Filling in the Gap: Enhancing Dental Diagnosis using Tooth-wise Semantic Segmentation with U-Net-DinoV2

Aimon Benfield-Chand
Department of Computer Science
University of Washington
Seattle, WA 98195
aimonbc@cs.washington.edu

Abstract

Dental and oral diseases significantly impact global health, necessitating advancements in diagnostic technologies. This paper introduces a novel approach to dental abnormality detection using panoramic X-ray radiograph segmentation and mask generation. Leveraging the Tufts Dental Database (TDD), I develop a multi-step method that segments dental radiographs into individual tooth crops and subsequently trains a semantic segmentation model to generate dental abnormality masks. This approach integrates a U-Net-inspired architecture with Meta's DinoV2 pretrained Vision Transformer (ViT), providing a balance between precise abnormality localization and robust feature extraction. The effectiveness of the U-Net method is assessed in comparison to a baseline model. Findings reveal that while the quantitative performances of the U-Net and baseline models are similar, the U-Net architecture provides superior qualitative benefits, particularly in the context of dental abnormality localization. Future enhancements, including the expansion of the dataset and further refinement of the DinoV2 backbone with a broader range of medical or dental images, hold the potential to elevate the model's performance even further. Code and models are available at https://github.com/aimonbc24/Dental-Abnormality-Detection-and-Segmentation.

1 Introduction

Dental and oral diseases significantly impact global health, affecting around 35% of the population with untreated tooth decay, 11% with severe gum disease, and 2% experiencing tooth loss (Gabbar et al., 2023). These statistics not only underscore the prevalence of oral health issues but also highlight the opportunity for designing more effective diagnostic methodologies. Traditional diagnostic practices predominantly depend on manual examinations by dental professionals. While these methods have been the cornerstone of oral healthcare for decades, they are not without their shortcomings, being subject to variability in diagnostic accuracy and sometimes leading to inconsistencies in patient outcomes. Additionally, these examinations often cause discomfort to patients, which can be a deterrent to seeking timely dental care.

In the wake of the deep learning revolution, there has been increased desire to leverage technological advancements to modernize dental diagnostics. Situated at the intersection of oral healthcare and artificial intelligence, this study aims to surmount the limitations of manual examinations by developing a deep learning system that achieves competitive diagnostic performance to traditional methods, all the while offering a more streamlined and patient-friendly experience.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

2 Related Works

Recent advancements in deep learning have shown promise in augmenting dental abnormality detection. Prior work has utilized convolutional neural networks (CNNs) for segmenting and classifying dental images, achieving notable results. For instance, Dhake and Ansari (2022) leveraged ResNet-101 to conduct comprehensive image evaluations. Alternatively, Sun and Chen (2022) adopted a transformer-based model to identify caries (loss of tooth substance, e.g. enamel or dentine) within panoramic x-ray images, using FPN as the backbone for feature fusion and a sparse RCNN. They achieve an AP50 [1] of up to 68.31 when adding an SE module.

Recently, a medical image segmentation framework proposed by Chen et al. (2021) called TransUNet built upon the success of vision transformers (ViT) by integrating the transformer architecture into the U-Net framework in order to capture fine-grained features along with global context. Compensating for the reduced feature resolution introduced by the Transformer, this hybrid architecture appears promising for medical computer vision tasks that require local precision.

Nashold, Pandya, and Lin (2022) from Stanford University proposed a multi-objective CNN approach for processing panoramic x-ray images, with one objective being the generation of teeth segmentation masks and the other being the binary classification of radiographs as “normal” or “abnormal”. Using ResNet-50 as a backbone, a two fully-connected layers for the segmentation head, and Atrous Spatial Pyramid Pooling layers for the abnormal classification head, they improve upon the baseline accuracy achieved by training a ResNet-18 architecture for abnormality detection and a ResNet-50 architecture for teeth segmentation. These dental radiographs were obtained from the 2022 Tufts Dental Database (TDD), a public database of 1,000 panoramic dental radiography images (Panetta et al., 2022), which I use for my project and describe in more detail in Section 3.

2.1 Proposed Approach

Extending on the work of Nashold et al., I propose a novel multi-step method utilizing the Tufts Dental Database to generate dental abnormality masks from panoramic x-ray radiographs. This approach involves initially segmenting the dental radiographs into individual teeth crops and subsequently fine-tuning a segmentation model to generate dental abnormality masks from these crops. Like Nashold et al., all abnormalities present in the TDD are treated identically such that the model must learn an abstraction of the features defining a dental abnormality, rather than the structural specifics differentiating individual types of abnormalities, such as dental caries and periodontal disease. This coarse-grained approach is sufficient since the model is only intended to aid professionals in abnormality diagnosis by narrowing down the potential regions of concern.

Similar to TransUNet, my segmentation model is a U-Net-inspired architecture (Noh et al., 2015) that leverages Meta’s DinoV2 pre-trained Vision Transformer (ViT) backbone in between a CNN encoder and decoder (Oquab et al., 2023; Noh et al., 2015). While the ViT backbone weights remain frozen, the U-Net encoder and decoder networks are trained on top of the backbone in a supervised procedure. With DinoV2 achieving state-of-the-art performance among open-source models in a variety of downstream vision tasks, I anticipate that its embedding quality will transfer to encoding dental abnormalities. To determine the effectiveness of the U-Net approach, the model is benchmarked against a baseline model composed of the DinoV2 backbone and a linear segmentation head.

