Dental Disease Classification in Panoramic X-Rays: A Multi-Stage Approach

Songyu Han
Institute for Computational & Mathematical Engineering
Stanford University
songyuh@stanford.edu

James Anderson
Department of Civil & Environmental Engineering
Stanford University
jamesta@stanford.edu

Sawyer Brundage
Department of Electrical Engineering
Stanford University
brundasa@stanford.edu

I. Introduction

Dental issues have been common health concerns around the world, and radiological images are being increasingly relied upon for diagnosing anomalies such as cavities, lesions, and impacts. However, manual diagnosis is subject to time constraints and dentists' availability, and may ultimately hinder treatment due to human error, especially in the case of inexperienced or less thorough dentists. To address this, we developed a multi-stage approach consisting of both deep learning and machine learning components aimed at automating dental X-ray analysis and providing decision support for dentists. Specifically, the goal of our machine learning based approach is automating tooth enumeration and quadrant classification, as well as identifying teeth that are affected by various dental abnormalities such as caries (cavities), lesions and impaction.

Towards this goal, we propose a multi-staged approach consisting of a deep object detection model, a enumeration classifier, and a disease diagnosis classifier. For the first stage our approach, we use grayscale panoramic dental X-ray images as inputs to our object detection model, which produces a series of bounding boxes highlighting detected teeth. In the second stage, we use these bounding boxes as inputs to the enumeration classifier to obtain quadrant and enumeration labels. In the final stage, we crop the overall panoramic image into smaller images containing each detected tooth which are input to our disease classifier model.

II. Related Works

In recent years, the availability of high quality clinical dental images as well as the development of deep learning methods have led to remarkable progress in the application of machine learning methods in dentistry. Across multiple approaches in dental image analysis, almost all used Convolutional Neural Networks (CNNs) as backbone feature extractors, with ResNet, VGG, and GoogLeNet being popular architectures. [10] More traditional approaches often combine a CNN with an SVM classifier in order to determine if an X-Ray image is "normal" or "anomalous", [12] while more cutting edge approaches may utilize Region-based CNN (R-CNN) and Feature Pyramid Networks (FPN) to extract regions of interest that enable separate diagnosis for each individual tooth. [6] Additionally, existing methods often take advantage of transfer learning, which involves fine-tuning a pretrained deep model (usually CNNs) that have learned relevant features from a larger image dataset. [10]

However, traditional hybrid approaches (such as CNN + SVM) are usually limited to providing one positive diagnosis per image, [12] and shallower CNNs often come with mediocre levels of detection accuracy on X-ray. [2] There was also little attempt made at streamlining disease detection and automatic enumeration on a single panoramic X-ray. [3] Therefore, our approach aims at combining the insights from existing approaches while addressing their limitations.

III. Dataset and Features

The overall dataset used in our project is sourced from the DENTEX benchmark dataset for abnormal tooth detection and diagnosis [4], which consists of grayscale X-ray images as raw features and annotations for these images as labels. A different subset of images is used for each task of the project: 693 for quadrant detection, 634 for enumeration, and 705 for disease detection and classification. The images in the raw dataset vary in aspect ratio and resolution, but the machine learning algorithms we used in the project expect inputs of fixed sizes. Therefore, to ensure data compatibility across models, we resized images to 960 by 448 pixels for quadrant detection and enumeration classification, and extracted image regions of size 150 by 300 for disease classification. Some examples of these cropped images are shown farther below. It can be easily seen that the X-ray images are noisy and the identification of caries, lesions, or impaction is highly nontrivial even for humans.

Annotations for these images are provided in the COCO dataset format. We first preprocessed all bounding box coordinates to $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$ coordinates with respect to the resized images to better suit PyTorch's workflow, and we parsed information in the annotation string in data structures expected by each model architecture before training. We used a 90-10 training-validation split for enumeration and quadrant detection, and an 80-20 split for classification.

Additional data cleaning and processing techniques will be outlined in detail in our **Methods** section.

