PGCOMP - Programa de Pós-Graduação em Ciência da Computação
Universidade Federal da Bahia (UFBA)
Av. Milton Santos, s/n - Ondina
Salvador, BA, Brasil, 40170-110

https://pgcomp.ufba.br
pgcomp@ufba.br

Dental panoramic radiographs are not only a highly valuable exam but also a versatile one. They can be used to diagnose periodontal bone loss, lesions, cysts, and tumors, as well as estimate the age and biological sex of the patient. The works that use deep learning to determine such conditions in panoramic radiographs are based on supervised approaches that require manual annotation of each attribute and condition considered. However, manual annotation of radiographs is demanding, as it requires qualified labor and is, consequently, expensive. This work seeks to overcome this difficulty by exploring the Human-in-the-Loop concept, a semi-supervised learning technique that expedites the labeling process through an interaction between human experts and machine learning models. To support this approach, special focus was given to teeth, as they are the main objects of attention and reference points for radiologists when reading panoramic radiographs. As a result, a dataset for tooth instance segmentation of panoramic radiographs was produced: the O^{2}PR dataset, containing 4,000 images. The remaining data of work consists of 4,795 radiographs in the Raw Panoramic Radiographs (RPR) dataset, with images in their crude format, and the Textual Report Panoramic Radiographs (TRPR) dataset, containing 8,029 pairs of radiograph images and textual reports. These groups of data comprise the most extensive dataset in the literature. Starting from these datasets, we classify thirteen dental conditions in the tooth or its surroundings. To classify all the considered conditions, a holistic approach was necessary. First, using the labeled radiographs of the O^{2}PR dataset, we trained an instance segmentation neural network to pseudolabel the teeth in the unlabeled radiographs. Subsequently, all tooth images were cropped to facilitate the classification of dental conditions. The O^{2}PR and RPR datasets do not include textual reports, making it impossible to generate labels for training or evaluating these images for dental conditions. Instead, the tooth crops from these datasets were used to pre-train Vision Transformers (which were later employed as classification networks for dental conditions) through a self-supervised learning technique called Masked Autoencoders. This approach proved effective as it allowed the use of unlabeled data to improve performance. The label extraction procedure follows a different branch. We explored the API of a Large Language Model, GPT-4, to avoid the pure manual labeling of the dental conditions. The goal of using it was to identify the noun phrases from the textual reports to find the dental conditions. Later, a heuristic associated each tooth present in the report sentences with all the dental conditions of the same sentence. We leverage the pretrained Vision Transformer to train several dental condition classification models. Encouragingly, the results consistently met or surpassed the baseline metrics for the Matthews correlation coefficient. A comparison of the proposed solution with human practitioners, supported by statistical analysis, highlighted its effectiveness and performance limitations; based on the degree of agreement among specialists, the solution demonstrated an accuracy level comparable to that of a junior specialist.

Keywords: deep learning; dental panoramic radiographs; instance segmentation models; large language models;

DSC | XXX | 2024

From Build-Up to Solid Foundations: Exploring Deep Learning for Classifying Dental Conditions on Panoramic Radiographs

Bernardo Peters Menezes Silva

UFBA

From Build-Up to Solid Foundations: Exploring Deep Learning for Classifying Dental Conditions on Panoramic Radiographs

Bernardo Peters Menezes Silva

Tese de Doutorado

Universidade Federal da Bahia

Programa de Pós-Graduação em
Ciência da Computação

Setembro | 2024

Universidade Federal da Bahia
Instituto de Computação

Programa de Pós-Graduação em Ciência da Computação

FROM BUILD-UP TO SOLID FOUNDATIONS: EXPLORING DEEP LEARNING FOR CLASSIFYING DENTAL CONDITIONS ON PANORAMIC RADIOGRAPHS

Bernardo Peters Menezes Silva

TESE DE DOUTORADO

Salvador
13 de Novembro de 2024

BERNARDO PETERS MENEZES SILVA

FROM BUILD-UP TO SOLID FOUNDATIONS: EXPLORING DEEP LEARNING FOR CLASSIFYING DENTAL CONDITIONS ON PANORAMIC RADIOGRAPHS

Esta Tese de Doutorado foi apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia, como requisito parcial para obtenção do grau de Doutor em Ciência da Computação.

Orientador: Luciano Rebouças de Oliveira
Co-orientadora: Patricia Ramos Cury

Salvador
13 de Novembro de 2024

Sistema de Bibliotecas - UFBA

XXXX Silva, Bernardo Peters Menezes.
From Build-Up to Solid Foundations: Exploring Deep Learning for Classifying Dental Conditions on Panoramic Radiographs / Bernardo Peters Menezes Silva – Salvador, 2024.
84p.: il.

Orientador: Prof. Dr. Luciano Rebouças de Oliveira.
Co-orientadora: Prof. Dr. Patricia Ramos Cury.
Tese (Doutorado) – Universidade Federal da Bahia, Instituto de Computação, 2024.

1. Dental Paranoramic Radiograph. 2. Human-in-the-Loop. 3. Large Language Models. I. Olivera, Luciano Rebouças de. II. Cury, Patricia Ramos. III. Universidade Federal da Bahia. Instituto de Computação. IV. Título.

CDU – YYY.YY.YYY

TERMO DE APROVAÇÃO

BERNARDO PETERS MENEZES SILVA

FROM BUILD-UP TO SOLID FOUNDATIONS: EXPLORING DEEP LEARNING FOR CLASSIFYING DENTAL CONDITIONS ON PANORAMIC RADIOGRAPHS

Esta Tese de Doutorado foi julgada adequada à obtenção do título de Doutor em Ciência da Computação e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia.

13 de Novembro de 2024

Prof. Dr. Luciano Rebouças de Oliveira
Universidade Federal da Bahia

Prof. Dr. Rodrigo de Melo Souza Veras
Universidade Federal do Piauí

Prof. Dr. Thiago Oliveira dos Santos
Universidade Federal do Espírito Santo

Prof. Dr. João Paulo Papa
Universidade Estadual Paulista

Profa. Dra. Flávia Caló de Aquino Xavier
Universidade Federal da Bahia

Aos meus amigos e familiares

ACKNOWLEDGEMENTS

Ninguém me disse que seria fácil. Eu tinha consciência disso. No entanto, eu jamais imaginei que seria tão difícil...

Terminar um doutorado foi mais um trecho árduo na trajetória da minha vida e não será o último. Ainda bem! Só escapamos das dificuldades quando morremos. Essa jornada durou pelo menos quatro revoluções e meia da Terra ao redor do Sol e teve baixos, muitos baixos, mas terminou no seu ápice.

Este capítulo da minha começou com um outro curso de doutorado, em engenharia elétrica. Minha baixa inclinação junto ao crescente descontentamento com a área me fazia ter crises de ansiedade. Mas eu queria ser professor, seguir a carreira acadêmica. Por isso insistia em fazer doutorado em engenharia elétrica, onde eu já tinha graduação e mestrado. Para aliviar minha aflição, comecei a fazer uma graduação EaD em ciência da computação, campo que eu adorava. Porém, a insatisfação nesse doutorado era grande demais, enorme, tornou-se insuportável. Foi aí que aconteceu o ponto de ruptura.

Largar o doutorado em engenharia elétrica não era apenas uma opção; tornou-se uma necessidade. Mas muita coisa precisava ser feita, a mais básica de todas: encontrar um orientador em ciência da computação. Foi aí que o professor Luciano me acolheu e uma boa parceria se formou.

Luciano acreditou no meu potencial e espero que eu tenha atendido suas expectativas. Como todos orientadores e orientandos nem sempre nos entendemos, mas no nosso caso foram, em sua maioria, sobre estrutura e título de artigos, e número de disciplinas a serem cursadas. No final das contas, essas experiências contribuíram imensamente para meu crescimento pessoal e profissional, e sou grato pelo tempo e paciência que ele investiu em minha formação.

Durante o período do doutorado, aconteceram muitas coisas. A pior delas foi uma pandemia. Isso impactou meu trabalho, mas nada que se compare a morte de milhões de pessoas. Muitas vidas foram salvas devido à ciência e aos cientistas. Esse contexto reforçou em mim a importância da ciência como meio de transformar o mundo, salvar vidas e amenizar sofrimentos, alimentando ainda mais meu desejo de me tornar professor e pesquisador. Espero que esse desejo se torne realidade e que eu possa fazer diferença na vida das pessoas.

Ao longo da jornada do doutorado, e também da vida, muitas pessoas desempenharam um papel essencial na minha trajetória. Por isso, gostaria de registrar meus agradecimentos:

Aos meus pais, Itamar e Maristela, pelo amor incondicional. Vocês são pessoas maravilhosas que renunciaram praticamente todas as suas vidas para cuidarem dos filhos. Essa infinita dedicação reflete o amor também infinito que tenho por vocês.

Aos meus irmãos Caroline, Fernanda e Leonardo, por todo o companheirismo, suporte e lealdade a mim. Também agradeço a todos os sobrinhos que vocês me deram: Gabriel, Júlio, Leonardo, Luísa e Mariana.

Aos amigos do Colégio São Paulo, de engenharia elétrica, do laboratório IVision, da vida... por todos os momentos compartilhados. A amizade de vocês fez com que essa caminhada fosse menos pesada e muito mais significativa.

À Universidade Federal da Bahia e aos órgãos de fomento que tornaram essa pesquisa possível, proporcionando recursos e oportunidades para que eu pudesse me dedicar ao meu trabalho.

A minha amiga e coorientadora, Patricia Cury, por todo o suporte, assistência e encorajamento. Conseguimos transformar uma parceria improvável, que parecia destinada ao fracasso, em diversas produções bem-sucedidas.

Ao amigo e orientador Luciano Rebouças, por confiar no meu potencial e me conduzir com paciência e sabedoria durante toda essa jornada. Sua orientação paciente e suas críticas construtivas foram fundamentais para o desenvolvimento do meu trabalho e para o meu amadurecimento pessoal e profissional.

Obrigado a todos vocês!

Ninguém me disse que seria fácil. Eu tinha consciência disso. No entanto, eu jamais imaginei que seria tão difícil, mas ficou mais fácil com vocês ;)

Think of the rivers of blood spilled by all those generals and emperors so that, in glory and triumph, they could become the momentary masters of a fraction of a dot

—PALE BLUE DOT (CARL SAGAN)

RESUMO

As radiografias panorâmicas dentárias não são apenas exames altamente valiosos, mas também versáteis. Elas podem ser utilizadas para diagnosticar perda óssea periodontal, lesões, cistos e tumores, além de estimar a idade e o sexo biológico do paciente. Os trabalhos que aplicam *deep learning* para determinar essas condições em radiografias panorâmicas se baseiam em abordagens supervisionadas que exigem a anotação manual de cada atributo e condição considerada. No entanto, a anotação manual dessas radiografias é exigente, pois demanda mão de obra qualificada, sendo, consequentemente, cara. Este trabalho busca superar essa dificuldade ao explorar o conceito de *Human-in-the-Loop*, uma técnica de aprendizado semi-supervisionado que acelera o processo de rotulagem por meio de uma interação entre especialistas humanos e modelos de aprendizado de máquina.

Para apoiar essa abordagem, deu-se foco especial aos dentes, por serem os principais objetos de atenção e pontos de referência para os radiologistas ao interpretar radiografias panorâmicas. Como resultado, foi produzido um conjunto de dados para segmentação de instâncias de dentes em radiografias panorâmicas: o conjunto O^{2}PR, contendo 4.000 imagens. Os demais dados do trabalho incluem 4.795 radiografias no conjunto *Raw Panoramic Radiographs* (RPR), com imagens em formato bruto, e o conjunto *Textual Report Panoramic Radiograph*s (TRPR), contendo 8.029 pares de imagens de radiografias e relatórios textuais. Esses grupos de dados compõem o maior conjunto de dados da literatura. Com base nesses conjuntos, classificamos treze condições dentárias presentes nos dentes ou em seus arredores.

Para classificar todas as condições consideradas, foi necessária uma abordagem holística. Primeiro, utilizamos as radiografias anotadas do conjunto O^{2}PR para treinar uma rede neural de segmentação de instâncias, a fim de pseudo-rotular os dentes nas radiografias não anotadas. Em seguida, todas as imagens dos dentes foram recortadas para facilitar a classificação das condições dentárias. Os conjuntos O^{2}PR e RPR não incluem relatórios textuais, impossibilitando a geração de rótulos para treinamento ou avaliação dessas imagens quanto a condições dentárias. Em vez disso, os recortes de dentes desses conjuntos foram usados para pré-treinar Vision Transformers (que posteriormente foram empregados como redes de classificação para as condições dentárias) por meio de uma técnica de aprendizado autossupervisionado chamada Masked Autoencoders. Essa abordagem se mostrou eficaz, pois permitiu o uso de dados não anotados para melhorar o desempenho.

O procedimento de extração de rótulos segue uma linha diferente. Exploramos a API de um Grande Modelo de Linguagem, o GPT-4, para evitar a rotulagem puramente manual das condições dentárias. O objetivo de sua utilização foi identificar os sintagmas nominais nos relatórios textuais para encontrar as condições dentárias. Em seguida,

uma heurística associou cada dente mencionado nas sentenças do relatório a todas as condições dentárias presentes na mesma sentença. Aproveitamos o Vision Transformer pré-treinado para treinar vários modelos de classificação de condições dentárias. De forma encorajadora, os resultados consistentemente atingiram ou superaram as métricas de referência para o coeficiente de correlação de Matthews. A comparação da solução proposta com profissionais humanos, respaldada por análise estatística, destacou sua eficácia e limitações de desempenho; com base no grau de concordância entre especialistas, a solução demonstrou um nível de precisão comparável ao de um especialista júnior.

Palavras-chave: *deep learning*; radiografias panorâmicas dentárias; modelos de segmentação por instância; grande modelos de linguagem;

ABSTRACT

Dental panoramic radiographs are not only a highly valuable exam but also a versatile one. They can be used to diagnose periodontal bone loss, lesions, cysts, and tumors, as well as estimate the age and biological sex of the patient. The works that use deep learning to determine such conditions in panoramic radiographs are based on supervised approaches that require manual annotation of each attribute and condition considered. However, manual annotation of radiographs is demanding, as it requires qualified labor and is, consequently, expensive. This work seeks to overcome this difficulty by exploring the Human-in-the-Loop concept, a semi-supervised learning technique that expedites the labeling process through an interaction between human experts and machine learning models. To support this approach, special focus was given to teeth, as they are the main objects of attention and reference points for radiologists when reading panoramic radiographs. As a result, a dataset for tooth instance segmentation of panoramic radiographs was produced: the O^{2}PR dataset, containing 4,000 images. The remaining data of work consists of 4,795 radiographs in the Raw Panoramic Radiographs (RPR) dataset, with images in their crude format, and the Textual Report Panoramic Radiographs (TRPR) dataset, containing 8,029 pairs of radiograph images and textual reports. These groups of data comprise the most extensive dataset in the literature. Starting from these datasets, we classify thirteen dental conditions in the tooth or its surroundings. To classify all the considered conditions, a holistic approach was necessary. First, using the labeled radiographs of the O^{2}PR dataset, we trained an instance segmentation neural network to pseudolabel the teeth in the unlabeled radiographs. Subsequently, all tooth images were cropped to facilitate the classification of dental conditions. The O^{2}PR and RPR datasets do not include textual reports, making it impossible to generate labels for training or evaluating these images for dental conditions. Instead, the tooth crops from these datasets were used to pre-train Vision Transformers (which were later employed as classification networks for dental conditions) through a self-supervised learning technique called Masked Autoencoders. This approach proved effective as it allowed the use of unlabeled data to improve performance. The label extraction procedure follows a different branch. We explored the API of a Large Language Model, GPT-4, to avoid the pure manual labeling of the dental conditions. The goal of using it was to identify the noun phrases from the textual reports to find the dental conditions. Later, a heuristic associated each tooth present in the report sentences with all the dental conditions of the same sentence. We leverage the pretrained Vision Transformer to train several dental condition classification models. Encouragingly, the results consistently met or surpassed the baseline metrics for the Matthews correlation coefficient. A comparison of the proposed solution with human practitioners, supported by statistical analysis, highlighted its effectiveness and performance limitations; based on the degree of agreement among

specialists, the solution demonstrated an accuracy level comparable to that of a junior specialist.

Keywords: deep learning; dental panoramic radiographs; instance segmentation models; large language models;

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AI	Artificial Intelligence	3
AP	Average Precision	39
HITL	Human-In-The-Loop	75
O^{2}PR	OdontoAI Open Panoramic Radiographs	75
MAE	Masked Autoencoders	76
TMJ	temporomandibular joint	2
ViT	Vision Transformers	35
mAP	mean Average Precision	39
AP	average precision	39

Chapter 1

INTRODUCTION

1.1 OVERVIEW

Oral health is a fundamental component of overall well-being, yet it remains a significant public health challenge worldwide. The World Health Organization (WHO) estimates that nearly 3.5 billion people suffer from oral diseases, making them among the most common health conditions globally (WHO, 2022). These conditions, which include tooth decay, periodontal disease, and oral cancer, can affect individuals across all age groups, leading to pain, discomfort, and social and psychological stress. Poor oral health not only reduces quality of life, but also contributes to other systemic health problems, such as cardiovascular disease, diabetes, and respiratory infections (TONETTI et al., 2017).