2.2 Expected Outcome

I expect the multi-step approach for detecting abnormalities to offer several advantages over existing methods. Though it relies on external tooth segmentation technologies, the preliminary tooth segmentation step is expected to enhance the focus of potentially abnormal regions in x-ray tooth crops, while also greatly decreasing the load on GPU memory and enabling larger batch sizes. Enhanced focus on the teeth may enable the model to learn more salient features associated with dental abnormalities. Furthermore, the prevalent binary classification methods, which merely flag the presence of abnormalities, lack the spatial granularity necessary for practical dental application. Unlike mere binary detection, generating entire segmentation masks for detected abnormalities not only identifies the existence of an abnormality but also pinpoints its exact location. Such

[1] AP50 is a common version of the Intersection over Union (or Jaccard Index) metric, where classifications are specifically determined by thresholding predicted probabilities at 0.5.

precise mapping is anticipated to expedite the identification process for dental professionals, saving considerable time and improving diagnostic efficiency.

3 Data

Like Nashold et al., I use the Tufts Dental Database (TDD) for the supervised-training procedure. The TDD includes 1,000 well-annotated panoramic radiographs and abnormality masks of pixel dimensions 840 x 1615 (Figure 1). For each radiograph, a set of bounding boxes detailing the location of individual teeth in the image is also provided, allowing for accurate tooth segmentation.

3.1 Image cropping

Using the provided bounding boxes,[2] I crop the panoramic x-rays and their respective abnormality masks into individual tooth crops and abnormality masks of pixel-size 140 x 56, an example of which can be seen in Figure 2. My choice of crop dimension arises from DinoV2'd restriction on the spatial dimensionality. To use the ViT, the height and width of an input image must be multiples of 14 so that the image can be processed into 14 x 14 pixel-patches. Across the TDD, the average tooth crop has dimensionality 144 x 64, leading me to choose crop dimensions of 140 x 56. As a result of preliminary tooth segmentation done in preprocessing, the number of sample radiographs increases approximately 26-fold from 1,000 panoramic images to 26,005 tooth crops. Of these initial croppings, 87 were discarded due to inaccurate bounding boxes that produced malformed images, yielding 25,918 successful crops.

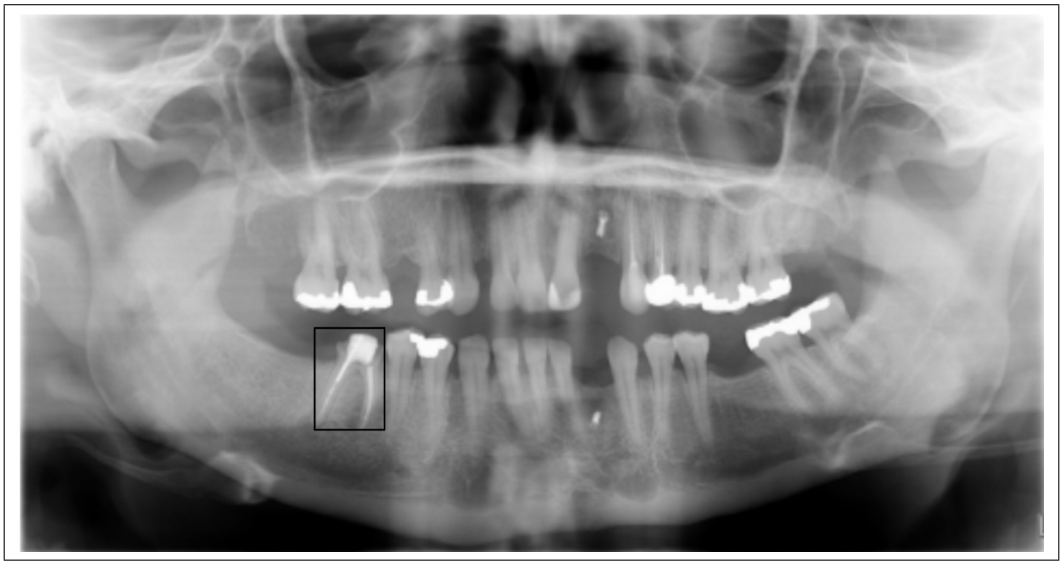

Figure 1: Example panoramic radiograph from the TDD, with a bounding box containing the bottom, left-most tooth.

3.2 Abnormality Masks

The data set includes descriptions of identified abnormalities at five different levels of granularity. These are relative anatomic location, peripheral characteristics, radio-density, effects on surrounding dental structure, and abnormality category (i.e. trauma, inflammation, dysplasia, developmental, benign tumor or cyst, malignant neoplasia, and systemic or metabolic conditions). Of these abnormality descriptions, I primarily focus on relative anatomic location, which is encoded as an abnormality segmentation mask for each radiograph.

[2] In real-world application, such bounding boxes would need to be generated from scratch. However, state-of-the-art tooth segmentation machine learning models have rendered this task trivial.

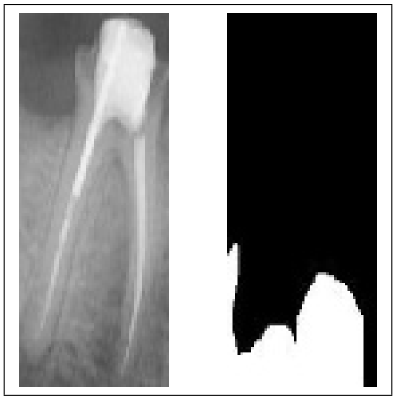

Figure 2: Scaled crop of the tooth contained in the bounding box in Figure 1 with its corresponding abnormality mask.

