Optimization methods for deep neural networks classifying OCT images to detect dental caries

Hassan S. Salehi,[1,*] Majd Barchini,[1] Mina Mahdian,[2]

[1]Department of Electrical and Computer Engineering, California State University, Chico, CA, USA
[2]School of Dental Medicine, Stony Brook University, Stony Brook, NY, USA
[*]hsalehi@csuchico.edu

ABSTRACT

Dental caries are common chronic infectious oral diseases affecting most teenagers and adults worldwide. Optical coherence tomography (OCT) has been studied extensively for the detection of early carious lesions. Deep learning techniques are a rapidly emerging new area of biomedical research and have yielded impressive results in diagnosis and prediction in the field of oral radiology. Deep learning models particularly deep convolutional neural networks (CNN) can be employed along with OCT imaging system to more accurately identify early dental caries. In this work, after OCT data acquisition, data augmentation was performed to obtain a large amount of training data in order to effectively learn, where collection of such training data is often expensive and laborious. For the backpropagation process, seven optimization methods, namely Adadelta, AdaGrad, Adam, AdaMax, Nadam, RMSProp, and Stochastic Gradient Descent (SGD) were utilized to improve the accuracy of a CNN classifier for diagnosing dental caries. In this study, 75% of the data were utilized for training and 25% for testing. The diagnostic accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and receiver operating characteristic (ROC) curve were calculated for detection and diagnostic performance of the deep CNN algorithm. This study highlighted the performance of various optimization methods for deep CNN models with OCT images to detect dental caries.

Keywords: optimization methods, image processing, machine learning, deep learning, convolutional neural networks, optical coherence tomography, dental caries detection.

1. INTRODUCTION

Early Detection of carious lesions can result in the implementation of non-surgical preventive approaches to reverse the demineralization process [1-3]. The conventional approach for diagnosing dental caries is clinical examination that supplemented by radiographs. However, studies based on the clinical and radiographic examination methods often show low sensitivity and high specificity. In the last few years, researches were studying a new approach to detect carious in early stage, because it is believed that by the time that a lesion is visualized in clinical or radiological examination, it will be advanced [3,4]. Optical coherence tomography (OCT) is a noninvasive imaging modality based on low-coherence interferometry that uses non-ionizing near-infrared laser to provide micrometer-resolution images. Previous studies using OCT have demonstrated the ability to evaluate characteristics of carious lesions, micro-fractures, pulpal inflammation, properties of dental materials, early dysplastic changes in oral malignancies, early inflammatory changes in the periodontal tissues, and PDL changes due to orthodontic tooth movement [5-15]. Deep learning is a rapidly emerging new area of biomedical research and have yielded impressive results in diagnosis and prediction in the fields of radiology and pathology. The deep learning employs computational models which are composed of a series of transforming and processing layers to learn representations of data with multiple levels of abstraction [16-18]. In deep learning, convolutional neural networks (CNN) [19] is the most commonly method applied to analyze medical imaging data and classify radiology images. Recent studies have demonstrated the CNN application for complex medical image analysis, such as automated

breast ultrasound lesions detection, classification of normal and pathologic OCT images, segmentation of nine retinal layer boundaries, cerebral microbleeds detection, and brain tumor segmentation [20-24].

In our previous publication [25-30], we presented OCT imaging for tissue characterization, as well as a novel approach combining OCT imaging modality and deep learning CNN model for the detection of occlusal carious lesions. In data acquisition and *ex vivo* OCT imaging, extracted human permanent teeth were collected and imaged. To the best of our knowledge, that study was the first one reporting deep learning-based classification of *ex vivo* OCT images of human carious and non-carious lesions for early detection of dental caries.