The burden of oral diseases is unevenly distributed, with low- and middle-income countries experiencing the highest prevalence. Limited access to dental care, inadequate public health policies, and socio-economic disparities contribute to the high rates of untreated oral conditions in these regions. Furthermore, lifestyle factors such as poor diet, tobacco use, and inadequate oral hygiene exacerbate the problem, creating a cycle of worsening health outcomes. The economic implications are also substantial, as dental treatments are costly, and the loss of productivity due to oral health problems further strains healthcare systems and economies.

Addressing the global oral health problem requires advancements in early detection, diagnosis, and treatment. In this context, medical imaging significantly impacts dentistry, allowing specialists to identify problems that might not be visible during a clinical examination. Modalities such as X-rays, computerized tomography scans, and magnetic resonance imaging provide detailed views of teeth, bones, and soft tissues (WHITE; PHAROAH, 2014). These tools enhance the precision of diagnoses and treatments, ensuring better patient outcomes. Among the current imaging exams, radiographs are the most common in dentistry (WHITE; PHAROAH, 2014; LANGLAIS; MILLER, 2016), being requested to identify various pathologies like cavities, periodontal disease, impacted teeth, and bone infections (CHANG et al., 2020; YÜKSEL et al., 2021) and track the progress of dental treatments.

(a) Sample of a periapical radiograph. (b) Sample of a bitewing radiograph.

Figure 1.1: Example of the two most common intraoral radiographs: periapical and bytewing. They are more focused on the teeth than the panoramic radiograph.

The three most common types of radiographs in dentistry are the periapical, bitewing, and panoramic (JADER et al., 2018). The first two are categorized as intraoral radiographs because the radiograph film is placed inside the mouth during the image retrieval. At the same time, the panoramic radiograph is classified as an extraoral radiograph because the film or sensor is outside the mouth during the during image capture. This difference results in more focused and detailed images in the intraoral case than in the extraoral one. Figures 1.1 (a) and (b) show periapical and bitewing radiograph samples, respectively. The comparison with the panoramic radiograph of Figure 1.2 evinces that the intraoral radiographs are the most focused on the teeth. Therefore, periapical and bitewing radiographs are the common choices of dentists to detect cavities, analyze restorations, and evaluate the health of the tooth in general. On the other hand, dentists use dental panoramic radiographs to diagnose conditions or plan treatments that do not require fine grained details, such as evaluating wisdom teeth, temporomandibular joint (TMJ), planning orthodontic treatments or implant placement.

Among the mentioned types of radiographs, the panoramic type is the most challenging to screen. This difficulty comes not only from the number of visible structures, but also from the amount of overlap (please refer to Figure 1.2). This characteristic stems from the method used to acquire the panoramic radiograph. The acquisition of the panoramic radiograph is done through a specialized imaging technique that provides a comprehensive view of the entire oral and maxillofacial region. During the procedure, the patient is positioned in front of a rotating X-ray machine, known as a panoramic unit. The patient bites down on a bite guide to ensure proper alignment of the teeth and jaws, as shown in Figure 1.3. As the machine rotates around the patient's head, a controlled beam is emitted from an X-ray tube, passing through the patient's mouth and surrounding structures. A detector on the opposite side captures the beam that emerges, creating a continuous image as the unit and the patient move in tandem. As a result, the imaging system generates a panoramic view that captures various body parts, and displays the teeth, jaws, temporomandibular joints, and other essential structures in one two-dimensional image. The resulting radiograph is very versatile; it can assist dental

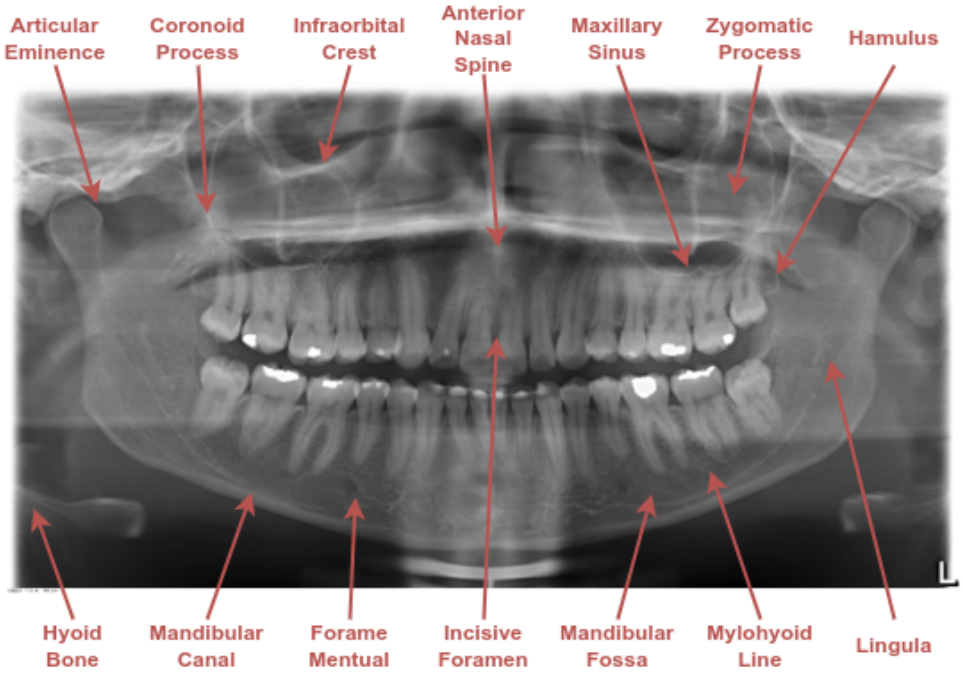


Figure 1.2: Sample of a panoramic radiograph and some of structures that are visible from it. The amount of visible structures makes this radiograph very versatile.

professionals in diagnosing various oral conditions, such as periodontal diseases, injuries, cysts, and tumors, though it can be challenging to interpret.

This difficulty has inspired many researchers to develop and propose tools to assist professionals in their work. Ten years ago, most proposed tools relied on unsupervised learning techniques and exhibited poor performance (JADER et al., 2018). More recently, deep learning, an Artificial Intelligence (AI) technique, has made a significant impact on the field. (SILVA et al., 2022). This data-driven technique has achieved significantly better results compared to previously used methods by employing several layers of neural networks.

The adjustment of the numerous parameters of neural networks first requires selecting an appropriate deep learning paradigm. Some of the deep learning paradigms commonly include supervised learning, unsupervised learning, semi-supervised learning, self-supervised learning, and reinforcement learning. Each one has its own specific applications and advantages depending on the nature of the data and the problem to be addressed. For example, supervised learning is often employed for tasks where labeled data is available, while unsupervised learning is useful for discovering hidden patterns in unlabeled data. The most successful approaches to developing AI tools in the healthcare field have utilized deep learning within a supervised learning paradigm. In this paradigm, the models are trained on labeled datasets where the input data are paired with the correct output, enabling the AI to learn from examples and make accurate predictions on new, unseen data. The disadvantage of this approach is that it requires large amounts of labeled data, which can be time-consuming and expensive to obtain.

Figure 1.3: A panoramic radiography machine also called panoramic unit. During the image acquisition, the patient stays in the center of rotation of the X-emitter and receptor. Both rotates in tandem to generate the dental panoramic radiograph (Wikipedia contributors, 2024).

This work investigates the automatic diagnosis of various tooth conditions using images and textual reports of dental panoramic radiographs and different deep learning paradigms (supervised, semi-supervised, and self-supervised). Although many structures are present on the panoramic radiograph, the teeth still play a unique role, since they serve as the primary focal points and reference points for radiologists. Beyond radiology, teeth hold significance in forensic sciences, particularly for identifying corpses. For software-based analyses, detecting, classifying, and segmenting teeth are crucial preprocessing steps for further analysis. Motivated by these considerations, this work delves into the automatic detection, classification, and segmentation of teeth.

1.2 MOTIVATION

Radiographs are priceless resources when diagnosing conditions that can not be analyzed by directly examining the patient. However, radiograph reading requires skilled labor. A radiologist's training takes numerous years and demands many skills from the professional (SILVA et al., 2020). The long shifts in this demanding job increase the risk of errors (JING; XIE; XING, 2017) Furthermore, there are several types of radiographs, each with peculiarities and challenges to read (JADER et al., 2018; SILVA et al., 2020). Developing tools that can aid professionals would be rather beneficial. Driven by these factors, this work explores the automatic detection, classification, and segmentation of teeth, along with the classification of dental conditions.

Our investigation begins in a supervised manner with the annotation of teeth on the

images. To address the slow and tedious nature of this task, we then progress to a second stage, adopting the Human-In-The-Loop (HITL) approach. Here, neural network predictions serve as provisional labels which are subsequently verified by human annotators. Using this enriched dataset, we further our investigation of several tooth conditions in combination with the noun phrases extracted from the textual reports.

Therefore, some questions guide this work:

1. How much time can we save using the HITL concept for annotating tooth instances in dental panoramic radiographs?

2. Is it possible to accurately detect, classify, and segment the teeth on dental panoramic radiographs?

3. Can we diagnose dental conditions precisely using the proposed pipeline?

In order to answer these questions, we:

1. Applied the HITL concept to expedite the annotation process of 3,150 images;

2. Performed a benchmark study to evaluate the use of two-stage detectors for detecting, classifying, and segmenting teeth through instance segmentation;

3. From the previous results, studied classifying each tooth condition separately.

1.3 GOALS

1.3.1 General goal

This study's primary goal is to contribute to the automatic dental panoramic analysis focusing on the radiologists' main targets: the teeth.

1.3.2 Specific goals

The specific goals of this thesis are:

- Employ instance segmentation two-stage detectors to automatically detect, number, and segment the teeth in a panoramic radiograph.

- Constructed a dataset to support the research community and establish a benchmark.

- Leverage all available data to train classifiers for dental conditions extracted from textual reports.

1.4 CONTRIBUTIONS

We believe that our work makes significant contributions to various aspects, ranging from data labeling to the final classification of dental conditions.

While reviewing an older dataset, we created a new one called the DNS Panoramic Images dataset. With this dataset, we conducted a benchmark of four distinct deep neural networks for instance segmentation: Mask R-CNN, PANet, HTC, and ResNeSt. This comparisson allowed us to release the new dataset and determine the best end-to-end neural network for segmenting and classifying teeth under the same conditions using the mAP metric. Our results are detailed in the following paper:

- SILVA, B.; PINHEIRO, L.; PITHON, M.; OLIVEIRA, L. (2020). A study on tooth segmentation and numbering using end-to-end deep neural networks. In 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 164-171.

Following the previous work, we segmented the teeth from scratch of 450 panoramic radiographs, considering tooth types and overlapping, resulting in the release of a new dataset (DNS Panoramic Images v2). A labeled radiograph repository for tooth instance segmentation, unmatched in size and quality in the literature. This dataset supported a benchmark, where we compared Mask R-CNNs using two configurations of segmentation heads: the traditional FCN module and the previously unused PointRend module. Our findings are detailed in the paper

- PINHEIRO, L.; SILVA, B.; PITHON, M.; OLIVEIRA, L. (2021). Numbering permanent and deciduous teeth via deep instance segmentation in panoramic x-rays. In 2021 17th International Symposium on Medical Information Processing and Analysis (SIPAIM), pp. 95-104.

We proceeded with our investigations expanding our dataset using the HITL concept. We benchmarked several instance segmentation neural networks trained from these images to fix the architecture for the HITL scheme, which we adopted to speed up the annotation process. This process resulted in our new dataset, so-called OdontoAI Open Panoramic Radiographs (O^2PR). The dataset comprises 4,000 images, from which 2,000 have their labels publicly available. We went a step further and released an online platform where researchers could submit their solutions for three different benchmarks that employ the labels of the remaining 2,000 radiographs. This access restriction was beneficial, as it will reduce assessment biases. These results are detailed in:

- SILVA, B.; PINHEIRO, L.; PITHON, M.; SOBRINHO, B.; LIMA, B.; OLIVEIRA, L.; ABDALLA, K.; CURY, P.; OLIVEIRA, L. (2023). Boosting research on dental panoramic radiographs: a challenging data set, baselines, and a task-central online platform for benchmarking. Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization, pp. 1327-1347.

We would also like to mention that we conducted some work not directly related to the main focus of the thesis.

- ANDRADE, K.; SILVA, B.; OLIVEIRA, L.; CURY, P. (2023). Automatic dental biofilm detection based on deep learning. Journal of Clinical Periodontology, 50, 571-581.
- SILVA, B.; PINHEIRO, L.; ANDRADE, K.; CURY, P.; OLIVEIRA, L. (2022). Dental Image Analysis: Where Deep Learning Meets Dentistry. In Convolutional Neural Networks for Medical Image Processing Applications (pp. 170-195). CRC Press.
- HOUGAZ, A.; LIMA, D.; PETERS, B.; CURY, P.; OLIVEIRA, L. (2023). Sex estimation on panoramic dental radiographs: A methodological approach. In Anais do XXIII Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS), pp. 115-125. SBC.
- PRADO, I.; LIMA, D.; LIANG, J.; HOUGAZ, A.; PETERS, B.; OLIVEIRA, L. (2024). Multi-Task Learning Based on Log Dynamic Loss Weighting for Sex Classification and Age Estimation on Panoramic Radiographs. In 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP), pp. 385-392.
- DA SILVA, J. F., SILVA, B., & OLIVEIRA, L. (2022, October). No boundary left behind in semantic segmentation. In 2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI) (Vol. 1, pp. 115-120). IEEE.

Finally, we would like to note that our new pre-print is currently undergoing its second round of review at the *Elsevier Medical Image Analysis* journal:

- SILVA, B.; FONTINELE, J.; VIEIRA, C. L. Z.; TAVARES, J. M. R. S.; CURY, P. R.; OLIVEIRA, L. (2024). Semi-supervised classification of dental conditions in panoramic radiographs using large language model and instance segmentation: A real-world dataset evaluation. *arXiv preprint arXiv:2406.17915*.

1.5 CHAPTER MAP

- **Chapter 2** reviews the main studies on tooth segmentation, classification, and dental condition detection. Additionally, we compare features of these studies, such as dataset size and the number of dental conditions considered, with our own work.
- **Chapter 3** describes the materials and methods used in this work, covering the starting databases, the construction of the datasets, and the adopted methodology.
- **Chapter 4** describes the steps involved in creating the O^2PR dataset for this work using the HITL concept. The Chapter also includes the evaluation and benchmarking of the instance segmentation neural networks used.

- **Chapter 5** introduces the experiments conducted using the framework developed in this study for classifying dental conditions. The Chapter also presents how the work was validated through assessments with dental specialists.
- **Chapter 6** presents the final conclusions of this work, discussing its strengths, shortcomings, applications, and future work.

Chapter 2

BACKGROUND AND RELATION WITH OUR WORK

Imaging is a fundamental tool for dentists and oral health experts who use photos, magnetic resonance imaging, ultrasound, and radiographs, among other techniques, to diagnose patients' conditions and diseases, as well as to monitor treatment progressions. When examining a panoramic radiograph, radiologists usually focus on the teeth, using them as landmarks to analyze the image and report their findings. Specialists also register missing teeth of patients, and the silhouettes of existing ones are helpful for forensic identification. Similar processes occur in computer-aided diagnostic tools.