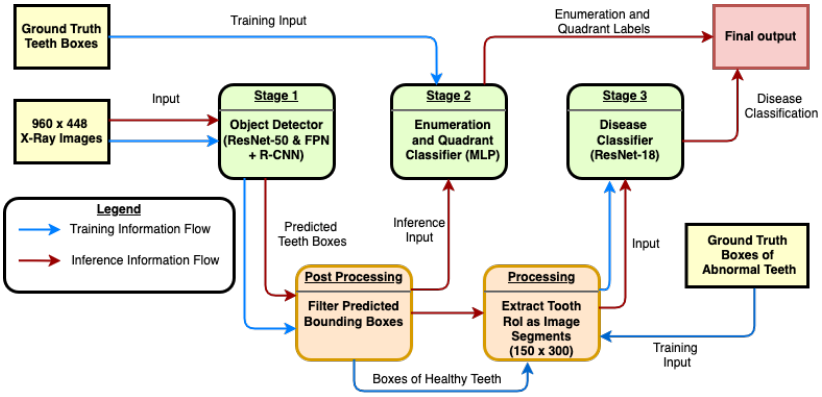

Figure 1: Overview of Model Pipeline

IV. Methods

A. *Initial Experiments & Overview of Final Approach*

We first considered using a two-layer feed-forward neural network (with 256 hidden units and ReLU activation) and a convolutional neural network inspired by the LeNet architecture [7] (2 convolution layers & 3 fully connected layers) for quadrant detection. Both networks take the resized images as inputs, and outputs a vector of 16 coordinate values for 4 bounding boxes. These models were trained for 40 epochs using the SGD optimizer. The CNN model required a learning rate of 0.001, while the MLP model required a larger learning rate of 0.01 to converge.

We initially considered using a two-staged approach for enumeration: An image is first passed through a binary classifier, giving us information about whether each tooth exists in the image. Then, we used this formation, combined with predictions given by 32 bounding box regressors, one for each tooth, to obtain enumerated bounding boxes. However, due to the shallower networks' tendencies to underfit and inability to learn complex image features, as well as the subpar performance of binary classifiers as a result of severe class imbalances, we decided to abandon this approach in favor of a multi-stage approach with both deep-learning and classical ML components as outlined in Figure 1.

This final approach would utilize an object detector that identifies all teeth in an image with no enumeration and disease information, and then feed the outputs of the object detector as the input to downstream models that will "fill in" the enumeration and disease classification as needed.

B. *Stage 1: Tooth Detection*

We fine-tuned a Faster R-CNN object detection model [11, 1] with pretrained ResNet-50 [5] and Feature Pyramid Networks [8] backbone feature extractors as our bespoke tooth detection model. The backbone modules were trained on the Common Objects in Context (COCO) dataset [9], and we fine-tuned the object detection model for 50 epochs with the enumeration dataset. (via the stochastic gradient descent / SGD optimization method, batch size of 16 and a learning rate of 0.0001) The fine-tuning task were carried out on the Google Colab cloud computing platform and took 3.5 hours of compute time on an Nvidia T4 GPU.

C. *Stage 2: Enumeration Classification*

We trained several multi-layer perceptron (MLP) models with various levels of model complexity. These models were trained using the ground-truth bounding box coor-

Impacted Caries Lesions Deep Caries

Figure 2: Randomly Selected Training Examples for Classification Model with Dental Anomalies