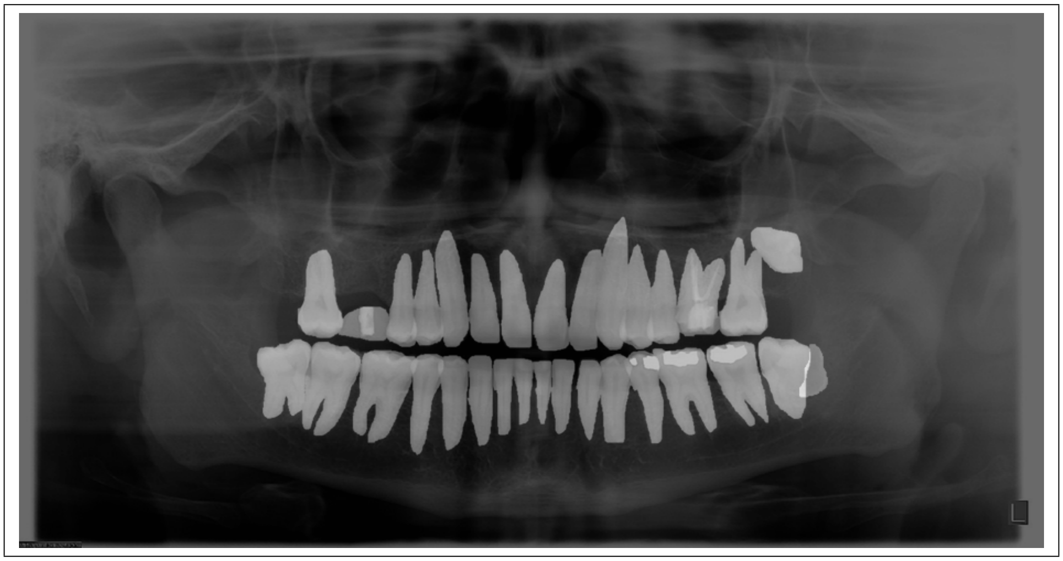

Figure 3: An example overlay of a ground-truth segmentation mask on top of the associated panoramic x-ray.

3.3 Challenges

One unavoidable challenge of my approach arises due to the binary and localized nature of image segmentation. This makes accurate mask generation challenging, as even slight shifts in abnormality localization can incur significant loss penalties.

Another major issue specific to the TDD is that of class imbalance, which is known to affect per-pixel segmentation accuracy. Of the original 1,000 panoramic radiographs, only 34% contain abnormalities and only roughly 0.28% of the total pixels are masked as abnormal. While the percentage of total abnormal pixels slightly increases to 0.43% upon preliminary tooth segmentation, the percentage of tooth crops containing abnormality decreases significantly from 34% down to only 3.2%. This result is due to the fact that most dental abnormalities only cover a small number of teeth in a panoramic dental radiograph. While some abnormalities may even be lost altogether after segmentation, others are split across multiple crops. This is especially common among abnormalities primarily affecting the gum and tooth root, since only those gum or root regions directly surrounding individual teeth are included in the tooth crops.

In an attempt to correct for image-wise class imbalance, a simple horizontal-flip is applied to crops containing abnormality. Following this data augmentation, 6.5% of tooth crops contain abnormality, while 0.87% of pixels are classified as abnormal.

4 Methods

4.1 Loss and Accuracy Metrics

Due to class imbalance in the data set, weighted binary cross-entropy (weighted BCE) is used as the training loss function.

$$\begin{aligned}\textbf{Weighted BCE} = -\frac{1}{D}\sum_{i=1}^{D}[w_p \cdot y_i \cdot \log(\hat{y}_i) \\ + w_n \cdot (1-y_i) \cdot \log(1-\hat{y}_i)]\end{aligned}$$

where D is the total number of pixels in the image, y_i is the ground-truth label of the i-th pixel, $\hat{y}_i$ is the predicted label of the i-th pixel, w_p is the weight for positive/abnormal pixels (when $y_i = 1$), and $w_n = 1 - w_p$ is the weight for negative/normal pixels (when $y_i = 0$).

The advantage of weighted BCE is that it explicitly compensates for the under-representation of abnormal pixels by giving them greater 'weight' when calculating the BCE. Although more complex and nuanced weighted loss functions such as focal loss exist, these generally depend on multiple hyperparameters, whereas simple weighted BCE only uses a single weight hyperparameter, making it preferable for its simplicity. As discussed in future sections, this weight hyperparameter is tuned to optimize for a balance of various metrics.

In order to use weighted BCE loss, the abnormality masks predicted by the models are first passed through a sigmoid operation, thereby converting the raw pixel values into probabilities of a binary classification between normal (black or 0) and abnormal (white or 1). After performing weighted BCE, these predicted probability masks are further converted into binary abnormality masks using the threshold of 0.5, where probabilities greater than or equal to 0.5 are classified as abnormal and probabilities of less than 0.5 are classified as normal.

Given a predicted binary abnormality mask $\hat{Y}$ and a true binary abnormality mask Y, pixel-wise accuracy is calculated as

$$\begin{aligned}\textbf{Pixel-wise Accuracy} &= \frac{|Y \cap \hat{Y}|}{|Y|} \\ &= \frac{TP + TN}{TP + TN + FN + FP}\end{aligned}$$

where TP, TN, FN, and FP are abbreviations for true positives, true negatives, false negatives, and false positives.

In addition to this basic accuracy metric, the Dice Coefficient and Jaccard Index (intersection over union) metrics are computed Eelbode et al. (2020), with both being commonly used in semantic segmentation to represent the difference between a predicted and ground-truth mask.

The Dice coefficient and Jaccard Index are defined respectively as

$$\textbf{Dice} = \frac{2 \times |Y_p \cap \hat{Y}_p| + \epsilon}{|Y_p| + |\hat{Y}_p| + \epsilon}$$

$$\begin{aligned}\textbf{Jaccard} &= \frac{|Y_p \cap \hat{Y}_p| + \epsilon}{|Y_p \cup \hat{Y}_p| + \epsilon} \\ &= \frac{TP}{TP + FP + FN}\end{aligned}$$

where $|Y_p \cap \hat{Y}_p|$ is the number of true positives and ϵ represents a smoothing factor added for numeric stability.