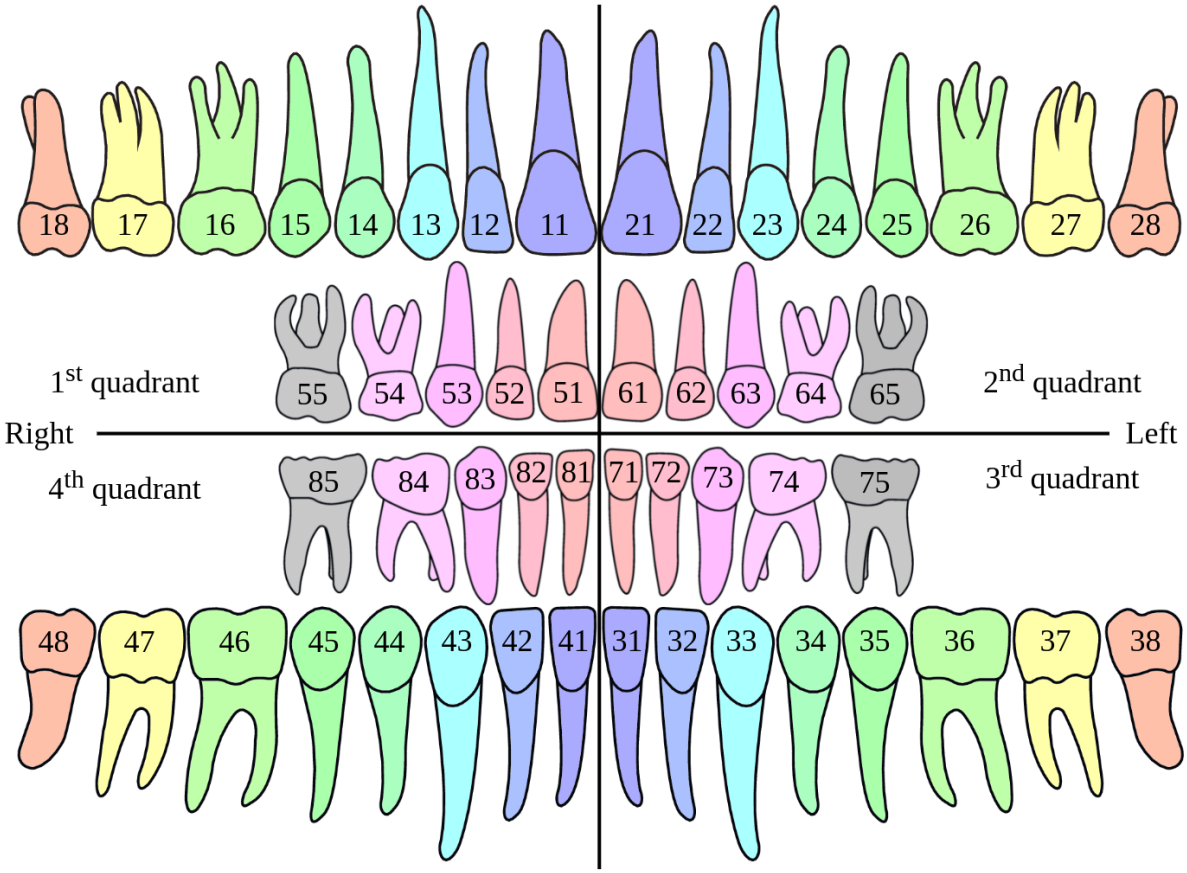
Dental professionals use a numerical notation in their written reports, as well as in their daily routines, to avoid citing the full name of the tooth and to expedite communication. The most common tooth numbering system is the FDI World Dental Federation notation, which represents each tooth by a two-digit number. The first digit specifies the quadrant and the dentition type (permanent or deciduous), while the second digit specifies the tooth type. In this work, we employ the FDI notation together with an additional custom color code system used to illustrate the qualitative results. We illustrate both systems in Figure 2.1, and we refer as "numbering" the act of identifying each tooth using the FDI notation[1].

2.1 TOOTH SEGMENTATION, DETECTION AND NUMBERING TIMELINE

Silva, Oliveira and Pithon (2018) were pioneers in applying deep learning to segment teeth on panoramic radiographs. They used a Mask R-CNN (HE et al., 2017) trained on binary masks that separated teeth from the background and showed that their approach outperformed traditional solutions to the task. The authors also made their data public under the name UFBA-UESB Dental Image dataset[2], which proved to be a valuable

[1]For simplicity's sake, we disregarded the supernumerary teeth in our analyses.

[2]The instructions on how to request the outcome data sets of our research were at:
https://github.com/IvisionLab/dental-image (UFBA-UESB Dental Images)
https://github.com/IvisionLab/deep-dental-image (UFBA-UESB Dental Images Deep)
https://github.com/IvisionLab/dns-panoramic-images (DNS Panoramic Images)
https://github.com/IvisionLab/dns-panoramic-images-v2 (DNS Panoramic Images v2)



Tooth code (2nd digit)

Permanent	1	2	3	4	5	6	7	8
Deciduous	1	2	3			4	5	
Tooth type	Central incisors	Lateral incisors	Canines	First premolars	Second premolars	First molars	Second molars	Third molars

Figure 2.1: The illustration of FDI World Dental Federation notation. The system designates each tooth by a two-digit number, in which the first digit determines the quadrant and the dentition type (permanent or deciduous), and the second digit determines the tooth type. We added a custom color code to identify each tooth in our qualitative results. Source: Adapted from Pinheiro et al. (2021).

resource to the community as it has been extensively used by many works (KOCH et al., 2019; ZHAO et al., 2020; OLIVEIRA; FERREIRA; SANTOS, 2020; CHEN et al., 2021; CUI et al., 2021; HSU; WANG, 2021).

Zhao et al. (2020) proposed a two-stage attention-based neural network for segmentation, referred to as TSAS-Net. The goal was to tackle the challenge of accurately segmenting cluttered borders, such as those found in teeth on dental panoramic radiographs. TSAS-Net consists of two stages. The first stage includes global and local attention modules that produce an initial segmentation map. The second stage is a segmentation network that refines the previous segmentation. The authors achieved state-of-the-art results with this configuration on the UFBA-UESB Dental Image Dataset.

Oliveira, Ferreira, and Santos (2020) proposed a new method based on Generative Adversarial Networks (GANs), called Conditional Domain Adaptation Generative Adversarial Network (CoDAGAN), designed to leverage both labeled and unlabeled data. The method aims to address the issue of disparate patterns caused by digitization techniques in biomedical images, a challenge that hinders the performance of data-driven techniques, such as machine learning models. To validate their approach, the authors tested CoDAGAN on multiple datasets, including the UFBA-UESB Dental Image Dataset.

In a similar vein to Zhao et al. (2020), Chen et al. (2021) explore the challenge of fuzzy root boundaries in panoramic radiographs, particularly in cases involving braces or root resorption. To address this issue, they proposed a multi-scale location perception solution. The key contributions of their work include: (i) a structural multi-scale similarity loss, (ii) a module that identifies tooth pixels from a global perspective, and (iii) a new module that aggregates multi-scale feature branches to reduce the semantic gap. Their approach achieved state-of-the-art results on the dataset provided by Silva, Oliveira and Pithon (2018).

Cui et al. (2021) also employ the UFBA-UESB Dental Image Dataset to conduct experiments using their proposed method, ToothPix, which leverages Generative Adversarial Networks (GANs) for tooth segmentation in panoramic radiographs. Their motivation aligns with previously mentioned works: improving the segmentation of tooth boundaries, particularly in cluttered regions of panoramic images. A key contribution of their approach is the use of wide residual blocks within an encoder-decoder setup, operating on image patches. This configuration, combined with data augmentation techniques, enabled the model to achieve state-of-the-art performance.

Hsu and Wang (2021) investigated the detection and segmentation of previous dental treatments in panoramic radiographs. To address this challenge, the authors proposed a system called DeepOPG, composed of three modules: a functional segmentation module, a tooth localization module, and a dental coherence module. By utilizing these segmentation and localization techniques, the system can perform instance-level segmentation of the objects of interest. The dental coherence module boost the system performance significantly. The system experiments were conducted on the UFBA-UESB Dental Image Dataset.

In an extension of the work of Silva, Oliveira and Pithon (2018) segmented tooth instances on radiographs of the UFBA-UESB Dental Image Data Set also using Mask R-CNN, though not numbering them. In order to perform instance segmentation, the

authors manually modified 276 binary masks from the original data set that separated the teeth from the background. This modification produced labels that disregarded tooth overlapping, but their results surpassed the preceding ones pronouncedly. The authors made their data publicly available under the name UFBA-UESB Dental Image **Deep** Data Set[2].

Silva et al. (2020) advanced the field by segmenting and numbering tooth instances. They conducted a benchmark with 543 radiographs from the UFBA-UESB Dental Image Data Set, modifying the original masks the same way Jader et al. (2018) did, also incorporating numbering labels to the permanent teeth. The benchmark assessed the performance of four end-to-end instance segmentation neural network architectures that achieved state-of-the-art performance on the COCO data set: Mask R-CNN, PANet (LIU et al., 2018), Hybrid Task Cascade (HTC) (CHEN et al., 2019a), and Cascade Mask R-CNN backboned by a ResNeSt (ZHANG et al., 2020). The benchmark winner architecture was the PANet, but the authors concluded that all architectures had satisfactory performances on the task. Their data are publicly available under the name DNS Panoramic Images[2].

Lastly, Pinheiro et al. (2021) labeled from scratch a subset of 450 radiographs from the UFBA-UESB Dental Image Data Set (SILVA et al., 2020) considering tooth overlapping and deciduous teeth, topics neglected by previous studies. The authors refined the Mask R-CNN prediction through the aid of the PointRend module (KIRILLOV et al., 2020). They demonstrated that it is feasible to accurately number and segment permanent and deciduous teeth through end-to-end deep learning solutions and that the PointRend module was more beneficial for segmenting more complex-shaped teeth. They named their data set DNS Panoramic Image **v2** and made it publicly available[2].

Other works from the community include the one of Tuzoff et al. (2019), which proposed a two-stage solution for detecting and numbering teeth. In the first stage, a Faster R-CNN network (REN et al., 2015) detects the teeth without numbering them. The detections define the areas used to generate the crops for the next stage. These crops are bigger than the tooth bounding boxes, which adds location context, easing the classification task. In the second stage, a VGG-16 classification network (SIMONYAN; ZISSERMAN, 2014) takes these crops as inputs and classifies the teeth. In total, the experiments relied on 1572 not publicly available images, labeled with bounding boxes by specialists.

Leite et al. (2021) proposed a two-stage solution to perform segmentation and numbering. In the first stage, a DeepLabv3 network (CHEN et al., 2017), backboned by a ResNet-101 (HE et al., 2016), segments 16 tooth classes (two incisors, one canine, two premolars, three molars for each dental arch). In the second stage, a fully convolutional network (FCN) (LONG; SHELHAMER; DARRELL, 2015) refines the segmentation predictions. For their experiments, the authors employed 153 panoramic radiographs labeled by an expert from a private data set. The two prior solutions had the inherent drawback of not allowing end-to-end training.

Koch et al. (2019) trained a U-Net (RONNEBERGER; FISCHER; BROX, 2015) on the UFBA-UESB Dental Images data set, where horizontal flipping and model ensemble improved performance. Both solutions surpassed the results of classic methods. However, semantic segmentation does not provide the necessary details for further processing steps

in most of the automatic dental analysis.

Chung et al. (2021) developed a new method for detecting and classifying teeth on panoramic radiographs. Firstly, through linear regression, the method localizes 32 points, each representing a single permanent tooth in an adult mouth regardless of its presence, automatically numbering them. In the second and final stage, the point coordinates are refined, and the tooth bounding boxes are predicted in a cascade manner. This approach ignores deciduous and supernumerary teeth.

Krois, Schneider and Schwendicke (2021) examined the impact of image context on tooth classification. The authors showed that a model performance can significantly increase with additional context around the tooth bounding boxes. They confirmed this fact by training and evaluating ResNet-34 networks to classify teeth with different contexts on a private data set comprising 5004 dental panoramic radiographs in total. More than 50 annotators were involved in labeling this large amount of data.

Finally, Panetta et al. (2021), constructed and published a multimodal dental panoramic radiograph data set. The data set comprises 1,000 radiographs and labels for tooth instance segmentation, abnormalities, eye-tracking, and textual description. The authors established some baselines only for semantic segmentation.

2.2 CLASSIFICATION OF DENTAL CONDITIONS

Ekert et al. (2019) employed a custom seven-layer neural network architecture to identify the presence of apical lesions across two levels in teeth, using panoramic radiographs. The data labels were determined by a majority vote among six experienced dentists, each of whom independently annotated the images. It is noteworthy that the teeth from the panoramic radiographs were not automatically detected, but were rather manually cropped prior to the application of the method. The area under the curve (AUC) result of 0.85 led the authors to conclude that the solution demonstrated adequate performance, even on a low-data regime (85 radiographs).

Fukuda et al. (2020) investigated the detection of vertical root fractures in teeth using a convolutional neural network (CNN), specifically the DetectNet architecture using the framework in DIGITS (TAO; BARKER; SARATHY, 2016). The experiments were conducted on data set comprising 330 panoramic radiographs that contained clearly visible fracture lines. Following a supervised learning approach, the data were annotated by two radiologists and one endodontist and radiographs with minor fracture lines were excluded. In a cross-validation setup, the model reached 0.83 of F1-score, which the authors deemed promising.

Lee, Kim and Jeong (2020) evaluated the effectiveness of CNNs, specifically the GoogLe- Net Inception-v3 architecture, in distinguishing between three types of odontogenic cystic lesions (OCLs): odontogenic keratocysts, dentigerous cysts, and periapical cysts. The study utilized panoramic radiographs and cone beam computed tomography (CBCT) images, which were first cropped and then resized to a uniform resolution of 299 by 299 pixels, maintaining the original aspect ratio. Images that were blurry, noisy, of low quality, or otherwise unsuitable were excluded, resulting in a total of 2,126 cropped images used in the analysis. The method was found to be effective, particularly when

trained with CBCT images.

The study conducted by Kwon et al. (2020) utilized the YOLOv3 architecture to identify the four types of dental cysts, including the three types previously studied by Lee, Kim and Jeong (2020) and ameloblastoma. The research was conducted using 1,282 panoramic radiographs, which were labeled by two radiologists. Data augmentation techniques were applied to enhance the data set. The attained AUC, sensitivity, specificity, and accuracy values were considered high, even with a limited number of data samples.

Chen et al. (2021) developed an auxiliary diagnosis system for dental periapical radiographs using CNNs. Their research focused on detecting lesions across various disease categories and severity levels (mild, moderate, and severe) for conditions such as periapical periodontitis and periodontitis. The study used 2,900 periapical radiographs in which the exclusion criteria were images with deciduous teeth, incorrect illumination, and severe distortion. The authors explored different system configurations to detect and classify conditions across multiple dimensions: the entire spectrum of diseases and their severity levels, all disease categories collectively, each disease category separately, and each severity level individually. All results were encouraging, especially in the severe level lesions, demonstrating the capability of CNNs in detecting and classifying multiple diseases in periapical radiographs.

Yüksel et al. (2021) explored the classification of five dental conditions using an elaborate pipeline, termed DENTECT. The pipeline comprises three stages, each powered by a distinct model. The first model divides the panoramic radiograph into four quadrants using a segmentation network. Two YOLO-based neural network models subsequently analyze each quadrant of the dental scans to detect and classify conditions. Additionally, a separate model is dedicated to number the teeth within each quadrant, enhancing the specificity of the diagnostic process. The study's data set consisted of 1,005 dental panoramic radiograph, initially annotated by intern dental students and subsequently validated by a specialist. The authors concluded that the tool is adoptable and capable of reaching the performance level of dental clinicians.

Khan et al. (2021) investigated the use of segmentation architectures for delineating three types of dental conditions in periapical radiographs. Three specialists annotated 206 periapical radiographs to detect features of caries, alveolar bone recession, and interradicular radiolucencies. The experiments demonstrated that U-Net-based neural networks surpassed the performance of other models under review. The authors conceded that the results achieved were not outstanding but were acceptable and promising, highlighting the necessity for further research and a more diverse data set. They concluded that less "off-the-shelf" and more "purpose-built" solutions might lead to a performance boost.

Vinayahalingam et al. (2021) employed a Mask R-CNN neural network for segmentation in panoramic radiographs, identifying not only the teeth but also five types of dental conditions: crowns, fillings, root canal fillings, implants, and root remnants. In accordance with a supervised approach, three clinicians labeled 2,000 radiographs. The solution attained F1-scores exceeding 0.95 for detection, segmentation and classification, which led the authors to conclude that deep learning-based methods may assist clinicians in diagnosing and planning treatments.

Liu et al. (2023) utilized four distinct neural network architectures – ResNet-50, VGG-16, InceptionV3, DenseNet-121 – to classify various dental conditions using periapical radiographs. The classifications included three common lesions: periapical periodontitis, dental caries, and periapical cysts, in addition to the normal condition. Their experiments involved 188 digital periapical radiographs, manually annotated. The images selected underwent a filtering process to exclude those of low quality or those depicting other lesions. In their experimental setup, the most performant neural network was DenseNet-121, reaching 99.5% accuracy. The authors deduced from their findings that employing CNNs for the analysis of periapical radiographs offers a reliable and effective method for assisting in the diagnosis of dental conditions

Bonfanti-Gris et al. (2022) evaluated a web-based software of Denti.AI™, designed to detect and classify five dental conditions using deep learning: metal restorations, resin-based restorations, endodontic treatment, crowns and implants. The research utilized a supervised learning approach and was conducted on 300 panoramic radiographs. The findings demonstrated effective performance in identifying implants, crowns, metal fillings, and endodontic treatments. However, it showed limitations in accurately classifying dental structures and resin-based restorations.

Amasya et al. (2024) proposed DiagnoCat, a deep learning software designed to identify periodontal bone loss in panoramic radiographs. To achieve this, the authors employed two neural networks: one for segmenting and numbering the teeth, a Mask R-CNN, and the other for directly detecting periodontal bone loss, a Cascade R-CNN (CAI; VASCONCELOS, 2018). The experiments were conducted on 6,000 selected images that had minimal image artifacts, at least 10 teeth, and no significant developmental anomalies. In this supervised learning approach, the data was labeled by drawing bounding boxes around relevant features. The authors compared the results attained by the proposed framework with the assessments of three clinicians and found that the framework was successful in accurately identifying periodontal bone loss.