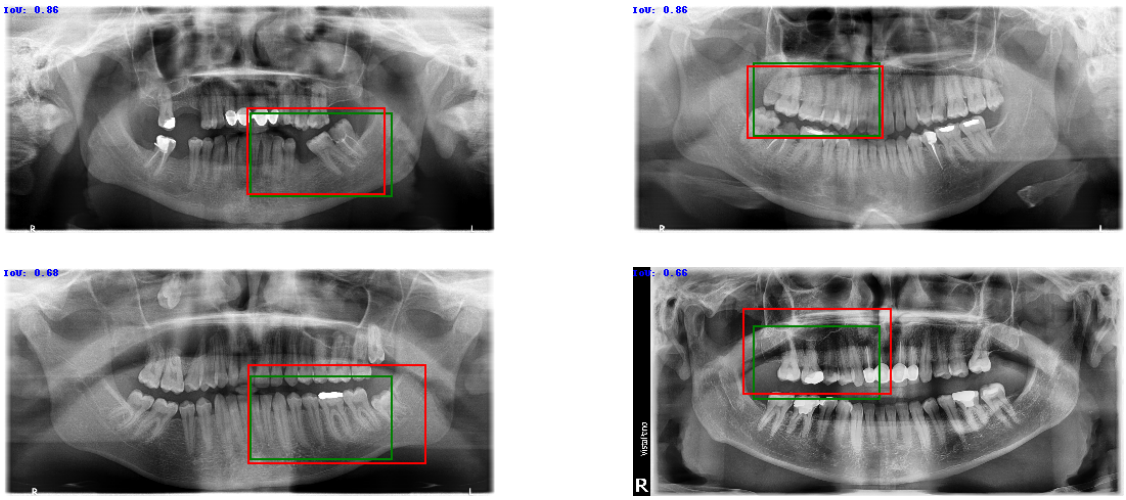

Figure 3: Underfitted Quadrant Bounding Boxes (Greed = Predicted, Red = Ground Truth)

dinates from the enumeration dataset as inputs. Models were trained for 30 epochs during candidate model selection and class label resolution study, and 50 epochs during final model selection with all 32 quadrant-enumeration classes. The SGD optimizer with a batch size of 16, a learning rate of 0.005 and momentum 0.9 was used for this section.

We were able to further increase the quality of enumeration of taking advantage of the spatial arrangement of teeth and the fact that each enumeration label can be matched to at most one bounding box in an image. This was implemented as a post-processing step during inference time only. The goal was to maximize the "quality" of assignments between labels and boxes given the output of our model and our prior knowledge about the space of possible labels. The softmax layer of the MLP is taken as the "quality" of a potential assignment. This optimization problem can be formally formulated as:

$$\begin{aligned} \max \quad & \sum_{i=1}^{n} P(i, x_i) \\ & \text{subject to:} \\ & x_i \in \{1, 2, \ldots, 16\} \\ & \qquad (x_i \text{ is the label assignment for box } i) \\ & x_a < x_b \quad \forall a, b \quad \text{s.t.} \quad \text{pos}[a] < \text{pos}[b] \end{aligned}$$

where n is the number of boxes, $P(i, j)$ is the softmax for the label corresponding to enum class j for tooth i, and pos[i] is the horizontal position for bounding box i. This optimization problem was solved via a dynamic program for both upper and lower rows of teeth.

D. Stage 3: Disease Classification

Being a relatively simpler task, our architecture for disease classification was much simpler than the quadrant detection/enumeration. ResNet-50 was initially considered for this task, but preliminary experiments showed a tendency to overfit the training dataset and a poorer f-score on the validation set. Thus, we elected to use a pretrained ResNet-18 CNN architecture for image classification, which was trained using Adam optimizer to minimize cross-entropy loss. It was trained for 40 epochs with a learning rate of 0.0005.

Our classification model takes in 150 by 300 Region of Interest segments of singular teeth extracted from the panoramic xray images. (See Figure 2) These are cropped using the bounding boxes generated from the object detection model. The training set includes not only the image tensors, but also label tensors with four boolean values each corresponding to a disease (with all false indicating a healthy tooth). During preprocessing, each tooth image is placed into a folder labeled healthy, caries, lesions, etc. depending on its label tensor. The images are resized to 224 by 224 to match the ResNet-18 input size and the pixel values are normalized. As an output, the model produces the predicted class label ("Caries," "Healthy," etc.)

The original dataset came with 4 different types of dental anomalies: Caries, Lesion, Impacted and Deep Caries. We have decided to merge the Caries and Deep Caries classes due to semantic similarities and a shortage of training examples with the Deep Caries diagnosis.

V. Experiments / Results / Discussion

A. Preliminary Study on Quadrant Detection

We initially treated the quadrant detection task as a multi-dimensional regression problem and experimented with two shallow networks. At convergence, both MLP and CNN have identical mean intersection over union metrics ranging from 0.695 to 0.733 on the validation set, and we found our that both models learned almost identical underfitted regressors. The learned bounding boxes are adequate for most images, but we did not notice visible variations in these bounding boxes across images. We have attached some prediction samples in Figure 3.

B. Object Detector Post Processing

Since we would like to use the predicted bounding boxes given by the object detector for downstream tasks, it is imperative to minimize the number of false-positive detections while also try reducing false-negatives. We can achieve this by filtering our the bounding boxes with low confidence scores. While it is possible to count false positives and false negatives by exhaustively checking pairwise overlaps between predicted and ground-truth bounding boxes, the computationally prohibitive nature of said method prompted us to estimate these quantities by counting the absolute difference in count between ground-truth and predicted boxes as false positives if the model predicts more boxes than actual, and as false negatives if otherwise. These counts are aggregated over the entire training plus validation set. Our findings in Table 1 shows that a filter threshold between 0.65 and 0.75 would be reasonable to strike a balance between FP minimization and FN minimization.