Lastly, the metrics of precision and recall are computed as

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

which are subsequently used to obtain an F1 score.

$$\mathbf{F}_1 = \frac{2 \cdot (\text{precision} \cdot \text{recall})}{\text{precision} + \text{recall}}$$

As with weighted BCE, the advantage of these evaluation metrics is their focus on pixels masked as abnormal. Due to the under-representation of abnormalities in the dataset, optimizing model performance with greater emphasis towards pixels masked as abnormal may help prevent the model from learning a degenerate solution that simply masks every pixel as normal and guarantees a $> 99\%$ pixel-wise accuracy.

Another metric of interest is whether the model would correctly classify a tooth crop image as being "normal" or "abnormal", where "abnormality" is defined by some predefined threshold percentage of abnormal pixels in the abnormality mask. In other words, for a given threshold t, a model would classify an tooth crop as "abnormal" if the percent of abnormal pixels in its generated abnormality mask is at least t. This image-wise binary classification accuracy is defined as the percent of tooth crops a model would correctly identify as "normal" or "abnormal" from the validation set for a given threshold.

4.2 Architecture

The primary model architecture is designed around the U-Net deep learning architecture. First introduced by Hyeonwoo Noh et al., the U-Net architecture has become a preferred framework for semantic segmentation due to its impressive performance on datasets like PASCAL VOC 2012 Noh et al. (2015). Its structure comprises an encoder network or "contractive path", which serves as a feature extractor, and a symmetrical decoder network or "expansive path", which upsamples the latent feature embeddings to generate segmentation masks. Another important feature of the U-Net architecture is its integration of skip-connections between the encoder and the decoder. Similar to residual connections, these skip-connections work by concatenating the outputs of encoder layers with the inputs to decoder layers at respective levels of spatial resolution. Rather than simply supporting gradient flow, these connections also help to transmit fine-grained, spatial information throughout the network and facilitate the precise reconstruction of local appearance and structural features in the generated segmentation masks. In a standard CNN-based approach, U-Net encoders often feature convolutional, pooling, Rectified Linear Unit (ReLU) and fully-connected layers, while the decoders include series of unpooling, convolutional, deconvolutional, and ReLU layers.

Modernizing this approach, I leverage the recent Transformer revolution in deep learning by incorporating Meta's pretrained DinoV2 Vision Transformer (ViT) backbone into the U-Net framework Oquab et al. (2023). Trained in a self-supervised fashion on a large corpus of image data, the DinoV2 backbone has achieved state-of-the-art performance among open-source models across a variety of downstream semantic segmentation benchmarks. Building on the success of dinov2, I train CNN-based encoder and decoder networks from scratch, which are sandwiched in between the dinov2 backbone, to generate dental abnormality masks. A diagram of the architecture is shown in Figure 4.

In my U-Net architecture, the encoder is composed of a sequence of three "Conv2d - LayerNorm - ReLU" blocks, which downsamples (1, 140, 56)-dimensional images into (1, 98, 28)-dimensional embeddings so as to reduce the memory usage of the ViT. These encoder blocks have channel outputs of 32, 64, and 1, respectively. The ViT backbone subsequently processes these embeddings into fourteen[3] (14, 14)-dimensional image patches, each of which are then encoded into 384-dimensional vectors. The (14, 384)-dimensional ViT output is then reshaped to (7, 2, 384) and permuted into an image of the form (C=384, H=7, W=2), which is next passed through an upsampling network to reconstruct the (98, 28) spatial resolution of the output from the final encoder block.

Mirroring the channel dimensions of the layers in the encoder, the decoder further upsamples this image using a sequence of three "Conv2d - LayerNorm - ReLU - TransposedConv2d" blocks. For

[3] The fourteen patches arises from the (98, 28) size of the encoded images. Processing these images into 14 x 14 pixel patches yields $98/14 \cdot 28/14 = 7 \cdot 2 = 14$ patches.

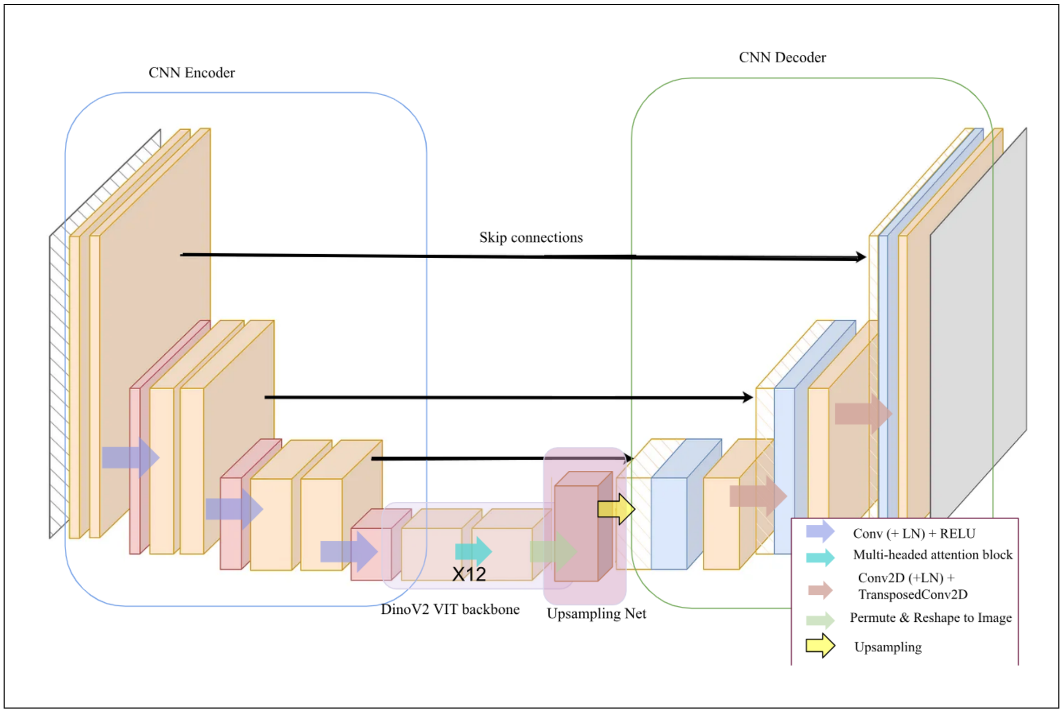

Figure 4: Architectural diagram of the ViT-extended U-Net framework.