In this study, for the backpropagation process, seven optimization methods, namely Adadelta, AdaGrad, Adam, AdaMax, Nadam, RMSProp, and Stochastic Gradient Descent (SGD) were utilized to improve the accuracy of a CNN classifier for diagnosing dental caries. 75% of the imaging data were utilized for training and 25% for testing. The diagnostic accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and receiver operating characteristic (ROC) curve were calculated for detection and diagnostic performance of the deep CNN algorithm. This study highlighted the performance of various optimization methods for deep CNN models with OCT images to detect dental caries. The Adam, AdaMax, and Nadam optimizers provided the highest accuracy of (95.45%-97.12%), and (86.86%-88.73%) for training and testing, respectively.

2. METHODS

2.1 Experimental setup and image acquisition

The OCT imaging system that was used in this study was a spectral-domain OCT from the TELESTO-series (Thorlabs Inc., Newton, NJ, USA) as demonstrated in Fig. 1.

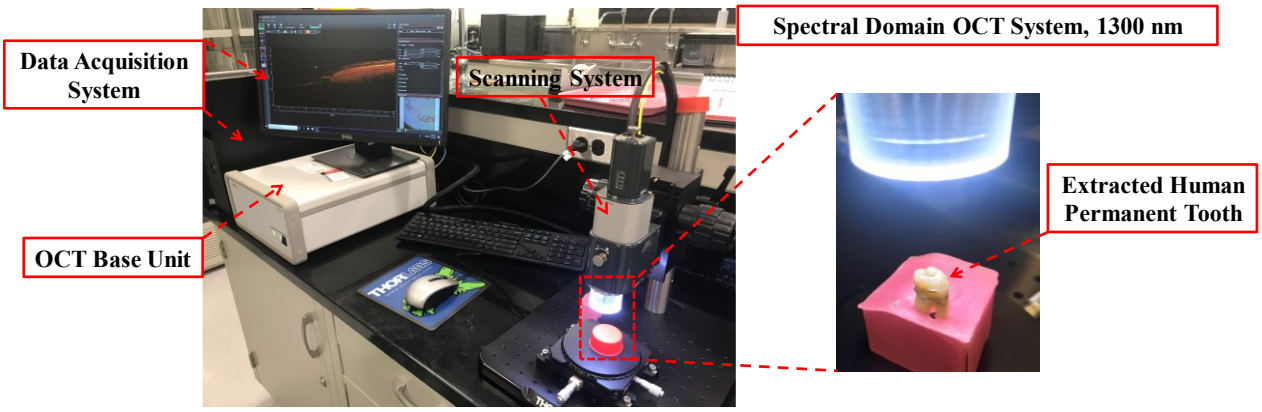

Figure 1 Data acquisition and OCT imaging system.

The OCT imaging system is operating at optical wavelengths ranging between 1300 nm and 1325 nm with an average power of 18 mW, a scan rate of 5.5-76 kHz, image depth of 3.5 mm in air, an OCT-LK4 objective, and axial resolution of 5.5 µm in air and 3.9 µm at n=1.4 [25]. For *ex vivo* OCT imaging, a total of 51 extracted human permanent teeth were collected and categorized into two groups: Non-carious teeth (NC), caries extending into enamel and dentin (C) as shown in Fig. 2. 782 cross-sections OCT images were obtained from the NC group, and 1357 images from the C group. The images were then subjected to machine learning using a CNN classifier.