C. Model Selection for Quadrant-Enumeration Classifier

We decided to combine the quadrant and enumeration labeling tasks as a single classification task, where our goal is to predict its quadrant label (0 - 3) and enumeration label (0 - 7) given the bounding box of a tooth. Spatial characteristics such as location and shape of bounding boxes are highly predictive of the associated quadrant and enumeration information, and the relatively low difficulty of this task suggests that shallow networks would be sufficient. Therefore, we used a multinomial logistic regression (LR) model as a baseline and compared it against MLP classifiers. Initially, we saw classifiers performing poorly on the full problem with all 32 labels, so we started with coarse-grained classes (adjacent teeth were designated to have the same label) and documented evolution of model performance as the class resolution increase in Figure 4. We used prediction accuracy on the **validation set** as evaluation metric.

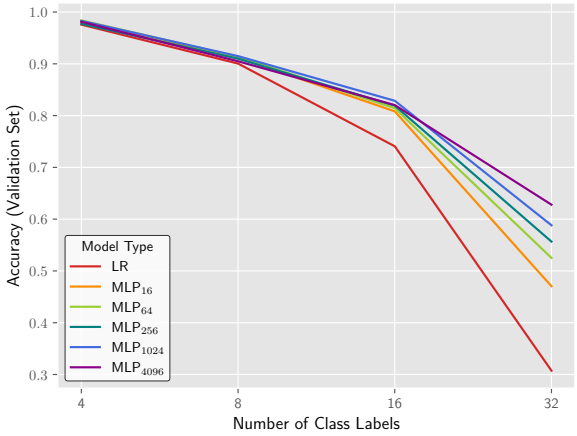

Figure 4: Validation Accuracy and Class Label Resolution

For easy tasks such as quadrant classification (with 4 classes total), even our simple baseline model would be able to achieve a satisfactory accuracy of $\geq 97.5\%$. However, there is a non-trivial performance gap between LR and MLPs for more fine-grained classification tasks. For the full problem with all 32 class labels, MLPs with higher model complexities also performed better. Since none of the models showed signs of overfitting, we decided to increase the model complexity to try to improve performance. We did this by constructing a 3-layer MLP with 4096 and 128 neurons in the 2 hidden layers.

We proposed 4 candidate models (Baseline LR, MLP with 256 hidden neurons, MLP with 4096 hidden neurons, and 3-Layer MLP) for the final model selection stage for the full enumeration tasks. These models were trained for 50 epochs before being evaluated using top-1 accuracy and top-3 accuracy (i.e. the fraction of predictions where the ground truth corresponds to one of the top 3 predicted class probabilities). Results are documented in Figure 5.

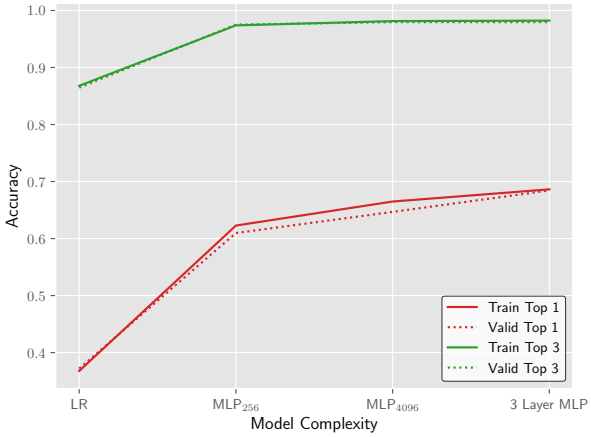

Figure 5: Accuracy and Model Size with Full Class Labels

Our experimentation showed that adding model complexity lead to non-trivial increases in top-1 accuracy. However, as the top-3 accuracy shows, if the model's down-stream application is able to tolerate some uncertainty, then all MLP candidate models would be sufficient. We selected the 3-layer MLP as our final stage 2 model, with a top-1 accuracy of 68% over all data for which enumeration labels are available. However, the post-processing algorithm taking advantage of spatial information increased this to 84%.