Ranjbar e Zamanifar (2023) took a different approach from most dental research by focusing on predicting eight future treatments instead of diagnosing existing conditions (filling, endodontic treatment, crown, extraction, bridge, implant, reendo and surgical extraction). A dentist with over 25 years of experience labeled the eight types of treatment in 1,025 panoramic radiographs for the experiments. The authors employed an off-the-shelf solution, a YOLOv7 neural network, and reached solid results. The employed model achieved high accuracy in its predictions showing promising potential for application in a clinical setting.

In the context of dental radiographic analysis, the study of Gao et al. (2024) presents an approach on periapical radiographs. The research is grounded on a data set comprising 413 radiographic images. Central to their study is the proposition of a YOLO-based network, the YOLO-DENTAL, outlined to detect and classify four distinct dental conditions: dental caries, dental defects, periapical lesions, and coronal restorations. The authors decided to exclude radiographs depicting distorted teeth or dental crowding. The YOLO-DENTAL achieved an mAP of 86.81%, compared to the 79.95% of YOLOv7-X, leading to the conclusion that the work's methodology can aid in clinical diagnosis.

Tassoker, Öziç and Yuce (2024) explored the use of a neural network (YOLOv5)

for detecting idiopathic osteosclerosis, which is characterized by increased bone density within the jaw, using panoramic radiographs. The research was conducted with a data set of 175 images. The approach was based on supervised learning, where two radiologists provided annotations for the radiographs. Despite the challenges posed by the limited size of the data set and the variability in the radiographs' contrasts and features, the authors reported that their model achieved a high level of accuracy in detection.

2.3 RELATION WITH OUR WORK

Our work is closely related to the studies reviewed in the previous sections. Section 2.1 focused on tooth numbering, segmentation, and detection—tasks that are crucial for developing diagnostic tools for dental conditions and have been a primary research focus of our group. A significant challenge for this field research is the availability and quality of data sets, which remains a major barrier for the research community. The labeling procedure is almost always completely manual and many researchers collect and label amounts of data only for their studies, with custom labeling standards. Consequently, the researchers' precious time is wasted at each new study. In addition, various metrics are used, hindering any possibility of comparing the performance of the proposed solutions. Our work tackled those problems by (i) introducing a large-scale, fine-labeled, and high-variability data set for tooth segmentation and numbering, comprising 4,000 dental panoramic radiographs built upon the HITL concept and (ii) releasing an online platform for benchmarking solutions to work as task central for instance segmentation, semantic segmentation, and numbering.

Section 2.2 reviewed research on the detection and classification of dental conditions, which is the primary aim of AI systems designed to assist dental professionals. Although the studies showed promising results, they primarily relied on off-the-shelf solutions and limited datasets. This highlights that many opportunities for advancing dental condition detection remain unexplored. To address this gap, we proposed a new framework that integrates multiple deep learning paradigms to effectively utilize both labeled and unlabeled panoramic radiographs

Table 2.1 presents the key features of the reviewed work on tooth segmentation and numbering compared to our data set. Our current study uses the largest data set size, with one exception (KROIS; SCHNEIDER; SCHWENDICKE, 2021). However, the latter only labels teeth for detection and numbering, not for segmentation, and is a private data set. Besides works from our research group, there are only two other studies that allow detection, numbering and segmentation (SILVA et al., 2020; LEITE et al., 2021; PINHEIRO et al., 2021; PANETTA et al., 2021). Finally, our constructed data set is public for the research community, different from many works (TUZOFF et al., 2019; CHUNG et al., 2020; LEITE et al., 2021; KROIS; SCHNEIDER; SCHWENDICKE, 2021).

Table 2.2 compares the current study created data set with other research in dental radiographic automation, highlighting common limitations. Most studies, except for one using a data set of 6,000 images (AMASYA et al., 2024), relied on data sets with 2,900 or fewer samples (CHEN et al., 2021; LEE; KIM; JEONG, 2020; VINAYAHALINGAM

Table 2.1: Main features of the previous works' data sets and ours. To the best of our knowledge, this work's data set is the largest on tooth instance segmentation and numbering of dental panoramic radiographs.

Authors	# Radiographs	Detection	Numbering	Segmentation	Image dimensions	Availability	Annotators
Silva, Oliveira and Pithon (2018)	1,500			✓	$1,991 \times 1,127$	Public	Lay people
Jader et al. (2018)	276	✓		✓	$1,991 \times 1,127$	Public	Lay people
Tuzoff et al. (2019)	1,572	✓	✓		N.A.	Private	Experts
Silva et al. (2020)	543	✓	✓	✓	$1,991 \times 1,127$	Public	Students
Chung et al. (2021)	818	✓	✓		Several	Private	Experts
Leite et al. (2021)	153	✓	✓	✓	$2,880 \times 1,504$	Private	Expert
Pinheiro et al. (2021)	450	✓	✓	✓	$1,876 \times 1,036$	Public	Mixed
Krois, Schneider and Schwendicke (2021)	5,008	✓	✓		N.A.	Private	Mixed
Panetta et al. (2021)	1,000	✓	✓	✓	$1,615 \times 840$	Public	Mixed
We	4,000	✓	✓	✓	$2,440 \times 1,292$	Public	Mixed

et al., 2021). Such small data sets can compromise model generalizability. Additionally, many studies excluded challenging cases, limiting the practical applicability of their findings (FUKUDA et al., 2020; LEE; KIM; JEONG, 2020; CHEN et al., 2021; LIU et al., 2023; AMASYA et al., 2024; GAO et al., 2024). The focus on a narrow range of target classes (EKERT et al., 2019; FUKUDA et al., 2020; LEE; KIM; JEONG, 2020; KHAN et al., 2021; LIU et al., 2023; AMASYA et al., 2024) further restricts the comprehensiveness of these models.

Technically, most studies relied on supervised learning (RANJBAR; ZAMANIFAR, 2023; TASSOKER; ÖZIÇ; YUCE, 2024; AMASYA et al., 2024; BONFANTI-GRIS et al., 2022), which requires extensive labeled data, which is a significant limitation. There was also a trend towards using off-the-shelf solutions (LEE; KIM; JEONG, 2020; KWON et al., 2020; VINAYAHALINGAM et al., 2021; LIU et al., 2023; RANJBAR; ZAMANIFAR, 2023).

This study introduces a novel framework for diagnosing various dental conditions from panoramic radiographs, leveraging the largest data set in the field (16,824 images). A semi-supervised approach, combining human-in-the-loop (HITL) strategy (SILVA et al., 2023) and Masked Autoencoders (MAE) (HE et al., 2022), is used to enhance performance and reliability. The proposed framework involves creating tooth crops from annotated and predicted teeth on radiographs, with an auto-labeler using LLMs to extract dental conditions from textual reports and map them to corresponding teeth using the FDI numbering system. This approach covers 13 dental conditions and includes a statistical agreement analysis to validate the results.

2.4 CLOSURE

This chapter reviews studies that have applied deep learning to panoramic radiographs. These works can be divided into two categories: (i) studies focused on tooth segmentation, detection, and numbering, and (ii) those investigating the detection and classification of dental conditions. Although the latter studies are not directly related to dental conditions, they can be particularly valuable to the former, as teeth are common targets for radiologists, and the detection, segmentation, and preprocessing steps are essential for many diagnostic tools. Therefore, it becomes imperative not only to research

dental classification but also to investigate tooth detection and segmentation, as will be addressed in the following chapters.

Table 2.2: Comparison of the current study with others regarding data set, task, learning paradigm, proposed solution, and investigated classes.

Reference	Radiograph	# Radiographs	Detection	Supervision	Semi-, Self-supervision	Architecture or framework	# Classes	Classes
Ekert et al. (2019)	Panoramic	85		✓		Custom	2	Two levels of apical lesions
Fukuda et al. (2020)	Panoramic	300	✓	✓		DetectNet version V5	1	Vertical root fracture
Lee, Kim and Jeong (2024)	CBCT and panoramic	2,126		✓		GoogLeNet Inception-v3	3	Odontogenic keratocysts Dentigerous cysts Periapical cysts
Kwon et al. (2020)	Panoramic	1,282	✓	✓		YOLOv3	4	Dentigerous cysts Periapical cysts Odontogenic keratocysts Ameloblastomas
Chen et al. (2021)	Periapical	2,900	✓	✓		Faster R-CNN	9	Severities for decay (3) Periapical periodontitis (3) Periodontitis (3)
Yüksel et al. (2021)	Panoramic	1,005	✓	✓		Custom framework DENTECT	5	Periapical lesion therapy Fillings Root canal treatment Surgical extraction Conventional extraction
Khan et al. (2021)	Periapical	206	✓	✓		U-Net	3	Caries Alveolar bone recession Interradicular radiolucencies
Vinayahalingam et al. (2021)	Panoramic	2,000	✓	✓		Mask R-CNN backboned by Resnet-50	6	Crowns Fillings Root canal fillings Implants Root remnants
Liu et al. (2023)	Periapical	188		✓		ResNet-50; Vgg-16; InceptionV3; DenseNet-121	3	Caries Periapical peridontites Periapical cyst
Bonfanti-Gris et al. (2022)	Panoramic	300	✓	✓		Faster R-CNN	5	Metal restorations Resin-based restorations Endodontic treatment Crowns Implants
Amasya et al. (2024)	Panoramic	6000	✓	✓		Mask R-CNN and Faster R-CNN	1	Periodontal bone loss detection
Ranjbar and Zamanifar (2023)	Panoramic	1,025	✓	✓		YOLOv7	8	Restoration Root canal treatment and filling Crown Conventional extraction Bridge Implant Root canal treatment and filling Surgical extraction
Gao et al. (2024)	Periapical	1,567	✓	✓		YOLOv7	4	Dental caries Dental defects Periapical lesions Coronal restorations
Tassoker, Öziç and Yuce (2024)	Panoramic	175	✓	✓		YOLOv5	1	Idiopathic osteosclerosis
Ours	**Panoramic**	**16,824**	✓	✓	✓	**Custom**	**13**	**Several**

Chapter 3

MATERIALS AND METHODS

In deep learning-based systems, the number of adjustable parameters of a neural network easily surpasses the million mark, demanding large amounts of data for training. Most domains, including computer vision, fundamentally rely on supervised learning techniques, which require labeled data to fit the deep learning model weights (LECUN; BENGIO; HINTON, 2015). The labeling procedure depends on human specialists who manually annotate the data according to the application purposes. This step is crucial and can take up more than 80% of a machine learning project's time (WU et al., 2021). Consequently, labeled publicly available data sets are valuable resources, and for academic research, they offer the additional benefit of creating benchmarks for model performance comparisons (MENZE; GEIGER, 2015; CORDTS et al., 2016; WANG et al., 2018). This scenario was no exception when it came to research on panoramic radiographs. To fill this gap, we constructed from scratch a tooth instance segmentation data set of dental panoramic radiographs that was publicly available to any researcher in the world upon request. This dataset, combined with the HITL approach, supported our studies by facilitating tooth instance segmentation in panoramic radiographs and enabling the classification of various dental conditions through the use of available textual reports.

3.1 METHODOLOGY

The experimental nature of our study led us to adopt a holistic research structure, guiding us from the inception of the deep learning project to its completion. We progressed through each step, including image data collection and labeling, design structuring, quantitative evaluation, and human assessment. One of the key steps in our work was the creation of a new tooth instance segmentation dataset from dental panoramic radiographs, which was essential for our experiments. This dataset enables direct investigation into tooth detection and segmentation in panoramic radiographs. However, constructing datasets is often costly and time-consuming, which can limit the size of collections and the robustness of experiments. To mitigate the time required for manual image annotation, we employed a Human-In-The-Loop (HITL) approach to accelerate the labeling process.

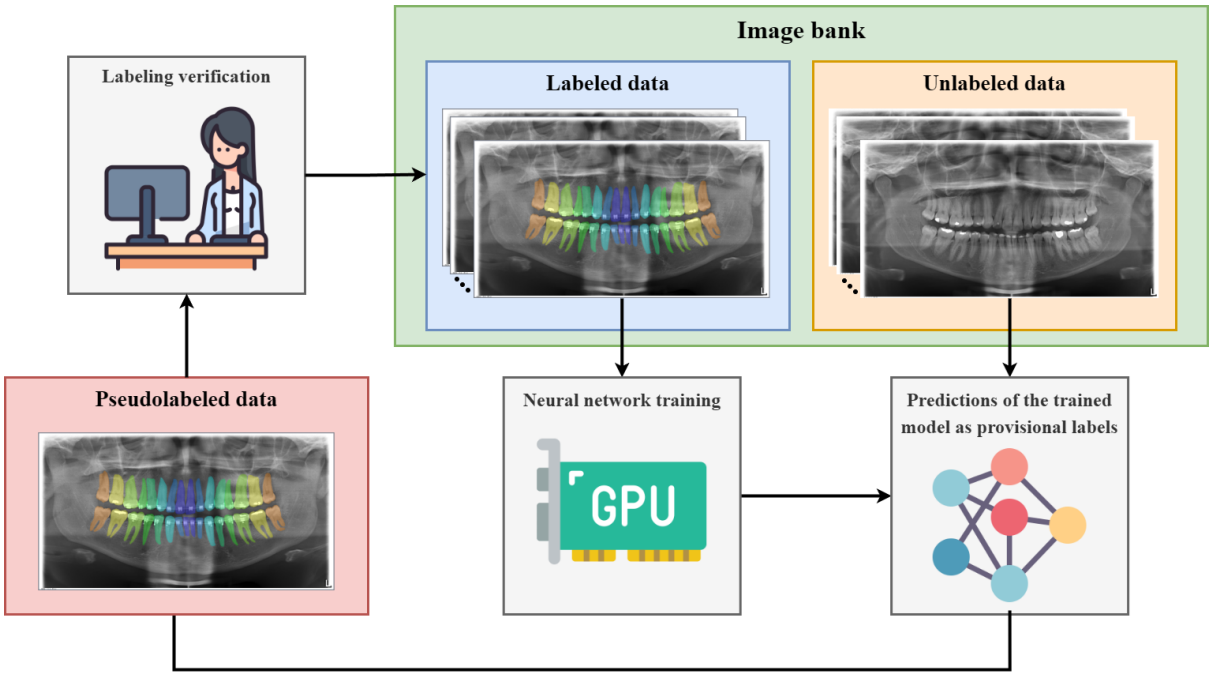


Figure 3.1: A general HITL diagram for the case of tooth instance segmentation on panoramic radiographs. The process works as follows: A neural network is fitted with available labeled data. This neural network produces annotations for the unlabeled data, which are later verified by humans. The human supervision qualifies the newly labeled data for the model training in subsequent HITL iterations.

The HITL approach aims to efficiently label the data by combining machine learning models and human supervision, expecting an overall reduction of time and costs (WU et al., 2021). Figure 3.1 displays a generic HITL pipeline with interventional training, using dental panoramic radiographs as example. In that setup, an initial set of labeled data are used to fit a machine learning model, which later produces annotations for new unlabeled data. Then, human experts verify (confirm or correct) these annotations, which qualifies them as suitable training and validation data for the next HITL iteration. After each training iteration, the model performance is expected to improve, increasing the label quality and lessening the verification time.

To effectively implement the HITL approach, it is important to choose a model that not only performs the task but also operates at optimal performance levels. A significant portion of our work focused on evaluating whether a two-stage detector could meet our objectives, identifying the optimal configuration, and selecting the best available neural networks (SILVA; OLIVEIRA; PITHON, 2018; PINHEIRO et al., 2021; SILVA et al., 2023). We found that several architectures (Mask-RCNN, PANET, HTC, ResNeSt) demonstrated reasonably good performance (SILVA et al., 2020), but the choice of architecture head and the backbone architecture had a significant impact on the results (PINHEIRO et al., 2021; SILVA et al., 2023). Rather than exhaustively testing every combination of backbone and head, we used the MMDetection (CHEN et al., 2019b),

library for object detection that is an open-source, to benchmark top-performing networks on the COCO dataset. These models were evaluated based on their mAP metric, allowing us to streamline the selection process.

Following the application of the HITL procedure, we thoroughly analyzed the results to assess its effectiveness. Part of this verification involved the traditional comparison of predictions from trained networks on panoramic radiographs with the ground truth, where improvement was expected with each iteration. The other part involved comparing the annotation time of human labelers with and without the use of the HITL concept. It would only make sense to use HITL if these variables showed improvement after each iteration. The improvement was evident, enabling us to move forward with constructing our tooth instance segmentation dataset and classifying dental conditions. With the HITL performance exhaustively validated, we proceeded with the setup for dental condition classification.

As in any supervised learning paradigm, we needed labeled data to train our network. A theoretical solution was to manually label each tooth, which was impracital, as our data contained more than 400,000 tooth instances. We approached the problem in another way: we labeled the data through the available textual reports, allowing we to classify several dental conditions dental conditons. We also used the radiographs with unavailable report as a pretraining strategy using tooth crops as data.

