Data Science and Engineering (DSE) Record, Volume 3, issue 1.

Detecting Caries Lesions with Bitewing Radiograph Using Ensemble of Convolutional Neural Network Model

Chawalit Chanintonsongkhla [1] and Varin Chaovatut [2,*]

[1] Double Master's Degree Program in Data Science and General Dentistry,
Chiang Mai University, Chiang Mai, Thailand
chawalit_chanin@cmu.ac.th
[2] Department of Computer Science, Faculty of Science,
Chiang Mai University, Chiang Mai, Thailand
varin.ch@cmu.ac.th

Abstract. This independent study aims to develop a model for segmenting proximal dental caries using a fully convolutional neural network in bitewing radiographs. The segmentation models were created with the explicit goal of helping dentists in segmenting dental caries in radiographs for a second opinion. To determine the most appropriate model architecture, we compared the performance of three fundamental segmentation models: U-Net, FPN (Feature Pyramid Network), DeepLabV3+, and XsembleNet, a combination of the three preceding models. The system is evaluated in two ways. The first is to assess segmentation quality using the dice coefficient; empirical experiments indicate that XsembleNet has the highest dice coefficient, followed by FPN. The second evaluation is to rate models' 12 testing bitewing radiographs segmentation. While all four models are comparable in accuracy and specificity, XsembleNet and FPN jointly achieve the highest classification metrics score. As a result, it can be concluded that a fully convolutional neural network could be used to detect dental proximal caries radiographs via computer-assisted diagnosis.

Keywords: Artificial intelligence, Semantic segmentation, Fully convolution network, Dental Caries, Dental radiograph

1 Introduction

Dental decay is a common disease affecting hundreds of millions of people worldwide. Having a carious tooth is known to affect health and quality of life. Dental caries causes the loss of mineral composition. If these lesions are not well-managed, a further mineral loss will lead to a breakdown of tooth structure and appear as a cavity [1].

Managing carious lesions requires early detection, proper diagnosis, and appropriate treatment [2]. Treatment of tooth cavities is typically more invasive as it progresses. Such as filling, pulpal treatment, or even extraction, it depends on the lesion's extent. Thus, Early detection enables intercept before the problem could become more severe over time.

* Corresponding author.

In the standard clinical setting, a dentist examinates teeth using visual-tactile inspection in conjunction with taking a bitewing radiograph. Bitewing radiographs are a complementary method for detecting proximal and occlusal cavities in teeth that visual examination alone might be unable to see [3].

However, there is a variance in the individual dentist's diagnosis. Table 1 shows that the level of conformity among the examiners is between moderate to substantial [4]–[6]. Different staging could result in varied treatment decisions as individual dentists have treatment strategies. This variability arose even in operative dentistry teachers, who lack standardized criteria for treatment decisions in operative dentistry. Some will start restoring teeth early, while others will monitor progression to some extent first [7].

Table 1. Kappa coefficient and level of agreement between the examiners

Author	Radiographic system	Kappa Coefficient	Level of agreement
Valachovic et al. 1986 [4]	Conventional	0.680 - 0.800	Substantial
Langlais et al. 1987 [5]	Conventional	0.565 - 0.599	Moderate
Naitoh et al. 1998 [6]	Conventional	0.424	Moderate
	Digital	0.439	Moderate

Computer-aided assistance (CAA) systems for dental radiography images have recently become an essential topic of study. CAA systems may aid dentists in making a more consistent and accurate diagnosis of dental caries in bitewing radiography images. Caries segmentation, i.e., classifying whether each point, specifically pixel, in a radiograph has caries or not, is considered a semantic segmentation task. Such segmentation tasks can be automated using deep learning, a branch of machine learning that excels on high-dimensional data such as text and images [8].

A fully convolutional network (FCN) is one of the deep learning model architectures capable of doing a segmentation task and found success in segmenting medical images. Previous works demonstrated that custom-made FCN and U-Net, a specific type of FCN model, could provide a consistent and accurate result [9], [10]. At present, There are many types of FCN models available. Our contribution is investigating different models and constructing an ensemble of several models to provide more reliable prediction results.