D. Experiments on Disease Classification

Our tooth classification model was evaluated based on the macro average f1-score on our validation set of about 5100 labeled tooth image segments. The f1-score of a predictor is defined accordingly:

$$F_1 = \frac{2TP}{2TP + FP + FN}$$

Threshold	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
Potential FP	663	500	389	316	245	183	133	75	38	7
Potential FN	117	161	218	270	331	408	521	698	1065	2406
FP + FN	780	661	607	586	576	591	654	773	1103	2413

Table 1: Tuning Filtering Threshold for Object Detection Model

where TP, FP, FN are the numbers of true positives, false positives, and false negatives, respectively. It is the harmonic mean of the precision (how many "positive" predictions made were correct) and the recall (how many "positive" class samples in the dataset were correctly identified) of the predictor. This is calculated for each of the four classes, and then averaged to compute the macro average f1-score. This metric encourages a balance between precision and recall, and is widely used to evaluate classification models.

We began with testing a pretrained ResNet-18 model as a baseline. The baseline model used Adam optimizer along with an initial learning rate of 0.001, and was trained for 40 epochs using a batch size of 32 to speed up training. This gave us a baseline macro average f1-score of 92%. Switching to ResNet-50 (a much deeper model) and changing the batch had a negligible effect on performance, so both remained untouched for the rest of our experiments. Replacing the model's optimizer with SGD brought down the f1-score by approximately 4%, and thus it also remained unchanged for our final model. Tweaking the learning rate of our model provided our model with a slight increase in performance. After decreasing it to 0.0005, we saw a 3% improvement in the model's macro average f1-score.

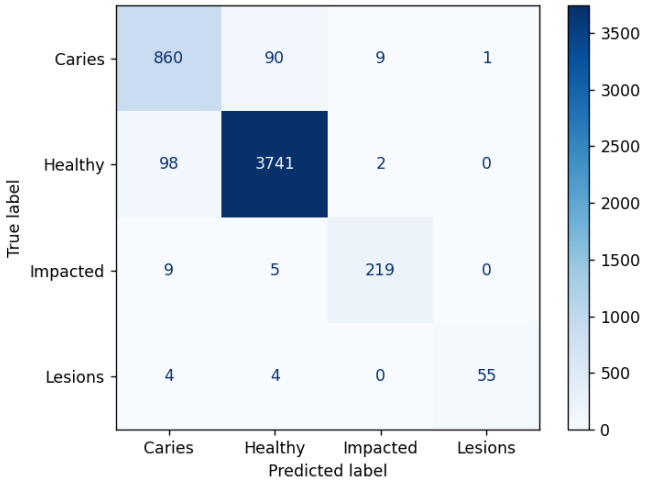

Figure 6: Confusion Matrix of Final Classification Model

E. *Final Results*

The final classification model, using a pretrained ResNet-18 architecture with a learning rate of 0.0005, obtained a final macro average f1-score of 95%. Figure 6 shows the confusion matrix of our final model for the validation set. The "Caries" class consistently shows a lower f1-score than the other classes, which is likely due to the greater variation in the X-rays of teeth with cavities compared to those that are impacted or have lesions.

VI. Conclusion / Future Work

We propose a 3-stage method for dental X-ray enumeration and diagnosis which aims to assist dentists by automating tooth detection, enumeration and disease classification. Through model architecture selection, fine-tuning and hyper-parameter search, we put together a pipeline that is accurate in both enumeration and classification tasks and robust to X-rays with widely varying numbers of missing teeth. An example prediction output is attached in Figures 7, 8.

During inference, we found out that our disease classification model could trigger false positives for a tooth if an adjacent tooth is has an anomaly, which could be fixed by removing irrelevant features with instance segmentation. Other limitations of our approach include the inability for the current classifier to handle abnormal cases in which more than 32 teeth are present, or rare cases where there are multiple diagnoses for a single tooth.

