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 Chouvatut [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, 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, depending 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 variability in individual dentists' diagnoses. 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 unique treatment strategies. This variability appears even among 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. These 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 radiograph contains 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, can provide consistent and accurate results. [9], [10] Currently, there are many types of FCN models available. Our contribution is in 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 et al. originally introduced fully convolutional networks (FCN) in 2014. [11] FCNs lack a flattening and dense (fully-connected) layer and rely solely on convolution operations. Most convolutional neural networks used in computer vision tasks follow a similar approach: they extract feature maps with a backbone, a feature extraction network, and then pass them forward to specific networks capable of utilizing those

features. In this article, we refer to each network as a "head," with the encoder head serving as a synonym for a network backbone.

The segmentation model can be divided into two principal heads—the encoder head and the decoder head. The encoder head's role is to extract features with a deep convolutional network structure. Feature maps in deep layers contain robust features but limited spatial information, contrasting with features in shallow layers, which are weaker but rich in spatial information. The decoder head can be considered an "inverse" process of the encoder, upsampling low-dimensional feature maps into larger ones. However, upsampling from deep feature maps alone can suffer from a loss of segmentation detail, so a connector is added to the encoder head to transfer spatial data. [11], [12]

The output of the decoder head is multiple feature maps with dimensions matching those of the input image. Finally, a convolution operation with a kernel size of one pixel is applied to generate the prediction result. Because the encoder head can be constructed from a competent image classification model, such as ResNet, [13] our primary interest is in the decoder head, where different approaches may yield different segmentation results. There are three FCN model variations within our scope of work: (1) U-Net implements cascade upsampling with fractionally stridden convolution and skip-connector [12]; (2) Feature pyramid networks (FPN), which convolve and decode at multiple dimensions [14]; and (3) DeepLabV3+, which probes multiple feature maps with dilated convolution to deliver 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 submodels, DivergentNets got better results.

Thambawitta's methods demonstrate two approaches. One is increasing the model components. Another is using majority voting, similar to 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 that has influenced the healthcare field, including dentistry, particularly in the detection of dental caries. One of the earlier implementations was to classify an image as having dental caries or not. Lee studied the efficacy of deep convolutional neural networks, specifically Inception v3, trained on over 3,000 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 categorize images into specific classes. Segmentation tasks, however, can provide additional diagnostic information by outlining lesions, similar to clinical diagnostic criteria, and accounting for anatomical structures in a radiograph. For example, a lesion reaching the middle half of the dentin is classified as moderate 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 it missed only a few caries cases but still produced some ambiguous false-positive results. When compared to dentists, the model outperformed them 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] This model was trained on 3,686 bitewing radiographs. Unlike most dentists, who have low sensitivity to detecting initial lesions, the caries detection model is robust against both initial and advanced caries. In terms of metrics, it demonstrated much higher overall accuracy and sensitivity than the dentists' average, although its specificity remains lower than that of 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 deep learning models can produce satisfactory results, yielding high sensitivity, correctly interpreting results, and maintaining consistency with the reference set. However, even though many segmentation models are now available, no efficacy comparison has been made specifically for dental caries. This gap motivated us to identify the most suitable model architecture for caries segmentation and to

develop an improved model to assist in treatment planning and reduce variability among dentists.

3 Data and Methodology

3.1 Data

We intentionally selected dental bitewing radiographs from various public sources to construct a small-scale pilot study test set. The criteria required that the images feature teeth in the adolescent age group, approximately 10-30 years old, with general characteristics such as the presence of multiple cavities, normal morphology, and no signs of tooth wear.

The dataset contains 326 images, divided into 270 radiographs with at least one caries site and 56 radiographs without caries, showcasing various tooth features and caries lesions. The researcher annotated the ground truth dataset using the Labelme application, resulting in a total of 657 identified 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. Baseline models used for benchmarking included U-Net, FPN, and DeepLabv3+. Subsequently, an ensemble model, named XsembleNet, was constructed by combining components of the three baseline models.