2 Literature Review

2.1 Segmentation models: U-Net, FPN, and DeepLabV3+

Long originally introduced FCN in 2014 [11]. FCN does not have a flattening and dense (fully-connected) layer and relies on only convolution operation. Most convolution neural networks used to perform computer vision tasks are similar. Extract feature maps with a backbone (feature extraction network) and then pass them forwards to

specific networks capable of using those features. This article will refer to each network as a head, and encoder head is a synonym for a network backbone.

The Segmentation model can be separated roughly into two principal heads — The encoder head and the decoder head. The encoder head's role is to extract a feature, much like a typical deep convolution network. Feature maps in deep layers contain robust features but low spatial information. Contrasting with the features in shallow layers were weak but rich in spatial information. The decoder head can be considered an "inverse" process of the encoder. It upsamples low dimension feature maps into a larger one. However, upsampling from deep feature maps alone suffers from loss of segmentation detail, so the connector is attached to the encoder head for transferring spatial data [11], [12].

The output of the decoder head is multiple feature maps with their dimensions same input image. Lastly, convolution with a kernel size of one pixel was done and returned as a prediction result. Because constructing an encoder head can be derived from a competent image classification model such as ResNet [13], Our primary interest is the decoder head, as the different approaches could provide pristine segmentation results. There are three variations of FCN models in our scope of work. (1) U-Net implements cascade upsampling with fractionally stridden convolution and skip-connector [12]. (2) Feature pyramid networks (FPN) are convolute and decoded at multiple dimensions [14]. (3) DeepLabV3+ probe into multiple feature maps with dilated convolution delivers a broader field of view [15].

2.2 Ensemble of Semantic Segmentation Models

Predictions from several models and techniques can improve efficiency and reduce prediction variability. There are several ways to ensemble prediction results, including averages, votes, and handcrafted machine learning algorithms [16]. The main disadvantage of using the ensemble model is expensive in terms of both time and computation cost and less interpretable as separate components.

Thambawitaa et al. applied an ensemble model concept to segment colon polyps in the EndoCV2021 challenge consisting of two rounds [17]. TriUNet consists of three U-Net models arranged in a triangular form which is the best performer in the first round. The model performs simultaneous learning of all three submodels, and loss values are passed back propagation to all three submodels.

In the second race, Thambawitaa created a series of models. DivergentNets consists of 5 sub-models: UNet++, FPN, TriUNet, Deeplabv3, and Deeplabv3+, but there is a difference from the first one. Each submodel learns and predicts separately, and the final prediction result obtains by using averaged results. DivergentNets were also the best model in the second round. When comparing the accuracy metrics among its sub-models, DivergentNets got better results.

Thambawitta's methods demonstrate two approaches. One is increasing the model components. Another is using majority voting, much like Condorcet's jury theorem that shows a large group is more likely to be correct than a decision attained by a single expert. However, training an end-to-end arbitrary spacious model is impossible as it

will approach physical limitations. Although not mentioned, we suspected this is one of the reasons why Thambawitta trained submodel separately in DivergenNets.

2.3 Deep learning in Dental Caries Segmentation

Deep learning is an emerging technology and influenced the health care field, including dentistry, specifically dental caries. One of the earlier implementations is to classify an image whether it has dental caries or not. Lee studied the efficacy of deep convolution neural networks, namely Inception v3, which train on over 3000 periapical radiographs [18]. The model can diagnose a cropped image of a tooth showing a promising result and may be helpful in clinical practice.

Such classification tasks can only to categorized images into specific classes. Segmentation tasks could provide additional diagnostic information because they could provide a lesion outline, much like clinical diagnostic criteria, which accounting structures in a radiograph. For example, the lesion that reaches the middle half of the dentin is in the moderate stage and requires immediate treatment [19].