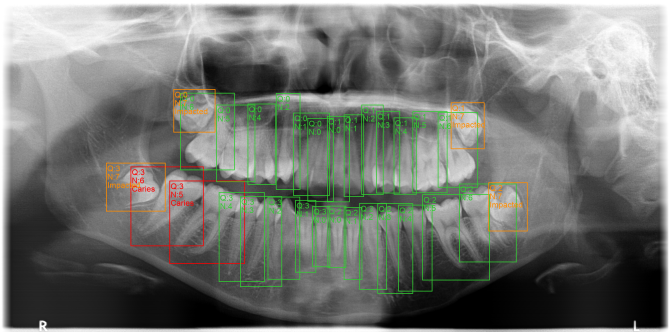

Figure 7: Prediction Output Sample

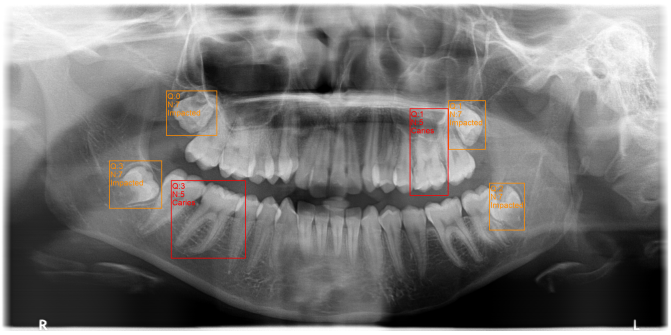

Figure 8: Corresponding ground-truth diagnosis

VII. Appendix

Contributions

In general, all team members worked together in the proposal of problem solving pipeline, selection of model architecture, and collection / analysis of experimental data. Additionally, team members worked together to streamline the code, and cross-checked for errors. For contributions for specific tasks, consult the sequential list below.

1. Data Processing: Songyu (Enumeration), James & Sawyer (Classification)
2. Initial Modeling and Experimentation for Quadrant: Songyu & Sawyer (GPU Code)
3. Stage 1 (Object Detection) Model Design and Fine-Tuning: James & Songyu
4. Stage 2 (Enumeration) Model Design and Experimentation: Songyu & James (post-processing)
5. Stage 3 (Disease Classification) Model Design and Experimentation: Sawyer & James
6. Model Inference Pipeline and Visualization: Songyu & James
7. Report: Songyu, James & Sawyer (Equally)

References

[1] Torchvision: fasterrcnn_resnet50_fpn documentation. `https://pytorch.org/vision/main/models/generated/torchvision.models.detection.fasterrcnn_resnet50_fpn.html`. Accessed: Feb 25, 2025.

[2] Thomas Ekert, Joachim Krois, Leonie Meinhold, Karim Elhennawy, Ramy Emara, Tatiana Golla, and Falk Schwendicke. Deep learning for the radiographic detection of apical lesions. *Journal of Endodontics*, 45(7):917–922.e5, 2019.

[3] Ibrahim Ethem Hamamci, Sezgin Er, Enis Simsar, Anjany Sekuboyina, Mustafa Gundogar, Bernd Stadlinger, Albert Mehl, and Bjoern Menze. Diffusion-based hierarchical multi-label object detection to analyze panoramic dental x-rays, 2023.

[4] Ibrahim Ethem Hamamci, Sezgin Er, Enis Simsar, Atif Emre Yuksel, Sadullah Gultekin, Serife Damla Ozdemir, Kaiyuan Yang, Hongwei Bran Li, Sarthak Pati, Bernd Stadlinger, Albert Mehl, Mustafa Gundogar, and Bjoern Menze. Dentex: An abnormal tooth detection with dental enumeration and diagnosis benchmark for panoramic x-rays, 2023.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[6] Gil Jader, Jefferson Fontineli, Marco Ruiz, Kalyf Abdalla, Matheus Pithon, and Luciano Oliveira. Deep instance segmentation of teeth in panoramic x-ray images. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 400–407, 2018.

[7] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[8] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2017.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

[10] Mónica Vieira Martins, Luís Baptista, Henrique Luís, Victor Assunção, Mário-Rui Araújo, and Valentim Realinho. Machine learning in x-ray diagnosis for oral health: A review of recent progress. *Computation*, 11(6), 2023.

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

[12] Dhruv Verma, Sunaina Puri, Srikanth Prabhu, and Komal Smriti. Anomaly detection in panoramic dental x-rays using a hybrid deep learning and machine learning approach. In *2020 IEEE REGION 10 CONFERENCE (TENCON)*, pages 263–268, 2020.