3.2 LIMITATIONS

We can point out some limitations of our work, such as the challenge of evenly sampling the dataset from the broader population. To address this, we utilized data collected directly from a dental clinic, allowing us to replicate patient distribution patterns and closely mirror the local population. However, focusing on a local population excludes ethnicities from regions farther away, possibly limiting the generalizability of the findings. Therefore, we must acknowledge that our results are biased toward the ethnicities represented in the patient population from which the data were collected.

In sum, the methodology of the study was designed to effectively achieve the research objectives. First, data were collected directly from dental clinics, ensuring a representative sample of the target population and mitigating potential biases. Next, deep learning techniques were employed to detect, segment, and classify dental conditions in panoramic radiographs, using a HITL approach to expedite data annotation. The combination of labeled and unlabeled data together with the textual reports enabled a more comprehensive utilization of available information for dental condition classification. This methodological approach ensures analytical accuracy and provides a solid foundation for the replicability and generalization of the results.

3.3 STARTING DATABASES

Machine learning and deep learning are heavily based on data, making the choice of data sample critical. In addition, the sample should accurately represent the target population. In this study, our goal was to apply the developed method to the general

Table 3.1: Specifications of the device used for X-ray image acquisition.

Property	Description
Model	ORTHOPHOS XG 5/XG 5 DS/Ceph
Nominal frequency	50 Hz/60 Hz
Pipe output power	1080W with any radiation duration
Tube voltage	60–90 kV (at 90 kV max. 12 mA)
Scale of images	For P1, normal dental arch approx. 1:1.19, i.e., the acquired image is enlarged by approx. 19%, on average, compared to reality.
X-ray tube	Siemens SR 90/15 FN or CEI OCX 100
Panoramic sensor	Digital CCD line sensor, repluggable for panoramic exposure technique
Active sensor area, panoramic type	138 mm × 6.48 mm
Widescreen sensor resolution	0.027 mm in size of the pixels
Size of images	2440 × 1292 pixels
Focus-sensor distance	497 mm

population, particularly to individuals who visit dental clinics. Thus, the use of data from these establishments was deemed the most appropriate approach.

Our data were obtained from two collections of dental panoramic radiographs sourced from a dental clinic. The first collection comprised 8,795 raw panoramic radiograph images, accompanied by the names of the patients, taken between January 2012 and April 2013. The second collection included 8,029 images, each paired with a textual report in Portuguese, containing the patient's name, age, and date of the radiograph, captured between January 2015 and December 2016, totaling 16,824 images. The textual reports were written in the `.odt` format. Both databases were generated using the same device: an ORTHOPHOS XG 5/XG 5 DS/Ceph. Table 3.1 outlines the properties of the device used.

3.3.1 Exploratory analyses

After collecting the data, we began our study with a preliminary analysis, which was done in Python. First, we curated the two image banks that were available to us, resulting in the 8795 and 8029 numbers previously mentioned. The images of these banks had widths ranging from 2,272 to 2,692 and heights ranging from 1,292 to 1,304. The most common size was 2,440 × 1,292, accounting for 79% of the images.

The textual reports were preprocessed from the original `.odt` format and converted to the `.txt` format for consistency and easier manipulation. The final `.txt` files had the format exemplified in Table 3.2 [1]: A two-digit number followed by a colon and the report line. In Table 3.2, the teeth are highlighted in bold according to the FDI system described in Fig. 2.1, to emphasize their significance for radiograph screening. In the

[1]For this article, all textual reports were translated from their original language (Portuguese) to English.

Table 3.2: Sample of a panoramic radiograph preprocessed report of the TRPR dataset. It is highlighted in bold the mentioning of the teeth according to the FDI system described in Fig 2.1 to reveal their importance.

Topic Number	Report line
01:	Anatomical modification in the right and left mandible condyle.
02:	Missing teeth: **18**, **28** and **48**.
03:	Teeth **13** and **38** included and impacted.
04:	Tooth **36** and **37**: endodontic treatment. Partially filled root canals.
05:	Mild bone loss in the region of the present teeth.
06:	Modification of the bone trabeculation in the region of tooth **48** compatible with a bone scar.
07:	Calcification of the right and left stylohyoid ligament complex.

textual reports used, as is commonly done, the digits of the textual reports denote a tooth through the FDI notation, while other numbers are written in full.

Although the biological sex of the patients was not available, we were able to infer it based on their names. In Portuguese, names generally provide a clear indication of an individual's biological sex, which allowed us to make this inference. The inference procedure revealed that the male and female sex was not evenly distributed. The female radiographs were the major one and acounted to 61.8% of the images in the first database of images and 60.2% on the second database. Furthermore, the age distribution is shown in Figure 3.2, which shows a wide range of values (1 to 90 years). The mean age is 32 years, with a standard deviation of 17.2 years.

3.3.2 Ethical Considerantions

The sensitive nature of the working data required special caution. We requested approval from the ethics committee for our project, ensuring that all procedures complied with the necessary ethical guidelines and standards. The National Commission for Research Ethics (CONEP) and the Research Ethics Committee (CEP) authorized the use of the radiographs for research under report number 646.050/2014. Strict measures were adopted to ensure data confidentiality: patient names and other identifiable information were anonymized in accordance with ethical guidelines and data protection regulations, meeting all ethical requirements for research involving human subjects.

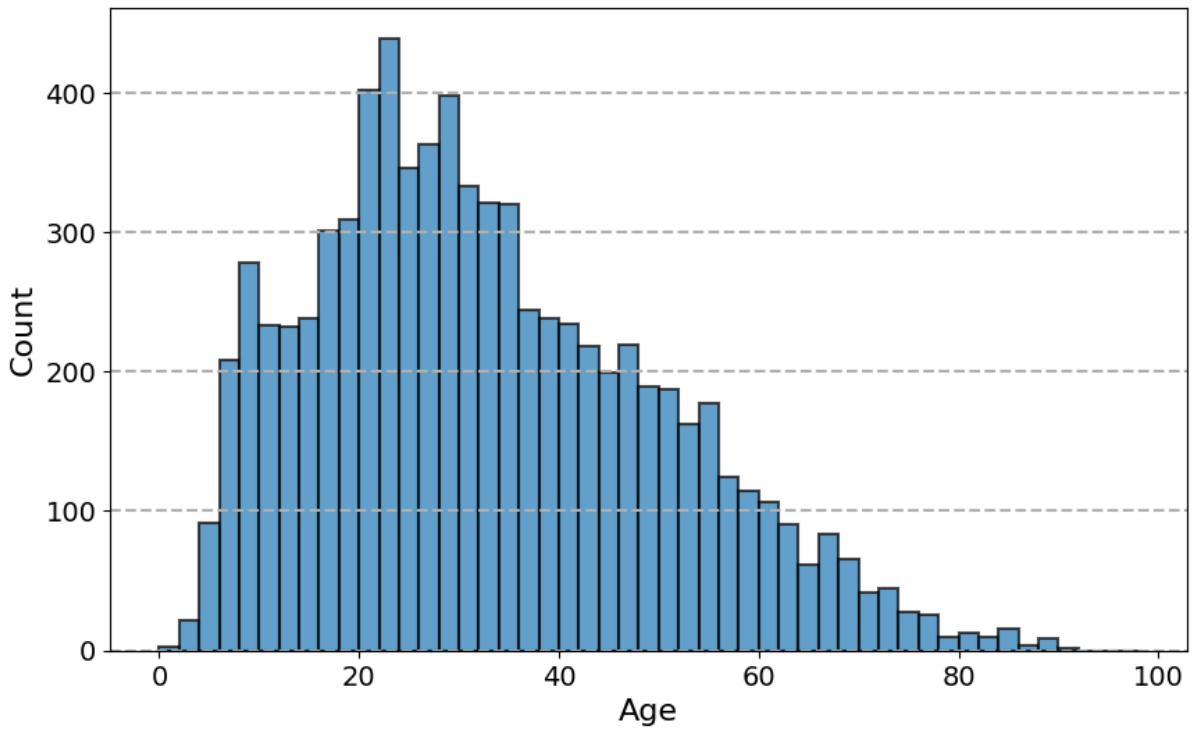


Figure 3.2: Age distribution of the patients of the second database ranged from 1 to 90 years, with a mean age of 32 years and a standard deviation of 17.2 years

3.4 PROPOSED SYSTEM

Our proposed solution consists of a four-step framework designed to automate tooth detection and classification of dental conditions from panoramic radiographs:

1. Construction of the panoramic radiograph datasets;
2. Construction of the crop datasets;
3. Classification network pretraining and label extraction;
4. Classification of Dental conditions.

These steps are depicted in Fig. 3.3. Steps 1 and 2 were designed to quickly generate a collection of pseudolabeled panoramic radiographs for creating tooth crops. From these groups, two distinct categories of datasets were built. The first group comprises a set of panoramic radiographs in total dimensions. The second group, derived from the first, comprises tooth crops for training binary classification models. Experiments of this process will be discussed in detail in Chapter 5. Step 3 focuses on using large language models to efficiently label teeth according to their dental conditions and to leverage data without textual reports for pretraining, while Step 4 constists in training binary classifiers. The experiments related to these phases will be covered in Chapter 6. In the following Sections, we discuss each step of our methodology in detail.

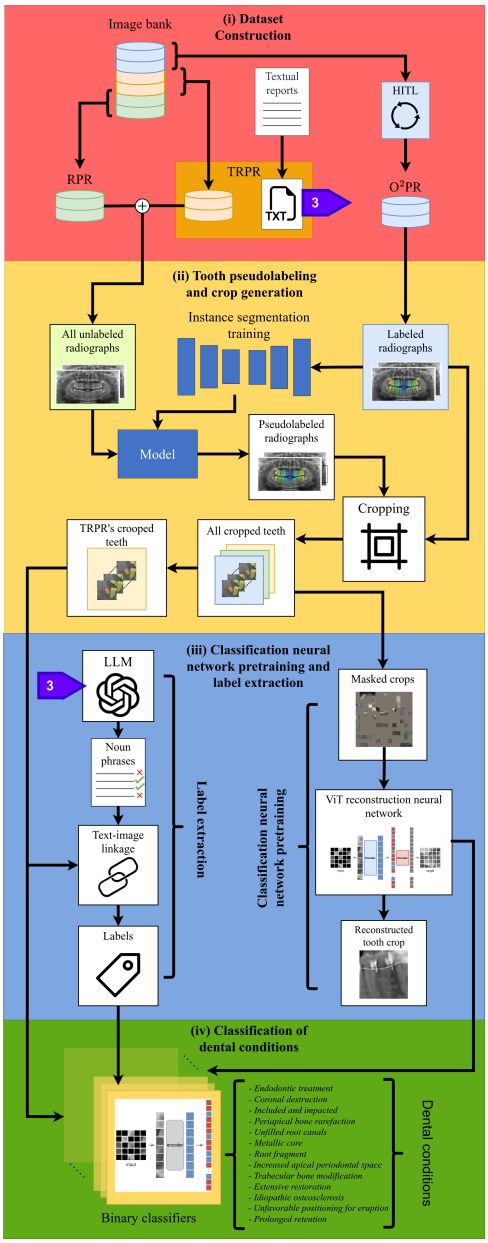


Figure 3.3: Proposed solution for classifying dental conditions: (i) **Construction of the panoramic radiograph datasets**: Combines the full-width radiographs into three non-overlapping groups, (ii) **Construction of the crop datasets**: Generates tooth crops centered on the teeth, (iii) **Classification neural network pretraining and label extraction**: Pretrain networks via Masked Autoencoders (MAE) and creates labels from noun phrases extracted of the reports, and (iv) **Classification of dental conditions**: Trains a binary classifier for each dental condition.

3.4.1 Construction of the Panoramic Radiograph Datasets

The construction of the Panoramic Radiograph datasets is illustrated in Fig. 3.3 (i). It began with all 16,824 image samples in their full dimensions. These images were gath-

Table 3.3: Feature summary of the Panoramic Radiograph datasets, whose images were obtained from the image bank in their full dimensions. These features include image counts, tooth instance segmentation labels, textual report availability, tooth pseudolabeling, image dimensions, and dataset splitting for training an instance segmentation model.

Dataset	# Images	Inst. Segm. Labels	Textual Reports	Pseudolabeling	Mode Dimensions	Inst. Segm. Training Data
RPR	4,795			✓	2,440×1,292	None
O²PR	4,000	✓				All
TRPR	8,029		✓	✓		None

ered into three distinct, non-overlapping datasets. The first subset, consisting of 4,795 radiographic images without textual reports and devoided of tooth segmentation labels, is called the **Raw Panoramic Radiographs** (RPR) dataset. The remaining 4,000 radiographs without textual reports have segmented and numbered teeth, labeled either manually or through the human-in-the-loop procedure. We called this one **OdontoAI Open Panoramic Radiographs (O²PR)**. Finally, we designated the subset of images accompanied by textual reports as the **Textual Report Panoramic Radiographs** (TRPR) dataset.

The O²PR dataset was constructed using the HITL[2]. The data set aimed to fill a gap in the dentistry field, where a large-size and consistently labeled panoramic radiograph data set was lacking, while deep learning applications were still in an incipient stage compared to other healthcare areas (SCHWENDICKE; SAMEK; KROIS, 2020). The objects of interest were the teeth, which were annotated and classified. However, the RPR and TRPR datasets lacked labels, a challenge we addressed by generating pseudolabels, as will be discussed in the following section. The main characteristics of these datasets are presented in Table 3.3.

3.4.2 Tooth pseudolabeling and construction of the crop datasets

This step serves two purposes, as illustrated in Fig. 3.3 (ii): (i) to automatically generate pseudo labels for the teeth in the radiographs of the RPR and TRPR datasets using an instance segmentation network, as they do not contain instance segmentation labels; and (ii) to create tooth crops from the tooth labeled and pseudo labeled images, which are later used to train and test the classification neural networks. These procedures allow for consistent, uniformly sized image crops over all datasets around each tooth. The standardization is crucial for subsequent steps, as the classification neural networks are trained on fixed-size inputs.

We aimed to use neural networks trained on the O²PR, developed using the HITL process, to pseudolabel the RPR and TRPR datasets. However, these networks had not yet undergone comprehensive validation. Verifying their effectiveness was crucial to ensure accurate pseudolabeling of additional radiographs. To address this, we conducted

[2]The O²PR dataset was publicly available until February 2024.

a validation process designed to assess both tooth detection and segmentation accuracy before and after applying HITL. The process is described in Chapter 4 and involved: (i) benchmarking several instance segmentation network architectures and selecting the best-performing model for further experiments; (ii) evaluating the progress of the networks across HITL iterations; and (iii) performing a numbering analysis of mean average precision (mAP) per tooth, as well as a confusion matrix analysis based on the Intersection over Union (IoU). The achieved reliability provided us with the confidence to utilize the neural network predictions as pseudolabels for the RPR and TRPR datasets.

An instance segmentation neural network was trained on all 4,000 labeled images of the O^{2}PR dataset. The hybrid task cascade (HTC) architecture (CHEN et al., 2019a) was selected because it was the best model in the benchmark conducted by Silva et al. (2023). HTC ensures more accurate object boundaries and improved detection results by leveraging information from tasks like semantic segmentation. Using this network, the teeth of all the remaining unlabeled radiographs from the RPR (4,795) and the train set of TRPR (8,029) datasets were segmented. Finally, 4,000 labeled radiographs and 12,824 pseudo labeled radiographs were obtained.

After training, our HTC instance segmentation neural network was employed to generate two distinct datasets of tooth crops from all labeled and pseudolabeled radiographs. The pseudolabels were essential, as our system needed to create crops around the teeth in the panoramic radiographs. This step was crucial for associating the TRPR tooth crops with their respective conditions via the textual reports, while the remaining crops, which lacked textual reports, were utilized as pre-training data. This approach minimized data wastage, as even crops without labeled dental conditions were effectively utilized.

The primary tooth crop variant spanned 224×224 centered around each tooth. This tooth crop type has the advantage of being more focused on the teeth but the disadvantage of having less context of the tooth surroundings, possibly excluding tooth parts. This dataset is termed "less context" crops. To address the context gap, a second crop category, termed "more context" crops, began with a broader 380×380 area, which was then resized to 224×224 to comply with the requirements of the employed neural network architecture used for classification. These data crops served as input for the subsequent step. After this procedure, approximately 460,000 crops were obtained for each configuration. Fig. 3.4 illustrates the cropping procedure, while Table 3.4 compiles the features of these datasets, termed **Crops**. This table also includes the dataset split for pretraining and training of the classification networks.

3.4.3 Classification network pretraining and label extraction

The next step is illustrated in Fig. 3.3 (iii). This stage consists of two distinct processes that can be executed independently: the neural network pretraining and label extraction (via auto-labeler). The neural network pretraining was designed to enhance the performance of classification models for the final tasks, leveraging even the unlabeled data to learn significant features.