The experiment was conducted in two stages. Initially, each model was trained, and quantitative metrics derived from a pixel-wise loss function were compared. Next, results were interpreted at the tooth level to assess the quality and characteristics of segmentation for each model. This section describes the experimental setup, including data preparation, training techniques, and deep learning network modeling.

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* library. [20] These networks served as the foundation for our planned XsembleNet. Each model's encoder was ResNet34, [13] initialized with ImageNet weights. The number of encoder parameters is 21.2 million across all three models. The total parameters for 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 output. 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, addresses repetitive components and leverages the relatively smaller size of the decoders compared to the encoder. Training multiple sub-models simultaneously increases model size and can lead to memory constraints, particularly in GPU memory. By creating a shared encoder, XsembleNet reduces the model size, and the encoder is trained with different gradients from multiple decoders.

However, some modifications to the encoder are necessary, as the desired feature map sizes vary among the decoders. For instance, the deep layer of the DeepLabV3+ decoder requires a constant feature map size for dilated convolution, which differs from the progressive downsampling used in U-Net and FPN decoders. To accommodate these differences, we designed a Y-shaped encoder, where the stem serves as a common encoder path for all three decoders, then splits into two branches: one for the DeepLabV3+ decoder and another for the U-Net and FPN decoders. This shared-component approach reduces the model size by 41 percent, enabling a more efficient use of the encoder, allowing for a larger model, and reducing computational costs.

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 was augmented through geometric transformations (horizontal flip) and altering pixels' intensity (brightness, contrast, and gaussian-blur). The data is pre-separated into training and validation sets with a 7:3 proportion to ensure all models are trained and validated on the same data.

All models were trained with the same configuration, using the Adam optimizer for 90 epochs to minimize a loss function. Training began with a learning rate of 5e-3, which was reduced by a factor of ten every 30 epochs. The loss function was a combination of binary cross-entropy loss (BCELoss) and Dice loss, defined in equations (1-3). [21] BCELoss and Dice loss contribute to distribution and region training, respectively. In this case, the data is skewed because positive pixels are vastly outnumbered by negative ones. To address this imbalance, BCELoss is multiplied by a factor of 5 (α) to bring it to the same level as Dice loss.

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

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

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

4) Model Evaluation

The model is evaluated in two ways. The first evaluation assesses segmentation quality at the pixel level using segmentation metrics: BCELoss, Dice loss, and Combo loss. The second evaluation rates the models' segmentation performance on 12 additional testing bitewing radiographs, analyzed at the tooth level. For this second evaluation, the prediction results are taken from the model state that achieved the lowest Dice loss, as this metric best correlates with the dentists' needs—segmenting the tooth decay area as accurately as possible.

4 Results

4.1 Quantitative Result

Table 3 shows the losses for the three baseline models and the ensemble model, XsembleNet. It was found that U-Net achieved the lowest BCELoss and Combo loss among all models. While FPN had the lowest Dice loss among the baseline models, XsembleNet yielded an even 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)

Figure 1 shows the plot of validation Combo loss. All performance metrics increased over the epochs until convergence. However, among the models provided, only DeepLabV3+ did not stabilize during training, as indicated by fluctuating loss throughout the entire run.

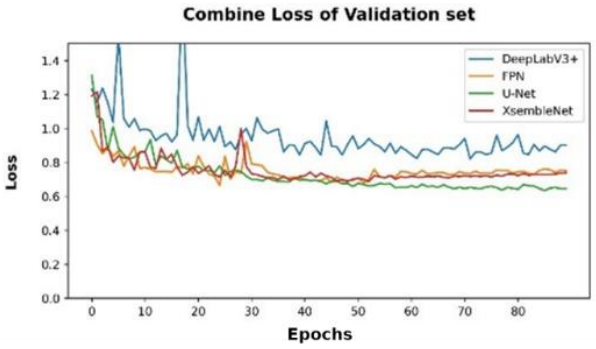


Fig. 1. Validation combo loss per epoch

4.2 Qualitative Result

Each model, at its lowest Dice score state, is then used for inference and analyzed at the tooth level, as illustrated in Figure 2. By merging ground truth with inference results, three distinct areas can be differentiated: true positive (TP), false positive (FP), and false negative (FN) findings. Lastly, true negatives (TN) are defined as teeth without any ground truth or prediction areas.