skip connections, each decoder block receives the output of the encoder block of corresponding spatial resolution as well as the output from the previous decoder block. These inputs are concatenated together and passed through a convolutional layer, layer normalization, and a transposed convolution. Finally, I apply a 1x1 convolution to the output of the last decoder layer to produce a (1, 140, 56)-dimensional logit image, which is passed through a sigmoid operation to produce the segmentation mask.

I assess the performance of the U-Net model by conducting a comparative analysis with a baseline architecture. In the baseline approach, I input each image into the DinoV2 backbone, flatten the outputted (40, 384) embedding matrix into a vector, apply a fully-connected linear layer and unflattening operation, and pass the resulting (1, 140, 56) image through a sigmoid operation.

4.3 Training and Cross Validation

The models are trained using the Adam optimizer, and cross validation is performed to select for several hyperparameters. I first performed a hyperparameter search over the learning rate and found 0.001 to be adequate.

In the U-Net model architecture, the output of the DinoV2 backbone produces an embedding matrix of shape (14, 382), and the first decoder block required height, width dimensions of (98, 28) as input. To upsample the embedding matrix to the appropriate shape, I experimented with both a one layer upsampling and a multi-layer upsampling. The multi-layer upsampling network consists of 3 convolutional layers with ReLU non-linearities, and was generally found to outperform the one-layer upsampling network in terms of loss after training for the same number of iterations. I therefore use the multi-layer upsampling network when comparing the U-Net model with the baseline in my experiments.

I also tried four different configurations for the number of channels in the encoder and decoder blocks. Listed in order of the encoder blocks, these output-channel configurations are [32, 64, 3], [32, 64, 1], [64, 128, 3], [64, 128, 1]. After cross validation using the multi-layer upsampling network, the simpler [32, 64, 1] output channel configuration of the U-Net was found to not only reduce training latency, but also achieve slightly better performance–a possible result of the small size of the input images and data set.

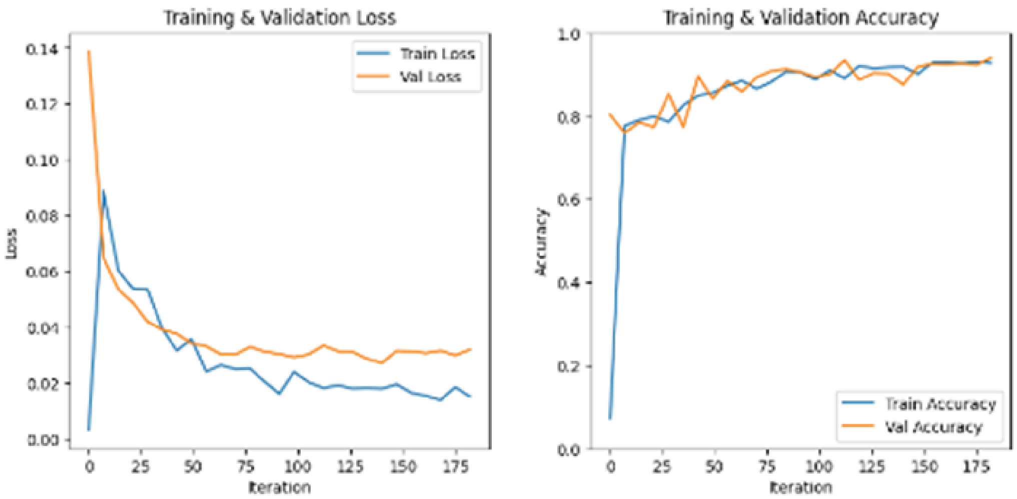

Figure 5: Training and validation loss and accuracy for the baseline model.

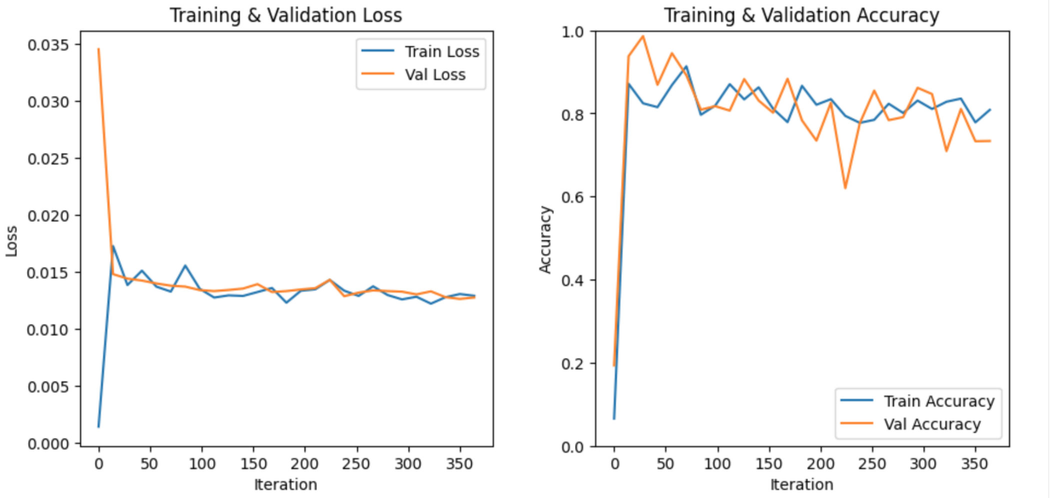

Figure 6: Training and validation loss and accuracy for the U-Net model.