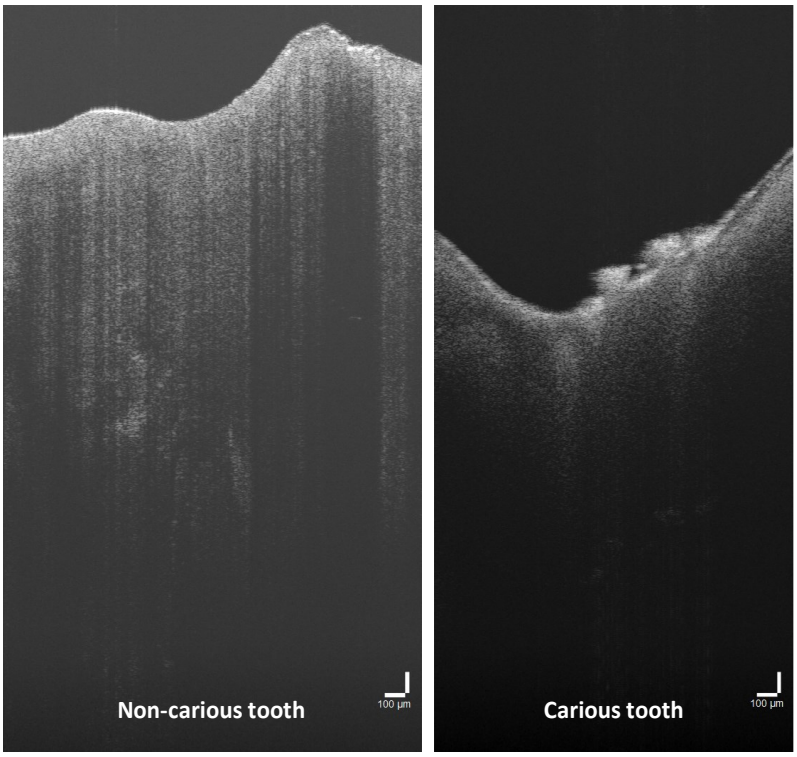

Figure 2 OCT images acquired from oral specimens.

2.2 Preprocessing and Image Augmentation

To acquire images with minimum inhomogeneity, imaging was performed multiple times at different points. OCT images with the least heterogeneous presentation were imported and saved in TIFF format. During preprocessing, the images were normalized to the size of 90 x 90 x 3 pixels and saved in the same format prior to training the CNN model. Deep artificial neural networks require a large amount of training data in order to effectively learn, where collection of such training data is often expensive and laborious. Data augmentation overcomes this issue by artificially inflating the training set with label preserving transformations [19]. We have applied image augmentation by perturbing an image using transformations that leave the underlying class unchanged (e.g. cropping and flipping) in order to generate additional examples of the class. Augmentation can be applied at training time, at testing time, or both. The augmented samples can either be taken as-is or combined to form a single feature, e.g. using sum/max-pooling or stacking.

2.3 Architecture of the Deep Convolutional Neural Network

Figure 3 demonstrates the CNN architecture that was developed in our previous publication [25] and was implemented in TensorFlow to compare the various optimization methods. Our CNN model contains two convolutional layers and three fully connected layers (FC). The first convolution layer contains 32 filters of size 3 x 3 followed by an activation layer of the rectified linear operator (ReLU), and a max pooling layer of size 2 x 2 and a local normalization layer. The output of the first convolution layer would be the input to the second convolution layer which has 256 filters of size 3 x 3 followed by ReLU and max pooling layer. The flatten applied to all feature's maps from the second convolution layers in one-dimensional vector. The first fully connected layer will have 128 neurons that will receive the one-dimensional vector from the second convolution layer, followed by a ReLU layer and dropout layer. The second FC has 128 neurons that receive the output from the first FC layer, followed by ReLU and dropout layer. The output layer with SoftMax function has 2 neurons that determine the output class (2 labels: non-carious and carious).