A ViT archictecture was selected for the used pretraining strategy because of its superior performance in benchmarks (KHAN et al., 2022), where it achieved state-of-the-

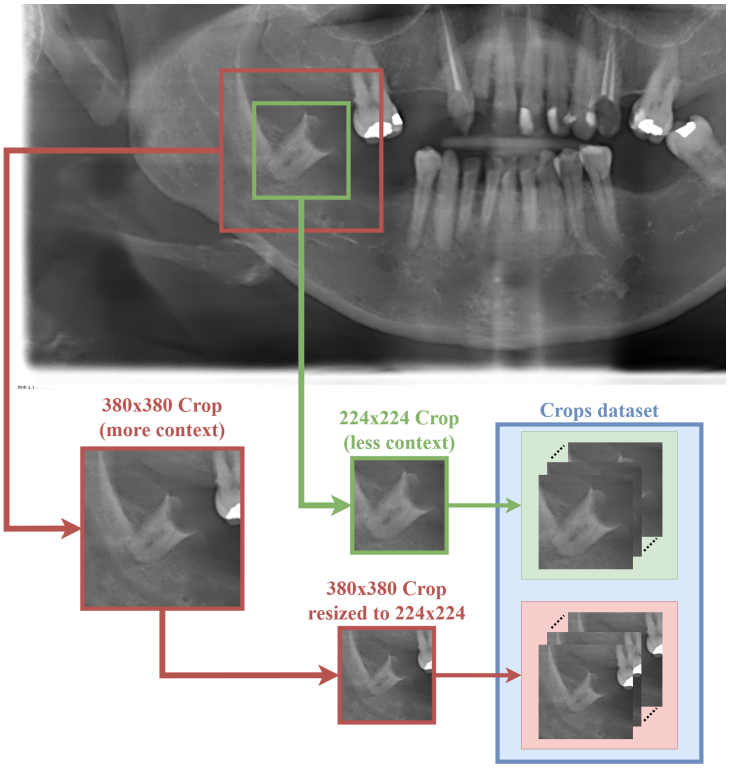


Figure 3.4: Two tooth crop variants used in this study. The first, termed the "less context" crop, was taken from a panoramic radiograph of a tooth and measures 224×224 pixels. The second, termed the "more context" crop, was resized to 224×224 pixels from an original crop of size 380×380 pixels. These two sets comprise the **Crops** dataset.

Table 3.4: Feature summary of the second group of datasets, termed **Crops**, whose images were used to pretrain and train binary classifiers for tooth conditions. These characteristics include the image dimensions, source dataset, crop counts, textual reports availability, pretraining usage, and dataset splitting to train and test the binary classifiers. All the images are tooth crops sourced from the first group of datasets.

Dataset	Crop Dimensions	Source Dataset	# Crops	Textual Reports	Pretraining	Train	Validation	Test
Less context	224×224	RPR	132,497		All	None	None	None
		O2PR	112,842					
More context	380×380 to 224×224	TRPR	213,395	✓	Train only (70%)	70%	15%	15%

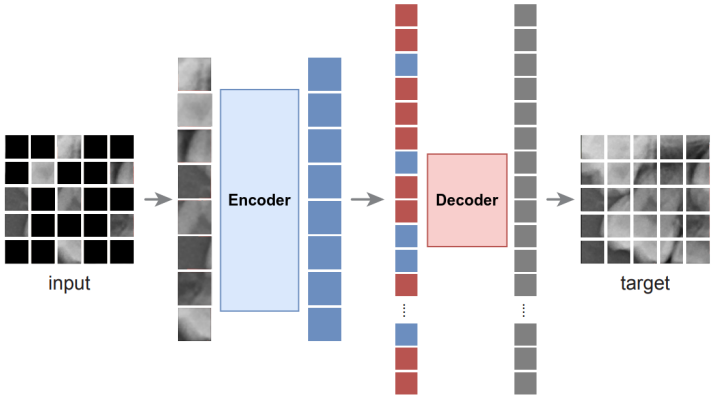


Figure 3.5: Illustration of a MAE: Selected patches from an input image are obscured, and the remaining visible patches are processed through an encoder. The obscured patches are subsequently reconstructed using a decoder from the latent space representations.

art results by exploiting transformers, and due to its convenience for employing the MAE strategy. The current study used MAE, a variant of deep-learning autoencoders trained by deliberately masking out portions of the input data, as a pretraining strategy. This "masking" approach challenges the network to reconstruct the original data, including the intentionally obscured (masked) patches, resulting in a more robust latent representation. Benefiting from transfer learning principles, MAEs can leverage pretrained models to further enhance their performance and generalization capabilities.

Fig. 3.5 illustrates an MAE within the context of the ViT architecture. Given an input image segmented into a grid with some sections obscured, the encoder compresses this partial image into a compact representation. These encoded data are then processed by the decoder, aiming to regenerate the full image. In the selected configuration, the decoder has fewer parameters than the encoder to focus on efficient reconstruction (HE et al., 2022). After completing the pretraining phase, the decoder is discarded, emphasizing its role in learning robust image representations without contributing to the final task performance.

The auto-labeler is the second process, executed in parallel with the pretraining pro-

cedure. The labels here represent dental conditions; therefore, only the crops from the TRPR dataset were used, as this dataset is the only dataset containing textual reports. This process aimed to extract the noun phrases to create the labels later used as ground truth for the classification neural networks. A noun phrase is a word or group of words with a noun as its head or main word. Noun phrases can function in a sentence as a subject, an object, or a complement. They can be single nouns or more complex structures with modifiers and related words. In the proposed framework, extracting noun phrases is necessary because all dental conditions are noun phrases, although not all noun phrases are dental conditions.

A large language model was used to automate and expedite noun phrase extraction. The adopted methodology is centered on prompt engineering – a technique where specific inputs are crafted to elicit desired outputs from language models. Guided by this approach, the following prompt was formulated to be executed on the textual reports:

> "You are an excellent English teacher who indicates for each item (sentences starting with two digits) of a textual report the noun phrases through vertical topics."

With the input above, for instance, the noun phrases extracted from sentence "03" in the report shown in Table 3.2 were:

- Tooth 36
- 37
- Endodontic treatment
- Partially filled root canals

However, in the lists of noun phrases, the teeth are not directly connected to their conditions. A linkage procedure was used to solve this issue. In this procedure, all teeth were associated with all present conditions in the sentence. For instance, using the linkage procedure, tooth 36 and tooth 37 both have the conditions of endodontic treatment and partially filled root canals. This example is illustrated in the following line:

The linkage process proves effective in the context of dental reports, which are organized by specific conditions. Radiologists are more favorable to note the presence of conditions rather than the absence. This tendency is illustrated in the case of the patient mentioned in Table 3.2, sentence "03". For example, it is less common for radiologists to document each tooth lacking conditions, such as 'Teeth 17, 16, 15, 14, 13, ... without endodontic treatment or unfilled root canals', focusing instead on those with notable conditions.

Sentences detailing the presence or absence of teeth were excluded from the analysis, as illustrated in sentence "02" of Table 3.2. This decision was based on the understanding

that this task is better conducted using object detection or instance segmentation techniques specifically designed for identifying and delineating objects within images (SILVA et al., 2020; PINHEIRO et al., 2021; SILVA et al., 2023). To filter out these sentences, straightforward regular expressions were implemented.

3.4.4 Dental conditions classification

The fourth and final step of our framework focused on training binary ViT classifiers, as shown in Fig. 3.3 (iv). Initially, we used a baseline model without pretraining. To enhance the approach, we also explored pretrained weights from MAEs. Specifically, MAE weights pretrained on two datasets—the widely recognized ImageNet and the previously constructed Crops dataset—were utilized. Following this pretraining, we trained a separate classifier for each selected tooth

The method was evaluated using the Matthews Correlation Coefficient (MCC) metric (MATTHEWS, 1975). The MCC takes into account true positives (TP), false positives (FP), and false negatives (FN), including the frequently overlooked true negatives (TN):

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}\,. \tag{3.1}$$

The main properties of MCC include:

- The denominator acts as a normalization factor, bounding MCC values between -1 and 1;
- A score of -1 indicates entirely incorrect predictions, while 1 signifies perfect predictions;
- An MCC of 0 implies predictions equal to random guessing; MCC treats TP, TN, FP, and FN symmetrically, an important feature when these values bear similar implications;
- MCC is preferable to the cases of imbalanced datasets as it does not give highly optimist results in opposition to other metrics, such as accuracy.

3.5 CLOSURE

In this chapter, we provided a comprehensive description of the methodology, which was thoroughly examined. While the procedure is somewhat complex, it has the potential to address and enhance various aspects. The panoramic radiographs were labeled with small effort through the HITL procedure and by using the predictions of an instance segmentation network. We could directly create the **Crop** dataset through the Radiograph Panoramic datasets and pretrain the binary classifiers with the MAE strategy to boost their final performance. With the support of a Large Language Model, we generated the classification labels directly from the radiological reports used to train the final binary classifiers.

Chapter 4

DATASETS THROUGH HUMAN-IN-THE-LOOP AND PSEUDOLABELING

The primary goal of this work was to automatically detect dental conditions from panoramic radiographs. A key step involved creating two datasets: full-size labeled panoramic radiographs and tooth-centered labeled crops. The panoramic images are used for training networks in tooth numbering, while the tooth crops are essential for training networks specialized in condition classification.

This section details the use of Human-In-The-Loop (HITL) and pseudolabeling techniques to efficiently labels of panoramic radiographs, and generate tooth crops. A dental instance segmentation network is employed to isolate individual tooth to create the tooth crops, which will be followed by a Vision Transformers (ViT) network to classify the conditions. By integrating these methods, we aim to improve the accuracy and efficiency of the automated detection and classification process.

4.1 PANORAMIC RADIOGRAPH DATASET CONSTRUCTION

We employed the HITL technique to perform instance-level tooth segmentation in panoramic radiographs and created a dataset called OdontoAI Open Panoramic Radiographs (O^2PR) dataset. The O^2PR dataset contains 850 manually annotated images and was constructed using 1,493 radiographs from the UFBA-UESB Dental Image Dataset (after discarding seven duplicates from the original 1,500) along with 2,507 additional images, totaling 4,000 radiographs. All radiographs were sourced from the previously mentioned image database, acquired using an ORTHOPHOS XG 5/XG 5 DS/Ceph device. Silva, Oliveira and Pithon (2018) grouped the original 1,493 images into ten radiograph categories, according to the presence of dental appliances, restorations, and 32 teeth. Two supplementary categories are exclusive for radiographs with dental implants and mouths with more than 32 teeth. This categorization demonstrated the high variability of the data. Following this categorization, we grouped the remaining radiographs of the image bank and noted that the database category proportions differed from the original 1,493 images subset. For instance, the radiographs of Category 1 were too oversampled

Table 4.1: Summary of the UFBA-UESB Dental Image and O^{2}PR data sets according to the number of images per radiograph category. We conducted the HITL procedure so the O^{2}PR's radiograph category proportions were similar to the image database's.

Category	32 Teeth	Restorations	Dental Applicance	UFBA-UESB Dental Image	**O^{2}PR**
1	✓	✓	✓	73	93
2	✓	✓		219	438
3	✓		✓	45	110
4	✓			138	274
5	Radiographs contaning dental implant(s)			120	228
6	Radiographs contaning more than 32 teeth			169	335
7		✓	✓	114	420
8		✓		455	1804
9			✓	45	93
10				115	205
		Total		1493	4000

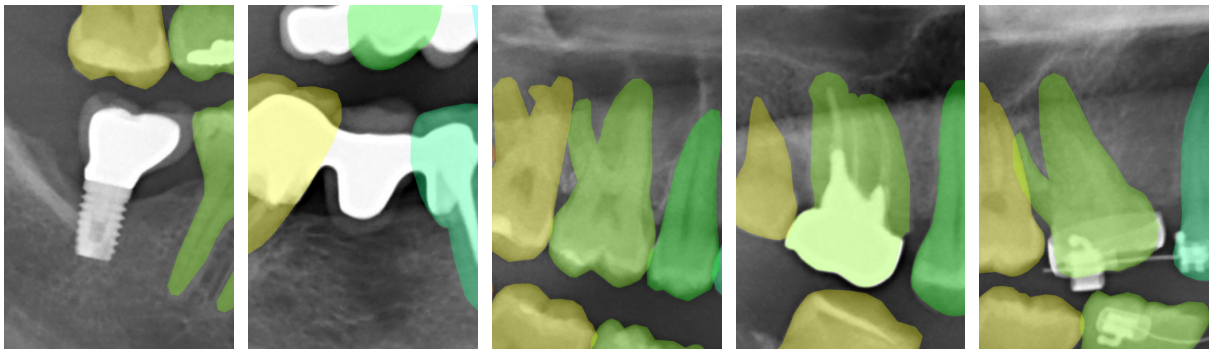
in the UFBA-UESB Dental Image Data Set while images of category 8 were subsampled. We conducted our HITL procedure so that, at the end of it, the 4,000 images' category proportions would be similar to the image database's. Table 4.1 summarizes the O^{2}PR dataset according to the number of images per category in the original dataset and the newly selected ones for the HITL.

4.2 MANUAL LABELING

An initial amount of labeled data is always required in the HITL semi-supervised procedure. Our work started from 450 images of the UFBA-UESB Dental Image dataset, in which four students labeled a subset, as detailed by Pinheiro et al. (2021). The students were two dentistry undergraduates and two STEM graduates experienced in the research of tooth segmentation and numbering on panoramic radiographs. An experienced radiologist supervised the students' work. Each student labeled about a fourth of the images using the COCO Annotator software and its polygon tool (BROOKS, 2019). The annotators should click on the tooth borders precisely as possible to delineate the teeth' outline, being expected crisper segmentations on sharp and well-focused images. In blurry images or regions, the students should picture the tooth contours based on their anatomical structure and label them accordingly, except when there was solid evidence for not doing so. Some criteria were defined as to be the standards for the labeling procedure:

(a) Implants should not be labeled;

(b) Prostheses should be incorporated into the tooth instances if they are associated with a single tooth root. If not, only the prothesis portions related to the tooth root in question should be considered;



(a) Implants. (b) Protheses. (c) Molar roots. (d) Restorations. (e) Appliances.

Figure 4.1: Label samples of the employed criteria for annotating implants, prostheses, molar roots, restorations, and dental appliances. In general, the labels should be more refined on sharp and well-focused images, while in blurry images, the annotators should rely more on the tooth anatomical structures.

(c) The palatine root of molars should be segmented, even if the spot is blurry;

(d) Restorations should be fused to the corresponding tooth instances;

(e) Dental appliances should be ignored. For labeling, the annotators should picture the tooth silhouettes when apparatuses, such as brackets and metal rings, blocked the visualization;

Figure 4.1 displays corresponding label samples of the aforementioned criteria. We followed the same criteria to label 400 additional images (40 per radiograph category), totaling 850 with the images labeled by Pinheiro et al. (2021) This last group of images compounded our test set for assessing the neural networks trained at each HITL iteration.

4.2.1 HITL setup

Our HITL methodology consisted of the cycle depicted in Figure 4.2: We trained a network with the available labeled radiographs and, subsequently, used its predictions as provisional labels for a new set of images. Our annotators verified these labels, which were incorporated in the next iteration into our training and validation sets for a new neural network training, restarting the cycle.