Srivastava et al. developed a custom-made FCNN with over 100 layers to segment dental cavities in the bitewing radiograph with 3000 images as training materials [10]. This model achieved high recall (80.5) and moderate sensitivity (61.5), implying that it could not detect only a few caries but still had ambiguous false-positive results. When compared to dentists, the model outperforms by a large margin.

Recently, Cantu et al. applied a particular type of FCN network, U-Net [9]. U-Net was first published by Ronneberger et al. and found to be successful and widely applied in medical imaging [12]. The model trained on 3,686 bitewing radiographs. The caries detection model is robust against initial and advanced caries in contrast to the dentist, where most have low sensitivity to the initial lesion. In terms of metrics, it showed much higher accuracy and sensitivity overall than the dentists' mean, while its specificity remains subpar to dentists.

Table 2. Metrics from previous works

Author	FCN architectures	Evaluation metrics	Score	Dentists' score (Mean, [Min-Max])
Srivastava et al. 2017 [10]	Handcrafted FCN (100+ layers)	Recall	0.80	0.41 [0.34-0.47]
		Precision	0.61	0.77 [0.63-0.89]
		F1-score	0.70	0.53 [0.50-0.56]
Cantu et al. 2020 [9]	U-Net with EfficientNet-B5 as backbone	Accuracy	0.80	0.71 [0.61-0.78]
		Sensitivity	0.75	0.36 [0.19-0.65]
		Specificity	0.83	0.91 [0.69-0.98]
		PPV	0.70	0.75 [0.41-0.88]
		NPV	0.86	0.72 [0.68-0.82]
		F1 score	0.73	0.41 [0.26-0.63]
		MCC	0.57	0.35 [0.14-0.51]

Table 2 shows that the deep learning models can produce satisfactory results, yield high sensitivity, correctly interpret results, and are consistent with the reference set.

However, even though many segmentation models are available nowadays, no efficacy comparison has been made in the dental caries section. This importance led us to discover appropriate model architecture for the caries segmentation task and develop a better model for assisting in treatment planning and a tool to reduce dentist variance.

3 Data and Methodology

3.1 Data

We purposively selected and obtained dental bitewing radiographs from various sources to construct a small-scale pilot study test set. The requirement is that teeth appear in the adolescent age group. Assuming the age range is about 10-30 years old, with a general appearance that they usually have many cavities, normal morphology, and do not exhibit signs of tooth wear.

The dataset contains 326 images divided into 270 radiographs with at least one caries site and 56 radiographs without caries with different tooth features and caries lesions. The researcher annotated the ground truth data set using the Labelme application, which yielded 657 cavities.

3.2 Methodology

This study is to train and evaluate the segmentation models. All models were constructed using PyTorch and trained with Google Colab Pro. Colab Pro consists of a Tesla T4 GPU with 16 GB of memory. The baseline models were used to benchmark including U-Net , FPN , DeepLabv3+. Then, the ensemble model name XsembleNet was constructed from three baseline models' components. There are two stages of experimenters. Initially, train each model and compare quantitative metrics derived from the pixel-wise loss function. Afterward, interpret at a tooth level to assess the quality and characteristics of segmentation for each model. The experimental setup, including data preparation, training techniques, and modeling of deep learning networks, is described in this part.

1) Baseline Models Construction (U-Net, FPN, and DeepLabV3+)

The baseline models were constructed using the implementations and pre-trained weights supplied in the *Segmentation Models* [20] library. These networks served as the foundation for our planned XsembleNet as well. The encoder of each model was ResNet34 [13], which was initialized with ImageNet weights. The number of encoder parameters is 21.2 million for all three models. The total parameter of the decoder part of U-Net, FPN, and DeepLabV3+ are 3.1 million, 1.8 million, and 1.1 million, respectively. In most cases, the decoder is considerably smaller than the encoder.