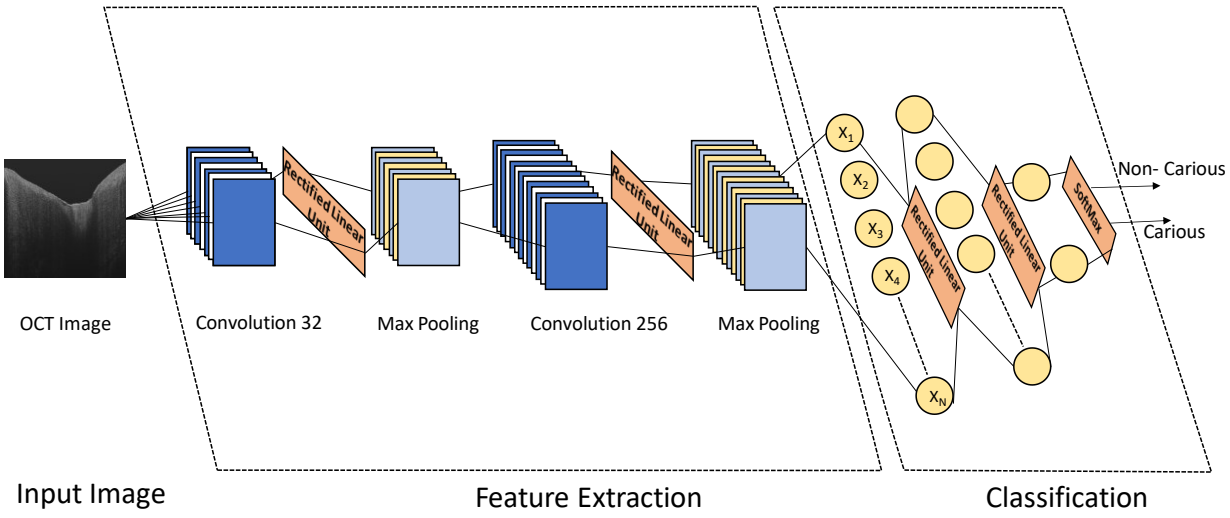

Figure 3 The CNN architecture for carious lesions detection.

2.4 Optimization Methods

For the backpropagation process, seven optimization methods were utilized which explained in the next subsections.

Adadelta Optimizer:

The Adadelta optimizer is a stochastic gradient descent (SGD) method base on per-dimension learning rate that doesn't required to have a manual selected global learning rate or continually decay learning rate through training [31]. The accumulation requires two steps, first is the accumulate of gradient descent square:

$$E[g^2]_t = \rho E[g^2]_{t-1} + (1-\rho) g_t^2 \qquad (1)$$

where ρ is the decay rate, and the initial accumulation variables is zero. The second step is to accumulate the updates:

$$E[\Delta x^2]_t := \rho E[\Delta x^2]_{t-1} + (1-\rho)\Delta x_t^2 \qquad (2)$$

where Δx_t^2 is the square of compute update and it is calculated:

$$\Delta x_t = -\frac{RMS[\Delta x]_{t-1}}{RMS[g]_t} g_t \qquad (3)$$

AdaGrad Optimizer:

AdaGrad optimizer is a parameter specific learning rate optimizer [32]. It updates FC weights per:

$$\theta_t := \theta_{t-1} - lr * \frac{g}{\sqrt{accum_{gt}+\epsilon}} \qquad (4)$$

where lr is the learning rate, g is the gradient, accum is the accumulator that calculated using

$$accum_{gt} := accum_{t-1} + g^2.$$

Adam Optimizer:

Adam optimizer is a combination of AdaGrad and RMSProp methods that is based on adaptive estimates of lower order moments [33]. The Adam optimizer update the FC layer's weights using:

$$\theta_t \leftarrow \theta_{t-1} - \frac{\alpha_t . m_t}{\sqrt{v_t}+\hat{\epsilon}} \qquad (5)$$

where m_t is the biased first moment estimate, v_t is the biased second raw moment estimate, α_t is the learning rate that calculated per:

$$\alpha_t \leftarrow \alpha.\frac{\sqrt{1-\beta_2^t}}{1-\beta_1^t} \quad (6)$$

where β_1, β_2 are the exponential decay rates for the moment estimate. From Eq. (5), the biased first raw moment updated per:

$$m_t \leftarrow \beta_1.m_{t-1} + (1-\beta_1).g_t \quad (7)$$

where g_t is the gradient with respect to stochastic objective at time t. From Eq. (5), the biased second raw moment updated per:

$$v_t \leftarrow \beta_2.v_{t-1} + (1-\beta_2).g_t^2 \quad (8)$$

AdaMax Optimizer:

AdaMax optimizer a variant of Adam based on the infinity order norm that makes the optimizer more stable [33]. It calculated the FC weight using:

$$\theta_t \leftarrow \theta_{t-1} - \frac{\frac{\alpha}{1-\beta_1^t}.m_t}{\mu_t} \quad (9)$$

where α is the step size, with β_1^t , β_1 is the exponential decay rates for the moment estimate denote to the power of t. The term $\frac{\alpha}{1-\beta_1^t}$ is the learning rate with the bias-correction term of the first moment. m_t is the biased first raw moment updated per Eq. (7). μ_t is the exponentially weighted infinity norm per $\mu_t \leftarrow \max(\beta_2.\mu_{t-1}, |g_t|)$.

Nadam Optimizer:

Nadam optimizer is similar to Adam optimizer, the difference is that the Nadam optimizer use Nesterov momentum [34]. The Nadam optimizer update the FC layer's weights per:

$$\theta_t \leftarrow \theta_{t-1} - \eta\frac{\bar{m}_t}{\sqrt{\hat{n}_t}+\epsilon} \quad (10)$$

where the $\bar{m}_t$ is the gradient update for current time step using $\bar{m}_t \leftarrow (1-\mu_t)\hat{g}_t + \mu_{t+1}\hat{m}_t$, η is the learning rate, $\hat{n}_t$ is the second moment vector estimator that calculated per $\hat{n}_t \leftarrow \frac{n_t}{1-v^t}$

RMSProp Optimizer:

RMSProp optimizer divide the learning rate for a weight by the average magnitudes of recent gradients for that weight [35]. It keeps a moving a mean average of the squared gradient for each weight by dividing the gradient by square root of mean square (MS). The mean square equation:

$$MS(w,t) = 0.9MS(w,t-1) + 0.1\left(\frac{\partial E}{\partial w}(t)\right)^2 \quad (11)$$

Stochastic Gradient Descent (SGD) Optimizer:

SGD with momentum optimizer with Nestrov's accelerated Gradient update weight per:

$$\theta_{t+1} = \theta_t + v_{t+1} \quad (12)$$

where v_t is the velocity vector that will be calculated by $v_{t+1} = \mu v_t - \varepsilon \nabla f(\theta_t + \mu v_t)$, ε is the learning rate, μ is the momentum coefficient, and $\nabla f(\theta_t + \mu v_t)$ is the gradient.

3. RESULTS

3.1 Training and testing the CNN classifier

Briefly, the training set is split into mini-batches, with 10 images per batch. Given a batch of training patches, the CNN uses two convolution and two pooling layers to extract features and then classify each patch based on the probabilities from the SoftMax classification layer. After that, the CNN calculates the error between the classification result and the reference label, and then utilizes the backpropagation process [36] to tune all the layer weights to minimize this error using seven optimizers that we have utilized in this study. The above process will be repeated several epochs, until the whole CNN model becomes convergent for each optimizer. In this study, 75% of the imaging data were utilized for training and 25% for testing. Note that, all neurons in layers used a Rectified Linear Unit [37] with the weights initially being initialized from a Gaussian distribution with a 0 mean and a standard deviation of 0.01. Overlapping pooling was deployed which increased CNN performance by reducing over-fitting [19].

3.2 Deep Learning Classification

The diagnostic accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and receiver operating characteristic (ROC) curve were calculated for detection and diagnostic performance of the deep CNN model with OCT images. Table 1, and table 2 compare the results of using seven optimizers for training and testing respectively. Figure 4 shows the Receiver Operating Characteristic (ROC) curves for comparison between seven optimizers (a) training and (b) testing. The Adam, AdaMax, and Nadam optimizers provided the highest accuracy of (95.45%-97.12%), and (86.86%-88.73%) for training and testing, respectively.