Using the hyparameter configuration selected above, further cross validation was performed to tune the abnormality weight hyperparameter w_p of the weighted BCE loss function. Following a coarse grid-search for w_p over the range $[0.90, 0.99]$, the range was refined to $[0.980, 0.990]$ using increments of 0.001. Weights within this range closely reflect the class imbalance in the data set and were most effective in balancing a range of evaluation metrics. Of particular importance were the metrics of recall and pixel-wise accuracy, which were found to be inversely proportional with respect to the weight hyperparameter. By placing a high penalty on false negatives (pixels incorrectly masked as normal), high values of the abnormality weight hyperparameter led to a greater percentage of pixels being masked as abnormal, which increased recall and prevented the model from learning a degenerate solution. However, without high precision on those pixels masked as abnormal, increasing the weight hyperparameter also decreased the pixel-wise accuracy. Given that false negatives are worse than false positives in the context of dental abnormality diagnosis and detection, I tune this hyperparameter for a balance of recall and accuracy, while allowing for relatively poor precision. After cross-validation, the models were found to achieve a reasonable balance between recall and pixel-wise accuracy with an abnormality weight of 0.984 for the baseline and 0.987 for the U-Net.

Figures 7 and 8 depict the image-wise binary classification accuracy, as described in the loss and accuracy metrics section, over several thresholds of abnormality for the final baseline and U-Net

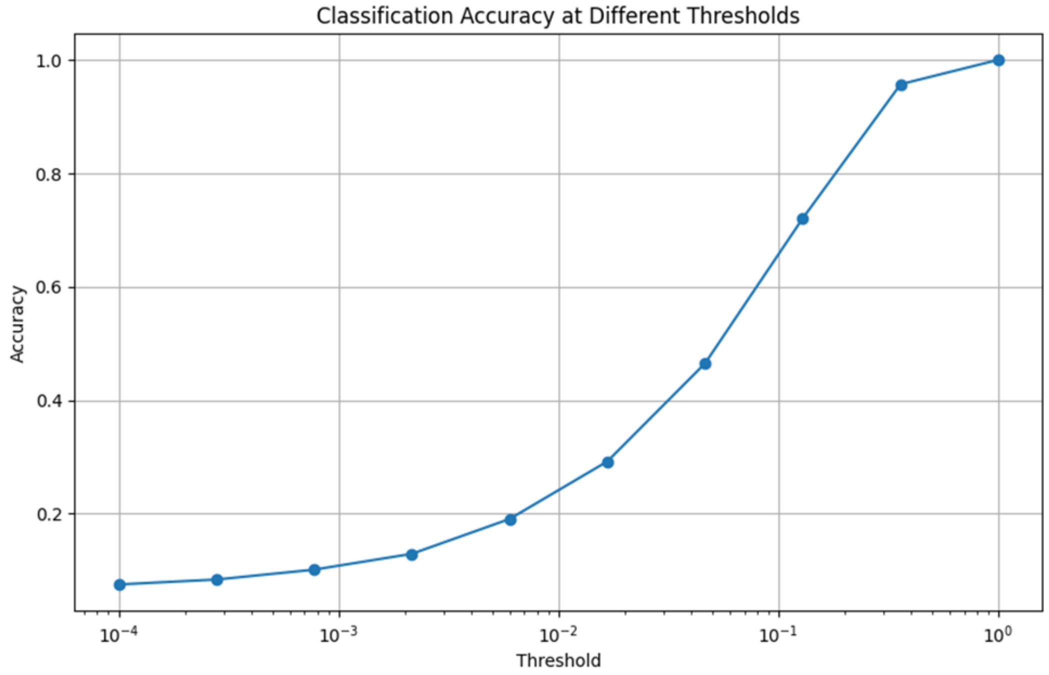

Figure 7: Image-wise binary classification accuracy by classification threshold for the baseline model.