2) Construction of Ensemble Model With Three Decoders

In the work of Thambawitta, DivergentNets is an ensemble of 5 different parallel pre-trained baseline models [17]. The prediction result of DivergentNets is obtained by averaging each model's result. Each sub-model is frozen and does not have a connection between models in the training process. Even though all models' encoders are the same, they can not be used interchangeably as the value of the encoder's weights, and biases are not the same. Hence, these components are repetitive, inefficient, and lack interaction between models, which could provide more information to enable more accurate predictions.

Our contribution, XsembleNet, remedial repetitive components and takes advantage of that decoder size is relatively small compared to the encoder. When training models such as multiple sub-model in one go, the model becomes large and could reach memory constraints, specifically, GPU memory. By creating a shared encoder, Model size will be lessened, and the encoder is trained with a different gradient from multiple decoders.

However, some modifications in the encoder should be made as desirable feature maps size is not all-purpose to every decoder. The deep layer of DeepLabV3+ decoder takes constant feature maps' size for dilated convolution. Constant feature maps size is different from U-Net and FPN decoder. The latter two models take progressive downsampling feature maps throughout every depth. In such wise, we designed a Y-shaped encoder where its stem is a common encoder path of three decoders and then separated into two branches. One is for DeepLabV3+ decoder, and the second is for U-Net and FPN decoder. Using shared components can reduce the model size by 41 percent. It provides efficient use of encoder, enabling training larger model and taking less computational cost.

3) Training of Segmentation Models

The available data for training is small-scale, and its amount is about one-tenth of previous studies [9], [10]. In order to avoid overfitting and increase the diversity of data available for the training model, The data is augmented by adjusting geometric transformation (horizontal flip) and altering pixels' intensity (brightness, contrast, and gaussian-blur). The data is pre-separated into training and validation sets with a proportion of 7:3 to ensure all models will be trained and validated with the same data.

All models are trained with the same configuration to minimize a loss function using the Adam optimizer for 90 epochs. They were starting with a learning rate of 5e-3. For every 30 epochs passed, the learning rate will be reduced by a factor of ten. The loss function is a combo loss between binary cross-entropy loss (BCELoss) and Dice loss, defined as equation (1-3) [21]. BCELoss and Dice loss will contribute to distribution and region training, respectively. In this case, the data is skewed because the positive pixels are overwhelmed by negative ones. BCELoss is multiplied (α) by a factor of 5, which will bring the BCELoss to the same level as Dice Loss.

$$BCELoss(Y,\widehat{Y}) = -(ylog(\hat{y}) + (1-y)log(1-\hat{y})); \ y \epsilon Y, \ \hat{y} \epsilon \widehat{Y} \quad (1)$$

$$DiceLoss(Y,\widehat{Y}) = 1 - \frac{2\times|Y\cap\widehat{Y}|}{|Y|+|\widehat{Y}|} \quad (2)$$

$$ComboLoss(Y,\widehat{Y},\alpha) = \alpha BCELoss(Y,\widehat{Y}) + DiceLoss(Y,\widehat{Y}) \quad (3)$$

4) Model Evaluation

The model is evaluated in two ways. The first is to assess segmentation quality at the pixel level using the segmentation metrics: BCELoss, Dice loss, and Combo loss. The second evaluation is to rate models' segmentation of 12 additional testing bitewing radiographs to analyze at the tooth level. For the second evaluation, Prediction results are from the model state that achieved the lowest Dice loss as it correlates to the dentists' needs—segmenting the tooth decay area as accurately as possible.

4 Results

4.1 Quantitative Result

Table 3 shows losses of three baseline models, one ensemble model named XsembleNet. It was found that U-Net achieved the lowest BCELoss and Combo loss among all models. While FPN has the lowest dice loss among the three baseline models, XsembleNet's yielded lower dice loss than FPN.

Table 3. Metrics from previous works

Model	Best Validation Loss (at epoch)		
	Binary cross entropy loss	Dice loss	Combo loss
U-Net	**0.0269 (30)**	0.5031 (77)	**0.6342 (85)**
FPN	0.0289 (24)	0.4830 (51)	0.6651 (51)
DeepLabV3+	0.0344 (12)	0.6648 (71)	0.8209 (71)
XsembleNet	0.0291 (17)	**0.4691 (52)**	0.6877 (52)

If inspecting the plot of validation combo loss shown in Figure 1, all performance metrics increased over the number of epochs until converging. However, in the different models provided, only DeepLabV3+ could not stabilize the training process as depicted by wiggle and fluctuated loss during the entire run.