The adopted methodology demanded the setting of some parameters and conventions. For instance, we had to define the image subset size of the HITL labeled images and how to conduct the neural network training. A reasonable choice to consider was to label a single image using the model prediction and, subsequently, fine-tune the current weights of the neural network using the available labeled radiographs incremented by the newly available one. We did not follow this approach due to theoretical concerns and practical issues. The theoretical concerns were primarily due to the possibility of bias induction towards the first employed images and labels because, at each cycle step, the neural network training would start from a state where the weights would better suit those images. The practical

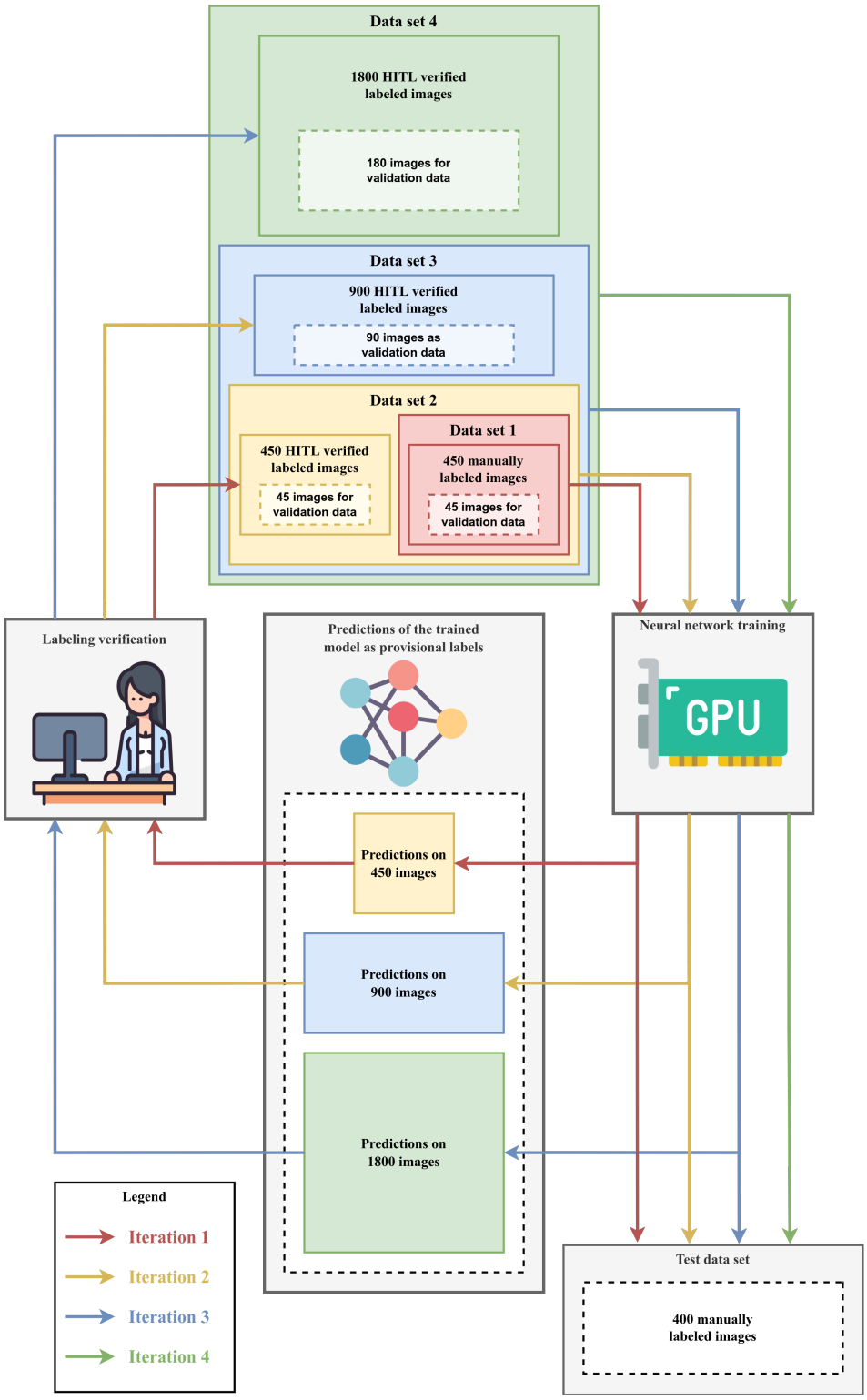


Figure 4.2: Our HITL setup: we trained a neural network with the available labeled radiographs and verify the model predictions on unlabeled images. In this setup, we doubled the number of labeled images at each iteration.

issues lay in the fact that our annotators should verify the model predictions in the same proportions. Our annotators had different time-availability and daily routines. It was impractical for them to verify the HITL annotations in small, consistent cycles (*e.g.*, the first annotator verifies the predictions over one image, the second annotator verifies the predictions over a second image, the third annotator validates another one, and so on). Letting the annotators verify the labels in any order would introduce untraceable biases, hindering us from performing in-depth analyses of the HITL outcomes.

In order to avoid any of the concerns mentioned above, we chose to verify the instance segmentation predictions for large sets of images at each cycle. Our setup labeled the same amount of images at each HITL iteration as the number of images used for training and validation. This way, we began with the 450 images from Pinheiro et al. (2021) split in training (405 images) and validation (45 images), doubling the set sizes at each cycle: The first iteration used 450 images (405 for training and 45 for validation), the second 900 images (810 for training and 90 for validation), continuing until we reached the fourth and final iteration containing 3600 images (3240 for training and 360 for validation). We hoped with this setup that, at each HITL iteration, we could perceive an improvement in the neural network performance, increasing the labeling quality.

4.2.2 Selecting the deep learning architecture for the HITL scheme

We conducted a benchmark with several state-of-the-art instance segmentation neural network architectures to define the model to be used in the HITL scheme. We selected the available architectures with the highest mean Average Precision (mAP) values on the COCO 2017 validation set. The mAP metric is a synonym for mAP@0.5:0.05:0.95, indicating that mAP is the mean value of the ten average precision (AP) scores with true positive thresholds of 0.5 up to 0.95 (inclusive) in steps of 0.05. The AP means, for each class, the area under the curve (AUC) of the precision-recall graph, where a true positive is computed when the prediction has an intersection over union (IoU) with a ground truth segmentation larger than the considered threshold. Usual AP metrics include AP50 (AP with 0.5 threshold value) and AP75 (threshold of 0.75). A threshold value equal to or larger than 0.85 is stringent, as well the final mAP metric. Throughout the HITL scheme, we used mAP as the primary metric in most of our experiments and analyses. In the end, our benchmark comprised seven architectures in total: A conventional Mask R-CNN (HE et al.; 2017), backboned by a ResNeXt-101-64x4d; Cascade Mask R-CNN (CAI; VASCONCELOS, 2019), backboned by a ResNeXt-101-64x4d; Mask R-CNN backboned by a ResNeSt-101 (ZHANG et al., 2020); Cascade Mask R-CNN with Deformable Convolutional Networks (DCN) (DAI et al., 2017) backboned by a ResNeXt-101-64x4d; Cascade Mask R-CNN backboned by a ResNeSt-101 (ZHANG et al., 2020); Hybrid Task Cascade (HTC) (CHEN et al., 2019a) with DCN backboned by a ResNeXt-101-64x4d; and DetectoRS (QIAO; CHEN; YUILLE, 2021) with HTC head backboned by a ResNet-50.

Each selected architecture introduced or adopted appropriate techniques to boost its COCO instance segmentation benchmark metrics. Mask R-CNN was the first architecture to extend the Faster R-CNN to instance segmentation by adding a mask branch. It also introduced the RoiAlign, a quantization-free layer, and employed a Feature Pyra-

mid Network (FPN) (LIN et al., 2017). The Cascade R-CNN or Cascade Mask R-CNN demonstrated the benefit of using a sequence of detectors with increasing IoU thresholds. DCNs are neural network modules that enhance the CNNs capabilities on transformation modeling, adding only a small computational overhead. The ResNeSt backbone stacks modular ResNet-like blocks that can attend to different feature-map groups. HTC's main contribution is a framework that interweaves the detection and segmentation tasks in a cascade fashion. Finally, DetectoRS improves object detection with two strategies: (i) Recursive Feature Pyramid, which modifies the FPNs through extra feedback connections to the bottom-up backbone layers, and (ii) Switchable Atrous Convolution, which are convolution operations with different atrous rates whose results are aggregated by switch functions. It is worth emphasizing that the benchmark ultimate goal was not to fairly compare methods and techniques (as the network and backbone sizes could vary significantly) but rather to specify a solid and reliable architecture to be used in our HITL scheme.

The benchmark protocol and the neural network training procedure for the HITL iterations were the same: It consisted in training each architecture for 150 epochs, with 90% of the available data as training set and 10% as validation set. We cropped all images to the reduced dimensions of 1876 $\times$ 1036 (159 pixels from the top and horizontally centered) to improve the network performances. These numbers and methodology came by roughly removing 80% of the extent between the outermost segmentations and the image borders of the 450 firstly labeled radiographs. This cropping may exclude tooth parts, or even the entire instances, hindering some applications, but, in the HITL, the human supervisor can catch these eventualities and correct them.

Each network performance was measured at the end of each epoch, and we saved the weights corresponding to the highest attained segmentation mAP. The optimizer was the stochastic gradient descent (SGD) with a 0.9 momentum value and no weight decay. We trained the models with eight Tesla V100 16GB GPUs with a batch size of 8 (one sample per GPU). We employed a linear warm-up strategy, linearly increasing the learning rate from 0 up to 0.024 in the first 40 epochs. Data augmentation was solely done through horizontal flipping, cautiously changing the tooth classes to their new corresponding numbers (right-sided teeth turned into left-sided teeth and vice-versa). Finally, we mention that the mask branches are class agnostic, *i.e.*, it only segments the object from the background. Table 4.2 summarizes the benchmark results (with the scores in green representing the highest one, while the smallest one are in red). The winner architecture was HTC, which also had the best values on all considered metrics, except on segmentation AP50 by a tiny margin. The HTC's final scores were also substantial, confirming it as a trustworthy option for our HITL scheme. We used the benchmark's resulting HTC neural network to start the labeling of new radiographs.

4.2.3 HITL labeling

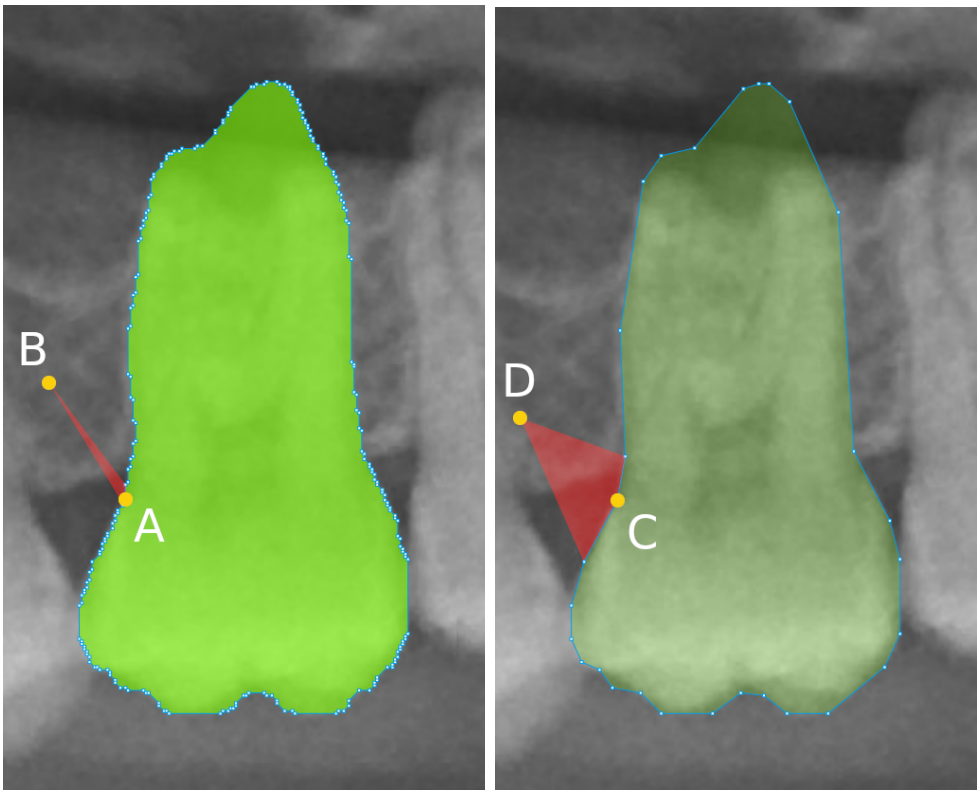
The HITL-based labeling started with the predictions of the HTC neural network. We labeled 450 radiographs in the first HITL iteration, as indicated in Figure 4.2. This iteration was considered experimental, as the annotators had not previously verified an-

Table 4.2: Summary of the benchmark results. The HTC architecture with DCN backboned by a ResNeXt-101-64x4d was the winner architecture and the used model for our HITL scheme.

Architecture	Backbone	Head	DCN	Detection			Segmentation		
				AP75	AP50	mAP	AP75	AP50	mAP
HTC	X-101-64x4d-FPN	HTC	✓	**0.913**	**0.983**	**0.795**	**0.958**	0.983	**0.802**
DetectoRS	ResNet-50	HTC		0.909	0.982	0.777	0.930	**0.984**	0.780
ResNeSt Cascade R-CNN	S-101-FPN	Cascade		0.871	**0.917**	**0.748**	**0.898**	**0.918**	**0.745**
Cascade R-CNN with DCN	X-101-32x4d-FPN	Cascade	✓	0.896	0.972	0.771	0.922	0.972	0.766
ResNeSt Mask R-CNN	S-101-FPN	FCN		0.873	0.931	0.753	0.913	0.931	0.755
Cascade R-CNN	X-101-64x4d-FPN	Cascade		0.901	0.982	0.763	0.939	0.982	0.768
Mask R-CNN	X-101-64x4d-FPN	FCN		**0.848**	0.969	0.752	0.916	0.978	0.758

notations from model predictions. Indeed, it quickly became notorious that manual image labeling is quite different from labeling verification. When labeling a radiograph from scratch, the annotator may promptly detect or localize the teeth and segment their instances using the annotator software mechanisms such as the polygon or brush tools. In the COCO Annotator software, the resulting area is filled with a colored layer to distinguish the already segmented objects from the others. On the other hand, when working on verifying neural network predictions, the human annotators must visually inspect the results and quickly confirm or correct the provisional labels. For that, the annotators can benefit from any software annotation tools, but in our case, they most frequently used the polygon point drag-and-drop feature. Two issues arise from this: (i) the filled segmented areas obstruct the instances, hampering the verification; (ii) the large number of points per segmentation slows down and hardens the corrections because point shift has less impact on the annotation. We mitigated these issues by changing the software source code of COCO Annotator, reducing the shape opacity, and lowering the number of control points through the Ramer–Douglas–Peucker algorithm with a tolerance of 2 pixels (DOUGLAS; PEUCKER, 1973). Figure 4.3 illustrates these modifications, evincing the new higher impact of point shift. Furthermore, we added a keyboard shortcut to toggle the annotation visualization, which was very helpful for the annotators.

We defined some correction criteria based on our observations during the labeling verification of the first HITL iteration. It was evident that the network predictions were outstanding, yet they were usually worse than manual annotations. This worse performance was mainly due to delicate details that could be polished such as the serrated segmentations originated from the network's low-resolution masks. Figure 4.4 shows samples of the serrated patterns on tooth crowns and on lower molars, which were highly frequent, especially on the former. For many applications, such as tooth detection and numbering, these tiny mistakes can be overlooked. However, we decided not to ignore those errors, as we want our data set to be general-purpose. In sum, the labels after correction should be as similar as possible to the manual labels. This determination slowed down the verification procedure significantly because our annotators had to make many tiny adjustments. With this main criterion defined, we proceeded with the other three iterations, reaching in the end 3,150 HITL labeled radiographs.

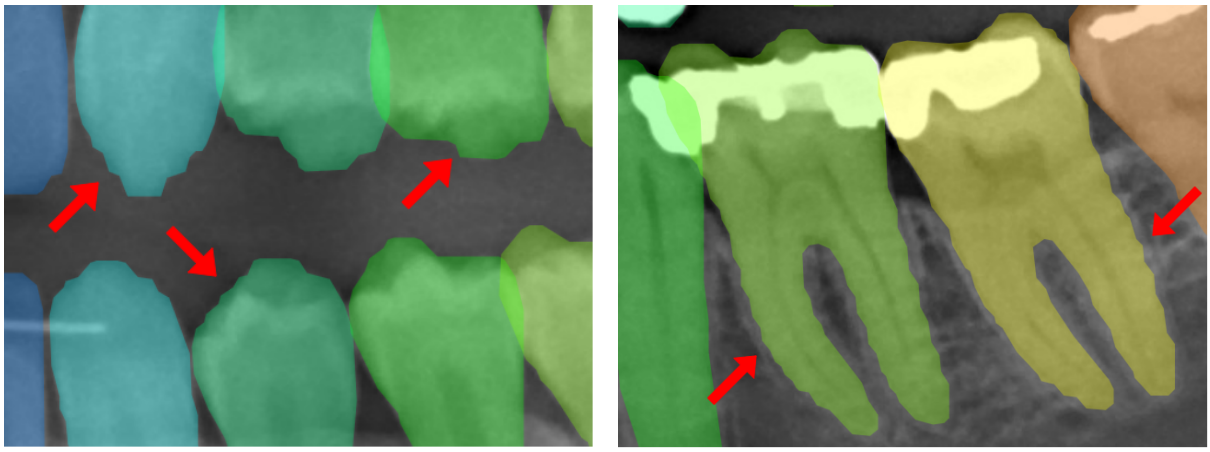


(a) Original code. (b) Changed code.

Figure 4.3: Illustration of the software visualization due to the code changes. The red area evinces the higher impact in the annotation when on a point shift. We reduced the shape opacities, easing the annotation verification, and lowered the number of control points. (a) Visualization of a tooth annotation with the original code and no tolerance in the Ramer–Douglas–Peucker algorithm and the impact on the annotation when point A moves to point B. (b) Visualization of a tooth annotation with the changed code and the impact on the annotation when point C moves to point D.

4.3 EVALUATION OF THE HITL RESULTS

The ultimate goal of our work was to create a labeled data set to boost research on dental panoramic radiographs. Under this perspective, the outcomes of the HITL procedure sufficed for our purpose, dispensing to report the performance of the trained models on a test set. However, we expect the deep learning community to heavily use our data set and increasingly employ the HITL concept to speed up the annotation process. Therefore, we performed detailed analyses on the HITL outcomes, including the evaluation of the trained networks on a separate manually labeled test data set. We aimed through these analyses to measure the HITL benefits and identify the main bottlenecks for better results.



(a) Serrated masks on tooth crowns. (b) Serrated masks on lower molars.

Figure 4.4: Samples of the serrated pattern due to the network's low-resolution mask predictions. This pattern frequently occurred on (a) tooth crowns and (b) lower molars, especially the former. The red arrows point to some spots with those patterns.