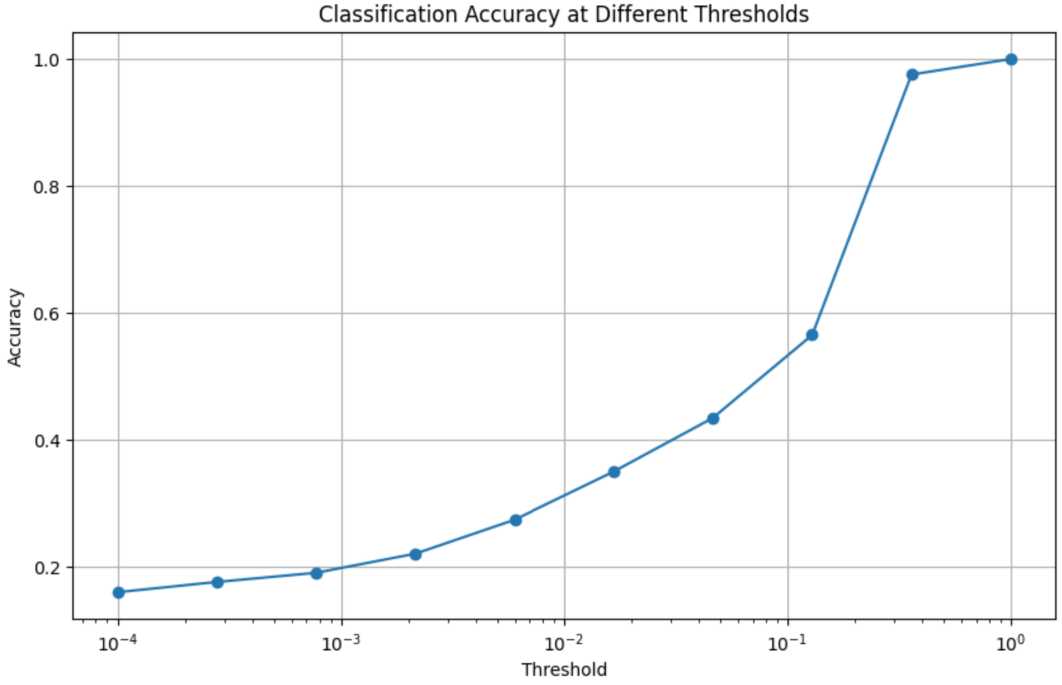

Figure 8: Image-wise binary classification accuracy by classication threshold for the U-Net model.

models, respectively. For high thresholds, the models predict that most of the tooth crops are normal, leading to high image-wise accuracy. This is expected given that the dataset has few abnormalities to begin with. Conversely, for extremely low thresholds, the accuracy is also quite low. This likely results from the fact that optimizing the models for recall caused them to mask more pixels as abnormal, thereby producing more false positive predictions.

5 Results

The tables of results (1 and 2) contain the test-time evaluation metrics achieved by the baseline (BASE) and U-Net models with the best performing hyperparameter configurations. Specifically, the baseline model uses a learning rate of 0.001 and abnormality weight of 0.984, while the U-Net model uses a multi-layer upsampling network, a learning rate of 0.001, and an abnormality weight of 0.987.

Table 1: Evaluation Metrics

	Loss	Pixel-wise Accuracy	Dice	Jaccard
BASE	0.030	0.984	0.031	0.025
U-Net	0.013	0.987	0.110	0.106

Table 2: Additional Metrics

	Precision	Recall	F1	$\lvert\hat{Y}_p\rvert/\lvert\hat{Y}\rvert$
BASE	0.054	0.457	0.096	0.077
U-Net	0.027	0.413	0.051	0.132

In the table, $|\hat{Y}_p|/|\hat{Y}|$ represents the total percentage of pixels masked as abnormal across the test data set. Generally, higher percentages of abnormal pixels in the predicted masks correlates with higher recall and slightly lower precision.

The baseline model and U-Net demonstrate similar pixel-wise accuracy and recall, with the U-Net having a slightly higher accuracy and loss. By examining the percentage of pixels masked as abnormal and the precision of the two models, it also appears that U-Net's predictions were slightly less precision than the baseline's. Though important, the reported pixel-wise accuracies are highly inflated due to the class imbalance in the data set. Further analysis of the low recall and precision scores indicates major room for improvement in the models' ability to correctly and precisely identify abnormality.

5.1 Discussion

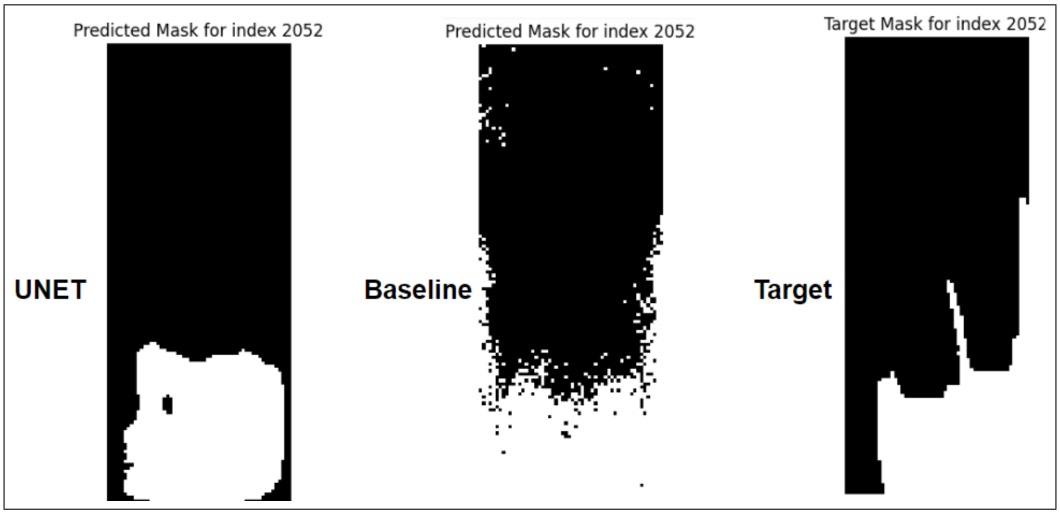
Predicted Mask for index 2052
Predicted Mask for index 2052
Target Mask for index 2052
UNET
Baseline
Target

Figure 9: From left to right: abnormality mask generated by U-Net, mask generated by baseline, and target mask.