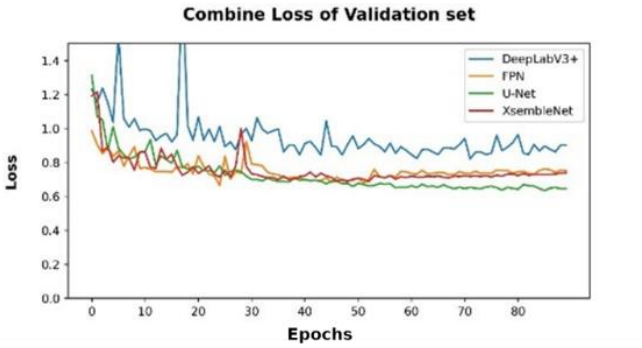

Fig. 1. Validation combo loss per epoch

4.2 Qualitative Result

Each model at the lowest dice score state is then inference and analyzed in tooth-level prediction, given an example in Figure 2. By merging ground truth and inference results, three distinguished areas could differentiate into true positive (TP), false positive (FP), and false negative (FN) findings. Lastly, true negatives (TN) are teeth free of either ground truth or prediction area.

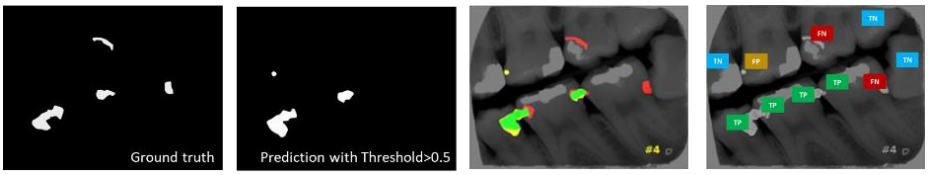

Fig. 2. Tooth-level interpretation

The testing set contains 12 additional bitewing radiographs with 24 cavities distributed over 77 teeth. We constructed confusion matrics and then derived four metrics to evaluate model efficacy: accuracy, F1-score, precision, and sensitivity, as shown in Table 4.

Table 4. Classification metrics and interference time from tooth-level interpretation

Model	Testing set classification metrics				Total CPU inference time
	Accuracy	F_1-Score	Precision	Sensitivity	
U-Net	0.9307	0.8571	0.8400	0.8750	08.38s
FPN	**0.9680**	**0.9200**	**0.8846**	**0.9580**	**07.05s**
DeepLabV3+	0.8713	0.6977	0.7895	0.6250	07.91s
XsembleNet	**0.9680**	**0.9200**	**0.8846**	**0.9583**	15.52s

Out of the four models, the XsembleNet and FPN models can detect cavities at the highest accuracy, 23 out of 24 positions, with three false positives. The U-Net model predicted cavities at 21 locations and four false positives. DeepLabV3+ could predict only 15 cavities and four false positives.

XsembleNet and FPN had the highest scores on all metrics as both models were more accurate in predicting and producing fewer false positives. The U-Net model was placed after those two because cavities were not detected in some areas. In contrast, the DeepLabV3+ had the lowest metrics score as the smallest amount of cavities can be seen and still have false positive results.