Table 1 Training Results.

Optimizer	Accuracy	Sensitivity	Specificity	PPV	NPV
AdaDelta %	66.19	16.63	94.64	64.00	66.42
AdaGrad %	82.92	67.46	91.81	82.53	92.10
Adam %	96.75	93.85	98.42	97.14	96.53
AdaMax %	95.45	89.55	98.84	97.79	94.28
Nadam %	97.12	95.07	98.29	96.96	97.20
RMSProp %	82.07	67.58	90.39	80.14	82.93
SGD %	74.92	51.03	88.63	72.02	75.93

Table 2 Testing Results.

Optimizer	Accuracy	Sensitivity	Specificity	PPV	NPV
AdaDelta %	66.54	17.21	95.02	66.67	66.53
AdaGrad %	81.60	65.06	91.15	80.84	81.88
Adam %	86.86	75.79	93.26	86.66	86.96
AdaMax %	88.73	75.78	96.21	92.03	87.31
Nadam %	88.70	81.25	93.01	87.03	89.57
RMSProp %	79.87	63.17	89.51	77.67	80.80
SGD %	75.03	51.42	88.67	72.82	75.97

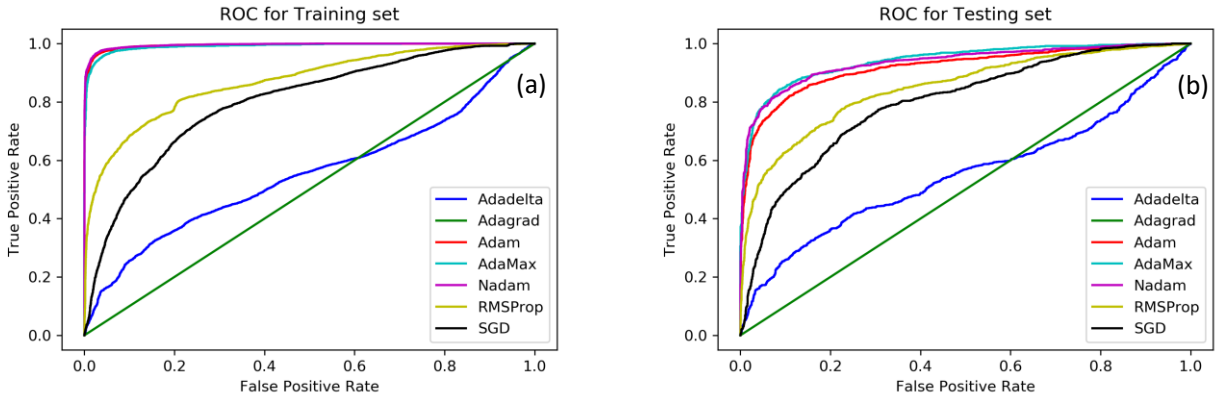

Figure 4 ROC Curves for training and testing using seven optimizers.

4. CONCLUSIONS

Deep learning is a rapidly emerging new area of biomedical research. The deep learning employs computational models which are composed of a series of transforming and processing layers to learn representations of data with multiple levels of abstraction. The deep learning techniques can be used to supplement optical coherence tomography (OCT) to more accurately identify diseased and damaged tissue. In this work, the CNN model was used to classify OCT images into non-carious and carious classes using seven optimization methods. In data acquisition and *ex vivo* OCT imaging, the samples were imaged using spectral-domain OCT imaging system operating at 1300 nm center wavelength with a scan rate of 5.5-76 kHz, and axial resolution of 5.5 µm in air. For deep learning, OCT images of extracted human carious and non-carious teeth were input to a CNN classifier to determine variations in tissue densities reflecting the demineralization process. The CNN model employs two convolutional and pooling layers to extract features and then classify each patch based on the probabilities from the SoftMax classification layer. The CNN calculates the error between the classification result and the reference label, and then utilizes the backpropagation process to tune all the layer weights to minimize this error using optimization algorithms. Seven optimization methods in TensorFlow with a learning rate of 0.001 were investigated and compared. In this study, 75% of the data were utilized for training and 25% for testing. The diagnostic accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and receiver operating characteristic (ROC) curve were calculated for detection and diagnostic performance of the deep CNN

algorithm. The Adam, AdaMax, and Nadam optimizers provided the highest accuracy of (95.45%-97.12%), and (86.86%-88.73%) for training and testing, respectively.