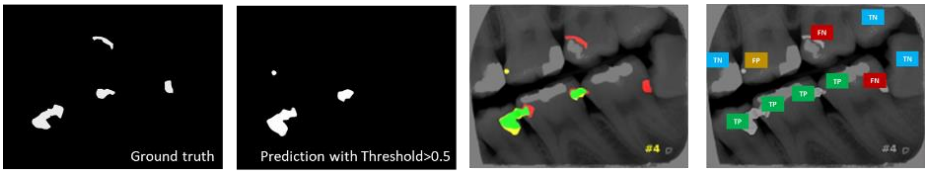


Fig. 2. Tooth-level interpretation

The testing set contains 12 additional bitewing radiographs with 24 cavities across 77 teeth. We constructed confusion matrices and 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

Among the four models, XsembleNet and FPN achieved the highest accuracy, detecting cavities in 23 out of 24 locations, with three false positives each. The U-Net model predicted cavities in 21 locations with four false positives, while DeepLabV3+ detected only 15 cavities and had four false positives.

XsembleNet and FPN scored the highest across all metrics because both models were more accurate in predicting cavities and produced fewer false positives. The U-Net model followed these two, as it missed detecting cavities in some areas. In contrast, DeepLabV3+ had the lowest metric scores due to detecting the fewest cavities while still producing false positives.

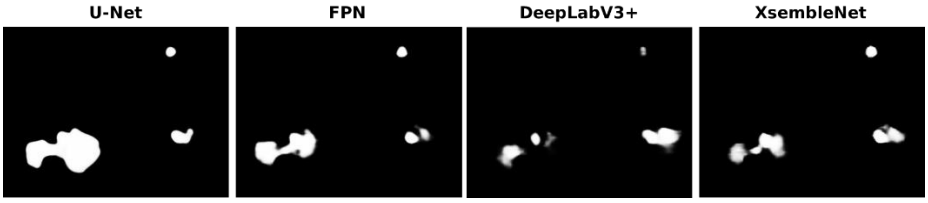


Fig. 3. Pre-thresholded prediction results

We further examined the segmentation characteristics of each model based on pre-threshold prediction results, as shown in Figure 3. Each model exhibited unique performance characteristics. Generally, U-Net produces well-defined, slightly rounded borders but does not adapt to concave areas as well as FPN. DeepLabV3+ differs from these two, as it generates numerous ambiguous patches that appear blurry and are typically below the threshold value (0.5). All models except U-Net exhibit some degree of blurriness. We suspect that multi-level prediction and dilated convolution are the primary causes.

Our implementation, XsembleNet, inherits characteristics from its components. Its predictions are slightly blurred but much less so than DeepLabV3+ results. We found that XsembleNet's results closely resemble a blend of the baseline models, as illustrated in Figure 4, where each color represents a corresponding model. However, XsembleNet's results tend to align more closely with the FPN model, as it is the most promising component within the ensemble.

5 Discussion

With the same encoder, different decoders can produce varying results. In this experiment, FPN and XsembleNet were found to be suitable for caries segmentation tasks, as they achieved satisfactory Dice loss and performed well in tooth-level prediction.

Our approach provided feature maps to different decoder architectures from the same backbone. XsembleNet differs from Thambawitta's TriUNet, which combines three U-Nets. [17] The ensemble model consists of three different base models trained end-to-end, achieving both quantitative and qualitative success. Notably, XsembleNet achieved a better Dice score than FPN, the best performer among the three base models.

The computational cost of adding decoders is not as high as adding encoders, making further extensions with additional decoder components feasible. However, while the ensemble model can enhance segmentation quality, it is complex to interpret as a whole; thus, each sub-model requires separate, detailed interpretation.

In our case, XsembleNet's pixel-wise metrics (binary cross-entropy loss and Dice loss) and tooth-level metrics are comparable to those of FPN. Ensemble techniques may sometimes increase efficacy by a small margin, making the modeling process a balance between model complexity and performance gain.