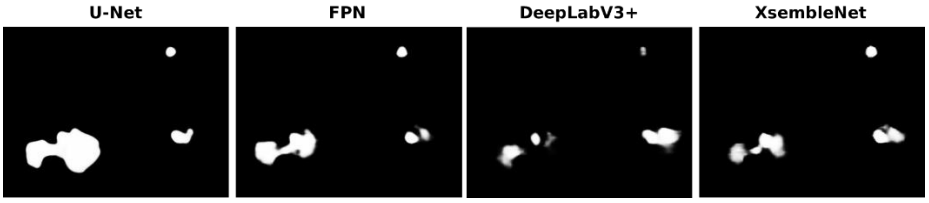

Fig. 3. Pre-thresholded prediction results

We further looked into segmentation characteristics for each model from pre-threshold prediction results. It found that each model performed differently, as shown in Figure 3. Generally, U-Net produces a well-defined, slightly rounded border but does not adapt to the concaved areas as well as FPN could. DeepLabV3+ is unlike those two as it makes numerous ambiguous patches that appear to be blurry areas, typically below the threshold value (0.5). All models except U-Net have blurry characteristics. We suspect that multi-level prediction and dilation convolution are the leading causes.

Our implementation, XsembleNet, has multiple traits from its components. Its prediction results are slightly blurred but much less than DeepLabV3+ results. We found that XsembleNet results are homogenous to the combination of baseline models illustrated in Figure 4, where each color represents the corresponding model below the figure. However, most XsembleNet's results incline toward the FPN model as it is the most promising component inside the ensemble model.

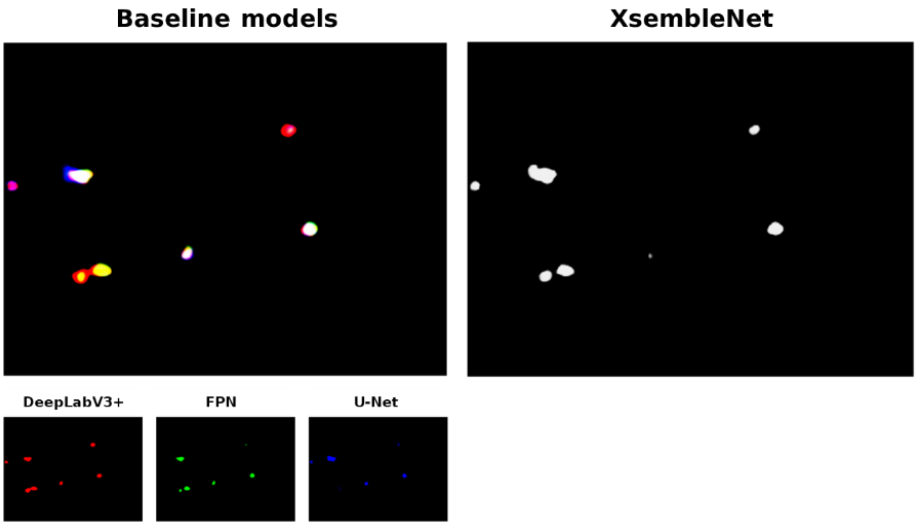

Fig. 4. Result comparison between baseline models and XsembleNet

5 Discussion and Conclusion

5.1 Segmentation results

With the same encoder, different decoders could produce different results. In this experiment, FPN and XsembleNet are appropriate for caries segmentation tasks, as they can achieve satisfactory dice loss and perform well in tooth-level prediction.

We tried to provide feature maps to different decoder architectures from the same backbone in our works. XsembleNet differs from Thambawitta's TriUNet implementation, in which he took three U-Net together [17]. Our ensemble model consists different three base models and training end-to-end. It achieves both good quantity and qualitative metrics. Interestingly, XsembleNet can gain a better Dice score than FPN, the best performer among the three base models.

The computation cost of adding the decoder is not as expensive as adding the decoder, making further extended decoder components possible. However, as previously stated, while the ensemble model could gain better segmentation quality, it is challenging to interpret as a component, so an exhaustive interpretation should be made separately for each submodel.

In our case, XsembleNet's pixel-wise metrics (binary cross-entropy loss and dice loss) and tooth-level metrics are comparable to FPN. As ensemble techniques might sometimes increase efficacy by a low margin, the modeling process is mainly a trade-off between model complexity and performance gain.