5. ACKNOWLEDGMENTS

This research was partially supported by College of ECC at California State University, Chico.

REFERENCES

[1] N. B. Pitts, "Are we ready to move from operative to non-operative/preventive treatment of dental caries in clinical practice?," *Caries Res.* 38, 294-304 (2004).
[2] JDB Featherstone, "Dental caries: a dynamic disease process," *Austr Dent J* 53 (3), 286-291 (2008).
[3] L. Karlsson "Caries detection method based on changes in optical properties between healthy and carious tissue," *Int J Dent* 2010, 1-9 (2010).
[4] R. H. Selwitz, A. I. Ismail, and N. B. Pitts, "Dental caries," *The Lancet* 369 (9555), 51-59 (2007).
[5] A. Calantog, L Hallajian, T Nabelsi, et al. "A prospective study to assess in vivo optical coherence tomography imaging for early detection of chemotherapy-induced oral mucositis," *Lasers Surg Med.* 45:22-27 (2013).
[6] JH. Baek, J Na, BH Lee, E Choi, and WS Son, "Optical approach to the periodontal ligament under orthodontic tooth movement: a preliminary study with optical coherence tomography," *Am J Orthod Dentofacial Orthop.* 135:252-259 (2009).
[7] EZ. Alsayed, I Hariri, A Sadr, et al. "Optical coherence tomography for evaluation of enamel and protective coatings," *Dent Mater J.* 34:98-107 (2015).
[8] Y. Nakajima, Y Shimada, A Sadr, et al. "Detection of occlusal caries in primary teeth using swept source optical coherence tomography," *J. Biomed Opt.* 19:16020 (2014).
[9] Y. Shimada, H Nakagawa, A Sadr, et al. "Noninvasive crosssectional imaging of proximal caries using swept-source optical coherence tomography (SS-OCT) in vivo," *J. Biophotonic* 7:506-513 (2014).
[10] I. Wada, Y Shimada, M Ikeda, et al. "Clinical assessment of noncarious cervical lesion using swept-source optical coherence tomography," *J. Biophotonics* 8:846-854 (2014).
[11] K. Imai, Y Shimada, A Sadr, Y Sumi, and J Tagami, "Noninvasive cross-sectional visualization of enamel cracks by optical coherence tomography in vitro," *J. Endod.* 38:1269-1274 (2012).
[12] YC. Ahn, J Chung, P Wilder-Smith, and Z Chen, "Multimodality approach to optical early detection and mapping of oral neoplasia," *J. Biomed Opt.* 16:076007 (2011).
[13] M. DeCoro, and P Wilder-Smith, "Potential of optical coherence tomography for early diagnosis of oral malignancies," *Expert Rev Anticancer Ther.* 10:321-329 (2010).
[14] H. Kawakami-Wong, S Gu, MJ Hammer-Wilson, JB Epstein, Z Chen, and P Wilder-Smith, "In vivo optical coherence tomography based scoring of oral mucositis in human subjects: a pilot study," *J. Biomed Opt.* 12:051702 (2007).
[15] ES. Matheny, NM Hanna, WG Jung, et al. "Optical coherence tomography of malignancy in hamster cheek pouches," *J. Biomed Opt.* 9:978-981 (2004).
[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature* 521(7553), 436-444 (2015).
[17] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science* 313(5786), 504-507 (2006).
[18] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.* 18(7), 1527–1554 (2006).
[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Proceedings of the Advances in Neural Information Processing Systems, (MIT Press, 2012), pp. 1097–1105.
[20] M. H. Yap, G. Pons, J. Mart´ı, S. Ganau, M. Sent´ıs, R. Zwiggelaar, A. K. Davison, and R. Mart´, "Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks," IEEE Journal of Biomedical and Health Informatics 22(4), 1218–1226 (2018).