Compared to previous works, our models achieved satisfactory tooth-level prediction accuracy. However, it is important to note that the training and testing data are relatively small-scale compared to previous studies. Interpreter bias may have influenced performance, as only one person annotated the dataset in this pilot study. Increasing the dataset size, constructing a gold standard, and performing K-fold cross-validation would help address this limitation.

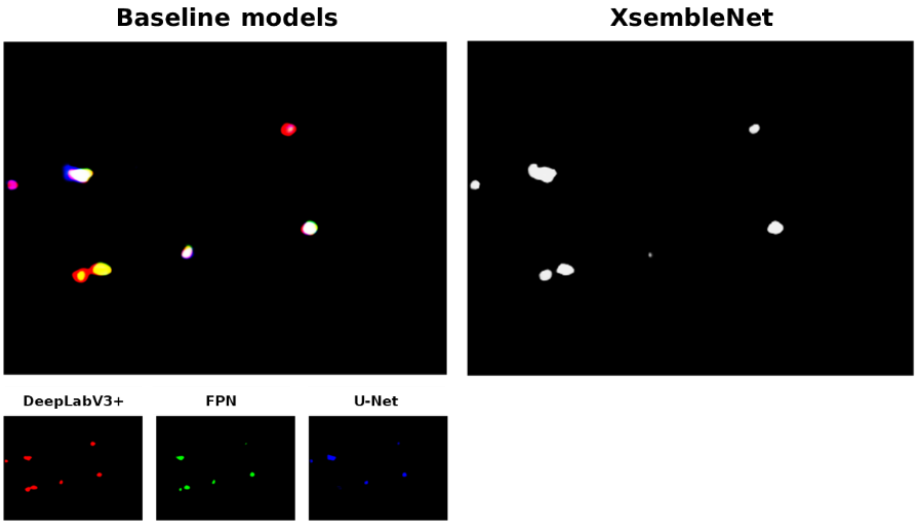


Fig. 4. Comparison result between baseline models and XsembleNet

6 Conclusion

The experiments demonstrate that fully convolutional neural networks can accurately detect dental proximal caries in radiographs. The implemented deep learning models show potential as tools for clinical routine use and for supporting adherence to reference diagnostic criteria.

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.

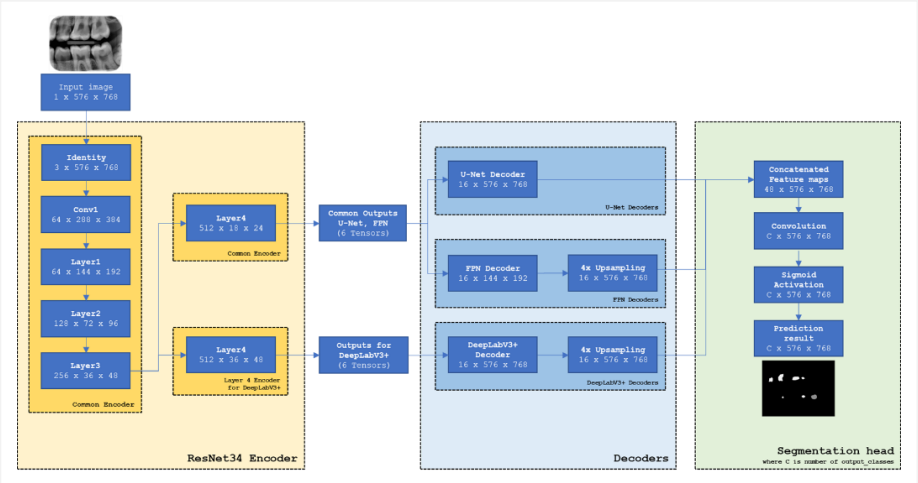
[18] J.-H. Lee, D.-H. Kim, S.-N. Jeong, and S.-H. Choi, "Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm," *J Dent*, vol. 77, pp. 106–111, Oct. 2018.

[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.

Appendix



Appendix Fig. 1. XsembleNet Arcitecture