Upon evaluating the baseline and U-Net in Tables 1 and 2, the U-Net model shows comparable results to the baseline for abnormality detection. While the quantitative comparison between the U-Net and baseline models appears inconclusive, the qualitative differences in the predicted abnormality

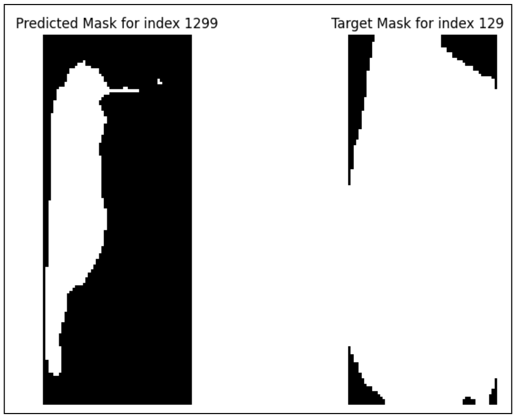

Figure 10: Mask predicted by U-Net resembling a hummingbird and its target mask.

masks is clear. The boundaries in the U-Net (Figure 9) are more distinctive than the linear baseline model which appears fairly accurate despite being more noisy. Some of the U-Net's abnormality masks also appear to have distinctive shapes that resemble everyday objects, such as a bird (Figure 10) or a face (Figure 9), present in the ImageNet-1k data set on which DinoV2 was pretrained Oquab et al. (2023). One possible reason why the U-Net model is more prone to hallucinating such objects than the baseline model is due to the difference in input received by the DinoV2 backbones in each setup. Whereas the ViT backbone receives high-level feature maps extracted by a CNN encoder in the U-Net model, it receives the actual radiograph inputs in the baseline model. The high-level features extracted by the CNN encoder (e.g. lines, curves, edges) are likely similar to those extracted from images of everyday objects, thereby coaxing the ViT backbone into detecting, and subsequently fitting, everyday objects within regions of dental abnormality.

5.2 Future Work

While well-annotated, the scale of the dataset proved to be a significant challenge throughout training with the class imbalance causing skewed metrics. As high-quality labeled data is scarce, adopting a self-supervised approach with unlabeled data may help the model learn more salient features in dental radiographs. Given that the U-Net seemed to hallucinate objects from its pretraining data set, further self-supervised fine-tuning of the ViT backbone on a large corpus of dental or other medical images would likely steer the model towards predicting shapes more characteristic of dental abnormalities themselves.

Given the similar quantative performance of the baseline to the U-Net, another possible extension could involve adding convolutional layers to the linear segmentation head in order to improve the qualitative quality of the baseline's predicted masks by reducing pixel noise and inducing smoothness. This would be a relatively simple addition that would likely bolster the usability of the baseline in practical abnormality diagnosis and detection.

A final extension would be to apply the U-Net model to the panoramic x-rays rather than the tooth crops. While segmenting the panoramic x-rays into tooth crops greatly decreased the memory and compute load of the model, only the baseline required this reduction, given that a weight matrix of dimensions ($6900^4 \cdot 384$, $840 \cdot 1610$) would be far too large to fit into memory. By comparison, the U-Net architecture avoids this memory constraint as the CNN encoder downsamples the large input image to a compressed latent embedding and the CNN decoder upsamples the latent embedding to the original spatial resolution. While the panoramic radiographs would likely need to be downsampled to $\frac{1}{64}$ their original dimension, this could be performed using max-pooling layers. In fact, the greater reduction in spatial resolution might benefit U-Net performance by enabling skip-connections at far broader ranges of spatial resolution.

[4] This results from rounding the panoramic radiograph dimensions from (840, 1615) to the nearest multiples of 14 which are (840, 1610). These images are processed by the ViT as 14x14 patches, resulting in $840/14 \cdot 1610/14 = 6900$ total patches.

5.3 Conclusion

In this study, I developed a novel approach for identifying abnormalities from dental radiographs. With limited data, the model generates an abnormality mask on a given tooth crop, which has applications in assisting with dental disease diagnosis and treatment planning. While quantitative assessment of the U-Net and baseline models yielded comparable results, the U-Net architecture demonstrated the qualitative advantage of less noisy abnormality localization, which is beneficial in the context of dental abnormality detection. With more data and fine-tuning of the DinoV2 backbone on medical images, I further expect that the performance of the U-Net model could be improved.

References

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation, 2021.

Tilottama Dhake and Namrata Ansari. A survey on dental disease detection based on deep learning algorithm performance using various radiographs. In *2022 5th International Conference on Advances in Science and Technology (ICAST)*, pages 291–296, 2022. doi: 10.1109/ICAST55766.2022.10039566.

Tom Eelbode, Jeroen Bertels, Maxim Berman, Dirk Vandermeulen, Frederik Maes, Raf Bisschops, and Matthew B. Blaschko. Optimization for medical image segmentation: Theory and practice when evaluating with dice score or jaccard index. *IEEE Transactions on Medical Imaging*, 39(11):3679–3690, 2020. doi: 10.1109/TMI.2020.3002417.

Hossam A. Gabbar, Abderrazak Chahid, Md. Jamiul Alam Khan, Oluwabukola Grace-Adegboro, and Matthew Immanuel Samson. Tooth.ai: Intelligent dental disease diagnosis and treatment support using semantic network. *IEEE Systems, Man, and Cybernetics Magazine*, 9(3):19–27, 2023. doi: 10.1109/MSMC.2023.3245814.

L. Nashold, P. Pandya, and T. Lin. Multi-objective processing of dental panoramic radiographs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2022.

Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation, 2015.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Karen Panetta, Rahul Rajendran, Aruna Ramesh, Shishir Paramathma Rao, and Sos Agaian. Tufts dental database: A multimodal panoramic x-ray dataset for benchmarking diagnostic systems. *IEEE Journal of Biomedical and Health Informatics*, 26(4):1650–1659, 2022. doi: 10.1109/JBHI.2021.3117575.

Che Sun and Hu Chen. An attention-based transformer model for dental caries detection. In Guoqiang Zhong, editor, *International Conference on Electronic Information Engineering, Big Data, and Computer Technology (EIBDCT 2022)*, volume 12256, page 122562R. International Society for Optics and Photonics, SPIE, 2022. doi: 10.1117/12.2635362. URL `https://doi.org/10.1117/12.2635362`.