[21] S. P. K. Karri, D. Chakraborty, and J. Chatterjee, "Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration," *Biomed. Opt. Express* 8(2), 579–592 (2017).
[22] L. Fang, D. Cunefare, C. Wang, R. H. Guymer, S. Li, and S. Farsiu, "Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search," *J. Bio. Optics Express* 8(5), 2732-2744 (2017).
[23] Qi Dou, Hao Chen, Lequan Yu, Lei Zhao, Jing Qin, V. C. Defeng Wang, Mok, Lin Shi, and Pheng-Ann Heng, "Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks," *IEEE Trans. Med. Imaging* 35(5), 1182–1195 (2016).
[24] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," *IEEE Trans. Med. Imaging* 35(5), 1240–1251 (2016).
[25] S. Nandy, H. S. Salehi, T. Wang, X. Wang, M. Sanders, A. Kueck, M. Brewer, and Q. Zhu, "Correlating optical coherence elastography based strain measurements with collagen content of the human ovarian tissue," Journal of Bio. Optics Express 6(10), 3806-3811 (2015).
[26] M. Mahdian, H. S. Salehi, A. G. Lurie, S. Yadav, A. Tadinada, "Tissue characterization using optical coherence tomography and cone beam computed tomography: A comparative pilot study," Elsevier Journal of Oral Surg. Med. Path. Radiology 122(1), 98-103 (2016).
[27] H. S. Salehi, M. Mahdian, H. Alnajjar, and A. Tadinada, "Utilizing Optical Coherence Tomography and Cone Beam Computed Tomography for Oral Tissues Characterization: Ex vivo Study," OSA International Biomed. Optics Congress, JTu3A.52 (2016).
[28] H. S. Salehi, A. Kosa, M. Mahdian, S. Moslehpour, H. Alnajjar, and A. Tadinada, "Characterization of human oral tissues based on quantitative analysis of optical coherence tomography images," Proc. SPIE BIOS 10044, Lasers in Dentistry XXIII, 1004406 (2017).
[29] N. Karimian, H. S. Salehi, M. Mahdian, H. Alnajjar, and A. Tadinada, "Deep learning classifier with optical coherence tomography images for early dental caries detection," Proc. SPIE BIOS 10473, Lasers in Dentistry XXIV, 1047304 (2018).
[30] H. S. Salehi, M. Mahdian, M. M. Murshid, S. Judex, and A. Tadinada, "Deep learning-based quantitative analysis of dental caries using optical coherence tomography: an ex vivo study," Proc. SPIE BIOS 10857, Lasers in Dentistry XXV, 108570H (2019).
[31] M. D. Zeiler, "Adadelta: an adaptive learning rate method," arXiv preprint arXiv:1212.5701 (2012).
[32] J. Duchi, E. Hazan, Y. Singer, "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization," Journal of Machine Learning Research 12 2121-2159 (2011).
[33] D. P. Kingma, J. Lei Ba, "Adam: A Method for Stochastic Optimization," ICLR (2015).
[34] T. Dozat "Incorporating Nesterov Momentum into Adam"
[35] G. Hinton "Neural Networks for Machine Learning -Lecture 6a- Overview of Mini-Batch Gradient Descent" (2012).
[36] C. M. Bishop, "Pattern Recognition and Machine Learning," Springer, New York, NY, (2006).
[37] R. H. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung, "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit," *Nature 405*(6789), 947-951 (2010).