± 0.13%
	0.8087 ± 0.0270	92.46% ± 1.10%		0.8573 ± 0.0055	79.89% ± 0.30%
	0.7991 ± 0.0240	87.52% ± 0.69%		0.8358 ± 0.0171	92.83% ± 0.58%
	0.6936 ± 0.0190	85.93% ± 0.85%		0.8233 ± 0.0048	84.89% ± 0.25%
	0.6476 ± 0.0350	72.80% ± 0.77%		0.8060 ± 0.0011	82.26% ± 0.15%
	0.6238 ± 0.0088	91.54% ± 6.50%		0.7785 ± 0.0054	87.70% ± 0.51%
	0.5890 ± 0.0150	87.63% ± 0.88%		0.7100 ± 0.0025	85.98% ± 1.28%
Mean	0.8177 ± 0.1335	88.54% ± 6.11%	Mean	0.8613 ± 0.0683	87.37% ± 4.48%

Table I: Classification results for each image in the test sets. Each row contains area under the receiver operating characteristic curve (AUC) and accuracy for all the patches in one white light image. STD: standard deviation.

the case in Fig. 4c.

Table I splits the test results for both classifiers by image. The classifiers performed much better on some images than others because the plaque in those images more closely resembled the distributions of plaque in the training sets.

The classifiers are limited to predicting the types of plaque they have been conditioned upon in their training sets. More training data that fully captures the variety of plaque formations and illumination conditions will improve accuracy and robustness. Additional training data under various illumination conditions would likely improve performance, especially of the RD classifier. More experiments can be done toward improving the accuracy by increasing the patch size and complicating the neural architecture. Reformulating the patch-based CNN as an FCN may provide a significant decrease in both training and testing time at the expense of requiring more training data.

Because they were trained on the union labels, the classifiers learn more comprehensive signatures of plaque, encompassing both biomarker locations and expert annotations which do not have significant overlap. Comparison with previous plaque segmentation work is difficult due to this enlarged notion of plaque. Patch-based CNNs have been used to to predict expert labels or image-based labels, such as dyes, in specific medical imaging modalities, but they have not, to our knowledge, been trained on union labels.

V. Conclusion

We have shown two classifiers which successfully predict the location of dental plaque in white light dental images to a degree beyond that of chance. Our fully trained and validated CNN, after learning from both fluorescent biomarker images as well as expert labels, accepts standard white light intraoral images as inputs and predicts the location of plaque pixels with high sensitivity and specificity without requiring device or expert intervention. This approach can be extended to the numerous other conditions that can be detected with fluorescent biomarker imaging, such as oral cancer and early periodontal disease, as well as to non-oral domains where diagnostic biomarker imaging is used to augment expert knowledge.

Acknowledgments

The authors would like to thank Otkrist Gupta, Nikhil Naik, Tristan Swedish, and Matthew Tancik.

References

[1] L. J. Lesko and A. J. Atkinson Jr, "Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: criteria, validation, strategies." *Annual Review Of Pharmacology And Toxicology*, vol. 41, pp. 347 – 366, 2001.

[2] K. Angelino, P. Shah, D. A. Edlund, M. Mohit, and G. Yauney, "Clinical validation and assessment of a modular fluorescent imaging system and algorithm for rapid detection and quantification of dental plaque," 2017, under review.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[4] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge MA: MIT Press, 2016.

[5] 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," *arXiv preprint arXiv:1702.05747*, 2017.

[6] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2424–2433.

[7] 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.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[9] I. Pretty, W. Edgar, P. Smith, and S. Higham, "Quantification of dental plaque in the research environment," *Journal of dentistry*, vol. 33, no. 3, pp. 193–207, 2005.

[10] M. H. Van Der Veen, C. M. Volgenant, B. Keijser, M. Jacob Bob, and W. Crielaard, "Dynamics of red fluorescent dental plaque during experimental gingivitisa cohort study," *Journal of dentistry*, vol. 48, pp. 71–76, 2016.

[11] P. Rechmann, S. W. Liou, B. M. Rechmann, and J. D. Featherstone, "Performance of a light fluorescence device for the detection of microbial plaque and gingival inflammation," *Clinical oral investigations*, vol. 20, no. 1, pp. 151–159, 2016.

[12] J. Kang and Z. Ji, "Dental plaque quantification using mean-shift-based image segmentation," in *Computer Communication Control and Automation (3CA), 2010 International Symposium on*, vol. 1. IEEE, 2010, pp. 470–473.

[13] J. Kang, L. Min, Q. Luan, X. Li, and J. Liu, "Novel modified fuzzy c-means algorithm with applications," *Digital Signal Processing*, vol. 19, no. 2, pp. 309–319, 2009.

[14] K. Carter, G. Landini, and A. D. Walmsley, “Automated quantification of dental plaque accumulation using digital imaging,” *Journal of dentistry*, vol. 32, no. 8, pp. 623–628, 2004.

[15] B. Joseph, C. S. Prasanth, J. L. Jayanthi, J. Presanthila, and N. Subhash, “Detection and quantification of dental plaque based on laser-induced autofluorescence intensity ratio values,” *Journal of Biomedical Optics*, vol. 20, no. 4, pp. 048 001–048 001, 2015.

[16] A. Mansoor, V. Patsekin, D. Scherl, J. P. Robinson, and B. Rajwa, “A statistical modeling approach to computer-aided quantification of dental biofilm,” *IEEE journal of biomedical and health informatics*, vol. 19, no. 1, pp. 358–366, 2015.

[17] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[18] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting.” *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[19] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, “Tensorflow: Large-scale machine learning on heterogeneous distributed systems,” *arXiv preprint arXiv:1603.04467*, 2016.

[20] L. G. Brown, “A survey of image registration techniques,” *ACM computing surveys (CSUR)*, vol. 24, no. 4, pp. 325–376, 1992.

[21] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets,” *PloS one*, vol. 10, no. 3, p. e0118432, 2015.