To conclude, these experiments show that fully convolutional neural networks can accurately detect dental proximal caries radiographs. The implemented networks of this contribution were a contextual process for utilizing the deep learning model in dentistry

and potentially valuable for clinical routine and means to achieve the reference diagnostic criteria.

5.2 Limitations

Compared to previous works, our models yield an astonishing tooth-level score. Nevertheless, there should be a consideration that training and testing data are relatively small-scale compared to previous studies. The interpreter's bias could affect performance as only one person fabricates the dataset in a pilot-study manner. Increasing dataset size, constructing a gold standard, and performing K-fold cross-validation should suffice for the flaw in this experiment.

References

[1] M. A. Peres et al., “Oral diseases: a global public health challenge,” Lancet, vol. 394, no. 10194, pp. 249–260, Jul. 2019.

[2] F. Chen and D. Wang, “Novel technologies for the prevention and treatment of dental caries: a patent survey,” *Expert Opin Ther Pat*, vol. 20, no. 5, pp. 681–694, May 2010.

[3] Gill J, "Dental Caries: The Disease and its Clinical Management," 3rd ed., NJ: Wiley-Blackwell, 2016.

[4] R.W. Valachovic, C.W. Douglass, C.S. Berkey, B.J. McNeil, H.H., “Examiner Reliability in Dental Radiography,” J Dent Res., vol. 65, no. 3 1986.

[5] R. P. Langlais, L. J. Skoczylas, T. J. Prihoda, O. E. Langland, and T. Schiff, “Interpretation of bitewing radiographs: Application of the kappa statistic to determine rater agreements,” Oral Surgery, Oral Medicine, Oral Pathology, vol. 64, no. 6, pp. 751–756, Dec. 1987.

[6] M. Naitoh et al., “Observer agreement in the detection of proximal caries with direct digital intraoral radiography,” Oral Surg Oral Med Oral Pathol Oral Radiol Endod, vol. 85, no. 1, pp. 107–112, Jan. 1998.

[7] J. Tubert, “Restorative Treatment Strategies Reported by French University Teachers,”, Journal of Dental Education, vol. 68, no. 10, pp. 1096-1103. 2004.

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, Illustrated edition. Cambridge, Massachusetts: The MIT Press, 2016.

[9] A. G. Cantu *et al.*, “Detecting caries lesions of different radiographic extension on bitewings using deep learning,” *Journal of Dentistry*, vol. 100, p. 103425, Sep. 2020.

[10] M. M. Srivastava, P. Kumar, L. Pradhan, and S. Varadarajan, “Detection of Tooth caries in Bitewing Radiographs using Deep Learning,” arXiv, arXiv:1711.07312, Nov. 2017.

[11] J. Long, E. Shelhamer, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation.” arXiv, Mar. 08, 2015.

[12] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation.” arXiv, May 18, 2015.

[13] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition.” arXiv, Dec. 10, 2015.

[14] A. Kirillov, K. He, R. Girshick, and P. Dollár, *A unified architecture for instance and semantic segmentation*. 2017.

[15] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation.” arXiv, Aug. 22, 2018.

[16] C. Zhang, *Ensemble Machine Learning*. Springer, 2012.

[17] V. Thambawita, S. A. Hicks, P. Halvorsen, and M. A. Riegler, “DivergentNets: Medical Image Segmentation by Network Ensemble.” arXiv, Jul. 01, 2021.

[18] J.-H. Lee, D.-H. Kim, S.-N. Jeong, and S.-H. Choi, “Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm,” *J Dent*, vol. 77, pp. 106–111, Oct. 2018.

[19] N. B. Pitts, A. I. Ismail, S. Martignon, K. Ekstrand, G. V. A. Douglas, and C. Longbottom, “ICCMS™ guide for practitioners and educators,” *London: King’s College London*, 2014.

[20] Pavel Iakubovskii, *Segmentation Models Pytorch*. 2020. Accessed: Jun. 21, 2022. [Online].

[21] S. Jadon, “A survey of loss functions for semantic segmentation,” in *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Oct. 2020, pp. 1–7.