Table 4.3: Results of the trained neural networks in our HITL system on their corresponding validation data sets. The validation data sets comprise 10% of the available data at their HITL iteration. We highlight the best (green) and worst (red) results per metric.

Neural	Detection			Segmentation		
Networks	AP50	AP75	mAP	AP50	AP75	mAP
HTC 1	**98.3**	**91.3**	**79.5**	**98.3**	**95.8**	**80.2**
HTC 2	98.7	94.8	81.6	98.7	96.7	82.1
HTC 3	**98.9**	**97.1**	83.6	**98.9**	97.1	83.6
HTC 4	**98.9**	96.6	**86**	**98.9**	**97.7**	**85.9**

4.3.1 Model results on validation data

Table 4.3 synthesizes the detection and segmentation metrics (AP50, AP75, and mAP) attained by the trained neural networks in our HITL system, highlighting the best (green) and worst (red) results per metric. These metrics come from the best networks according to the segmentation mAP over the validation data sets, which comprised 10% of the available data at each HITL iteration. We call these networks HTC 1, HTC 2, HTC 3, and HTC 4. The number in their names corresponds to the iteration in which the network was trained.

When looking at the results of Table 4.3, we perceive an unmistakable increasing trend on the considered metrics, especially on the mAP ones, which contain the primary metric (mAP for segmentation). The increasing trend exists in the other looser metrics (AP50 and AP75) but is less pronounced. This difference was no surprise, as the selection of the network weights was according to the segmentation mAP, and the AP50 and AP75 values were already pretty high on the first HITL iteration (not much room for improvement).

Table 4.4: Results of HTC 1, 2, and 3 on the verified labeled from their predictions over 450, 900, and 1800 images, respectively. We highlight the best (green) and worst (red) results per metric.

Neural Networks	Detection			Segmentation		
	AP50	AP75	mAP	AP50	AP75	mAP
HTC 1	**82.0**	**79.2**	**71.2**	**82.0**	**80.3**	**74.0**
HTC 2	88.4	87.5	79.6	88.4	87.6	82.1
HTC 3	**89.4**	**88.3**	**80.9**	**89.4**	**88.7**	**82.7**

4.3.2 Model results on HITL data

The HITL labeled data is an alternative to the validation data sets for model evaluation. In this case, we evaluate the model performances on the verified annotations from the model predictions. The main advantage here is that we do model assessment in unseen and large data. We performed this analysis using the threshold values computed with the procedure described in Section 4.2.3. Table 4.4 synthesizes the results of HTC 1, 2, and 3 on, respectively, 450, 900, and 1800 radiographs labeled from their corresponding predictions, also highlighting the best and worst results (we did not assess HTC 4 as no labels came from its predictions). All metrics increased at each iteration, but the most significant performance boost came from HTC 1 to HTC 2, when there was still significant room for improvement. The mAPs of HTC 3 were the best and surpassed the 80 points on both the detection and segmentation tasks.

Using the HITL labeled data as test data mitigated the problem of the biased estimation of the network results and the computed metrics revealed consistent results. However, some issues persisted: we evaluated the networks on distinct images with different radiograph category proportions and disregarded HTC 4. For those reasons, it proved imperative to label a separate set of images for a consistent comparison. For that, we assessed the networks on 400 images (40 for each radiograph category), which we manually labeled from scratch and comprised our test data set as pointed out in Section 4.2.

4.3.3 Model results on test data

Besides a consistent comparison, our test data set allows unbiased model assessment, as we manually labeled 40 images per radiograph category exclusively for model evaluation. Table 4.5 synthesizes the results of each trained HTC network over the test data set accordingly to the detection and segmentation AP50, AP75, and mAP metrics. One can observe that all segmentation metrics increased at each HITL cycle, being a favorable indication for the HITL results. The detection metrics also display a prominent increasing tendency, but they may oscillate slightly. These aggregate results give no insights on the specifics of the network performances. In order to solve that, we analyzed the segmentation mAP per dentition and tooth type.

Table 4.6 split the segmentation metrics into the dentition types: permanent and

Table 4.5: Performance metrics of each trained neural network in our HITL system on the manually annotated test data set. The test data set comprised 400 images (40 per radiograph category). We highlight the best (green) and worst (red) results per metric.

Neural	Detection			Segmentation		
Networks	AP50	AP75	mAP	AP50	AP75	mAP
HTC 1	**91.9**	**83.4**	**72.0**	**92.2**	**87.2**	**72.0**
HTC 2	95.4	87.8	75.5	95.7	89.5	74.6
HTC 3	97.0	**88.0**	75.4	97.6	90.3	75.6
HTC 4	**98.4**	87.4	**76.0**	**98.9**	**91.8**	**77.4**

Table 4.6: Segmentation results on the test data set according to the dentition tooth types: Permanent and deciduous. We highlight the best (green) and worst (red) results per metric. The metrics over the deciduous teeth were worse but improved significantly over the HITL iterations.

Neural	Permanent			Deciduous		
Network	AP50	AP75	mAP	AP50	AP75	mAP
HTC 1	**99.0**	**97.0**	**82.0**	**81.5**	**71.4**	**56.1**
HTC 2	**99.0**	97.3	82.3	90.4	77.0	62.3
HTC 3	**99.1**	**97.5**	82.4	95.3	78.7	64.7
HTC 4	**99.1**	**97.6**	**82.7**	**98.5**	**82.7**	**69.0**

deciduous. The highest (green) and smallest (red) values per metric indicate that the best predictions are from HTC 4, while the worst are from HTC 1. These metrics demonstrate small but consistent increasing performances over the HITL iterations on permanent dentitions. On the other hand, the segmentation results on deciduous teeth improved significantly: at least 15% on all metrics. The segmentation mAP increased 12.9 points, which represents a 23.0% gain. This improvement is due to the initial lack of training data increment (the deciduous teeth constitute approximately 3% of the training instances). The rare occurrences of deciduous teeth, especially the central and lower lateral incisors, hindered the first trained networks from generalizing on those tooth types.

Finally, we broke the segmentation mAP by dentition and tooth type in Tables 4.7 (permanent teeth) and 4.8 (deciduous teeth), highlighting the best (green) and worst (red) results of the networks per tooth type. Table 4.8 also brings the number of instances presented in the training data sets to illustrate the initial lack of training data. No lower central incisor was present in the first training iteration (HTC 1), and only 14 were present in the last (HTC 4). The additional data on those less frequent tooth types resulted in significantly higher metric values. Table 4.7 unveils that, on the permanent teeth, the HITL was more beneficial on the segmentation of the upper teeth than the lower ones. HTC 4 performed better on permanent lower incisors and permanent upper teeth than HTC 1. According to our annotators, the upper teeth, particularly the premolars and molars, are harder to segment. We consider this fact, along with the improved metrics on those tooth types propitiated by the HITL scheme, to subsidize the use of deep learning-

Table 4.7: The mAP results on the test data set per permanent tooth type. We highlight the best (green) and worst (red) results per tooth. The HITL benefited more the metrics over the more challenging to segment teeth, such as the upper premolars and molars.

Dental Arch	Neural Network	Incisors		Canines	Premolars		Molars		
		Central	Lateral		1st	2nd	1st	2nd	3rd
Upper	HTC 1	**85.1**	**83.8**	**84.0**	**72.9**	**79.9**	**78.4**	**80.7**	**77.2**
	HTC 2	85.8	83.9	84.4	73.4	**79.9**	80.0	80.8	77.7
	HTC 3	85.8	84.0	**84.8**	73.4	81.0	**80.1**	81.1	77.8
	HTC 4	**85.9**	**84.4**	84.5	**74.7**	**81.1**	79.9	**81.6**	**78.0**
Lower	HTC 1	**79.0**	**80.8**	**85.3**	**84.7**	**87.3**	**85.1**	**84.5**	**82.9**
	HTC 2	79.6	80.9	**85.9**	**85.4**	87.1	84.4	84.1	83.5
	HTC 3	**79.9**	81.2	85.8	85.0	**87.0**	**84.0**	**83.8**	**82.9**
	HTC 4	79.6	**81.6**	85.7	85.3	**87.3**	84.6	**84.5**	**83.8**

Table 4.8: The mAP results on test data set and instance count (in parentheses) on training sets per deciduous tooth type. We highlight the best (green) and worst (red) results per metric. The metrics over the deciduous teeth were worse on average but improved significantly over the HITL iterations.

Dental Arch	Neural Network	Incisors		Canines	Molars	
		Central	Lateral		1st	2nd
Upper	HTC 1	**35.2 (3)**	58.8 (22)	**64.8 (70)**	64.7 (58)	65.5 (78)
	HTC 2	64.6 (7)	**55.1 (28)**	64.9 (110)	**59.8 (79)**	**62.8 (117)**
	HTC 3	66.4 (19)	62 (56)	**65.2 (206)**	59.2 (152)	65.9 (225)
	HTC 4	**74.7 (45)**	**64.4 (106)**	**64.8 (423)**	**65.4 (322)**	**68.8 (437)**
Lower	HTC 1	**0 (0)**	63 (10)	**69 (52)**	**67.8 (61)**	72.2 (73)
	HTC 2	41.2 (2)	**62.8 (12)**	**73.2 (74)**	66.4 (83)	72.4 (110)
	HTC 3	53.7 (8)	66.4 (26)	73 (149)	**64.7 (161)**	**71 (212)**
	HTC 4	**66.8 (14)**	**72.6 (54)**	72.8 (304)	67 (321)	**73 (440)**

based assist tools to aid in challenging cases. In contrast, the metrics on permanent lower premolars and molars stagnated or oscillated a bit. The metrics on those teeth, which are large and straightforward to segment teeth, were already pretty high on the first iteration, resulting in less room for improvement.

4.3.4 Numbering analysis on test data

The numbering task consists in detecting all tooth instances and correctly classifying them. This task has a direct practical value, as the radiologists must inform the patients' missing teeth in the reports. Additionally to this practical application, numbering is helpful to assess the HITL benefits on a task less sensitive to coarse predictions.

We evaluated the model performances according to their errors, which in the numbering task may be grouped in three types:

Table 4.9: Neural network performances according to the error types on the numbering task for a 0.5 IoU detection threshold over the test data set. All errors shrank at each HITL iteration.

Network	False Negatives	False Positives	Misclassifications	Total Errors	True Positives
HTC 1	111	74	216	401	11,390
HTC 2	85	53	179	317	11,453
HTC 3	78	52	166	296	11,473
HTC 4	41	42	158	241	11,518

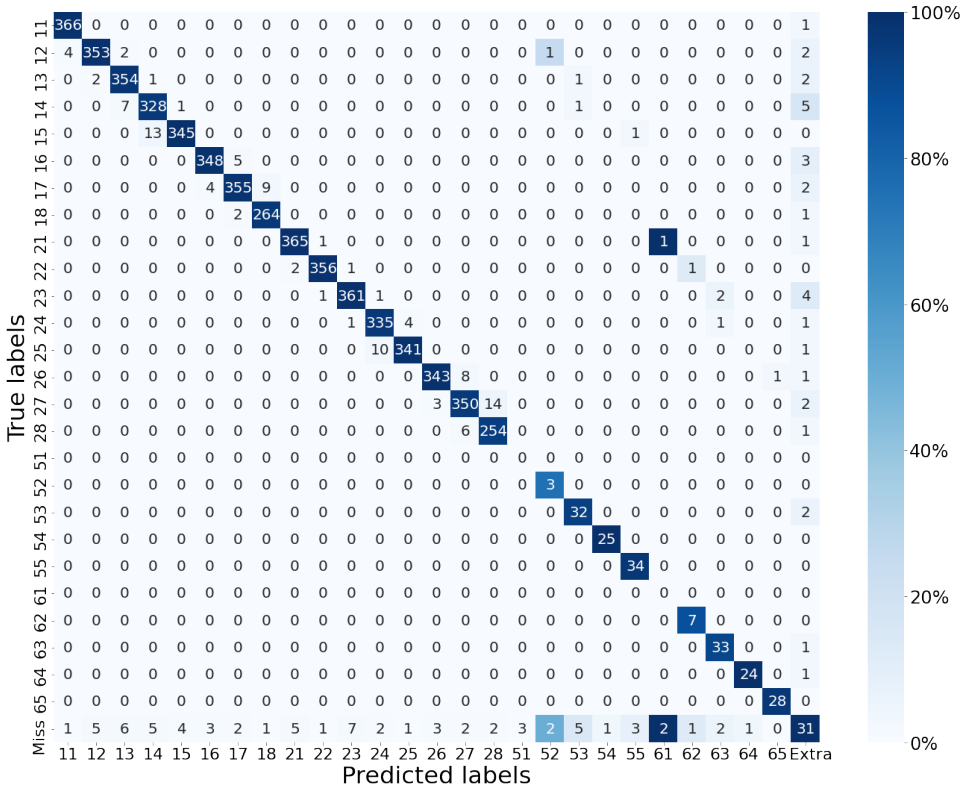
- False negatives (when the model does not detect an instance);
- False positives (when the model detects something that is not an instance of an object of interest);
- Misclassifications (when the model correctly detects an object instance but classifies it wrongly).

We synthesize these values along with the total errors and the true positives for the trained networks in Table 4.9 using a 0.5 IoU detection threshold. One can perceive a consistent performance improvement trend in all values over the iterations. The most significant advancement was over the false negatives, reduced by 63%. The misclassification errors shrank 27%.

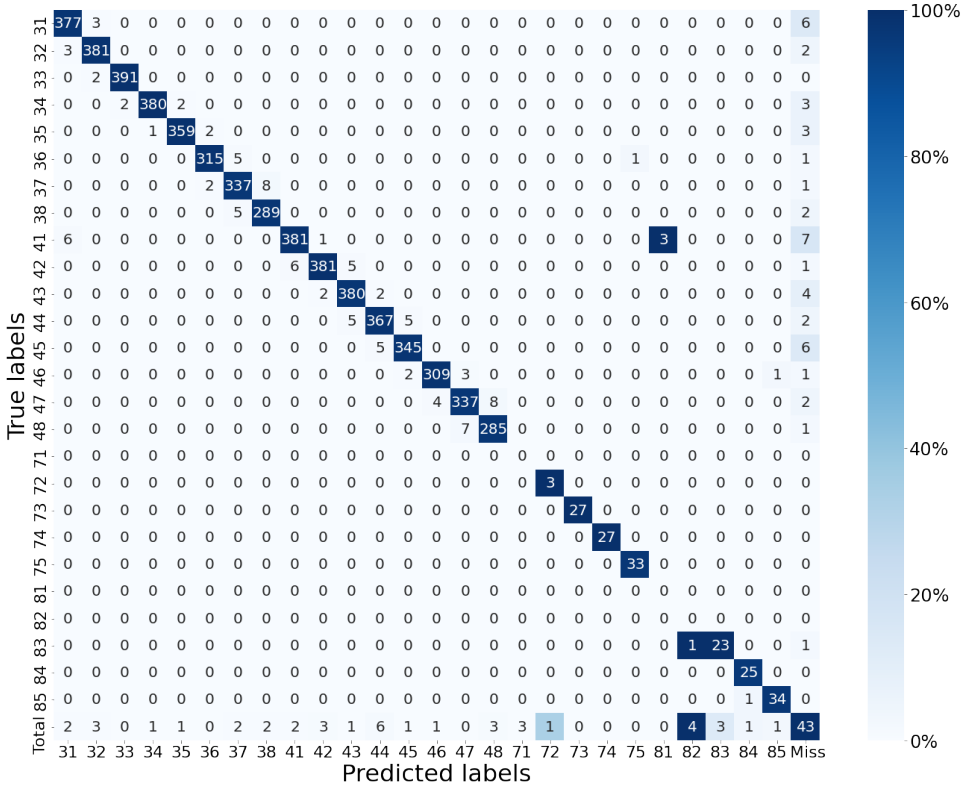
The aggregated results from Table 4.9 do not allow a detailed analysis of the numbering errors. To solve that, we plotted the confusion matrices according to tooth types. For brevity, we depict only HTC 1's and HTC 4's detection confusion matrices in Figures 4.5 and 4.6, in which we split the matrices into the upper teeth and lower teeth parts for visualization purposes. This division is not harmful to the analyses because misclassifications between those groups are rare. A performance boost can be observed in all tooth groups by comparing HTC 1's and HTC4's confusion matrices. The upper teeth were slightly easier to detect and classify for both networks. The deciduous teeth were more challenging to detect correctly but easier to classify than permanent ones. The misclassifications were essentially among nearby, same-function teeth, especially the premolars and molars. The numbering of premolars and molars may be quite challenging in some circumstances, such as on unhealthy missing-tooth mouths, where dubious situations may occur even for human experts.

4.3.5 Labeling time analysis

In a HITL setup, we are not only interested in labeling quality, but also labeling speed-up. Therefore, we monitored the HITL labeling verification and the radiograph manual labeling times. Figure 4.7 compares those two labeling approaches according to each annotators' average time. These time values were measured during the third iteration, in which we asked our annotators to clock their correction and manual labeling. For the

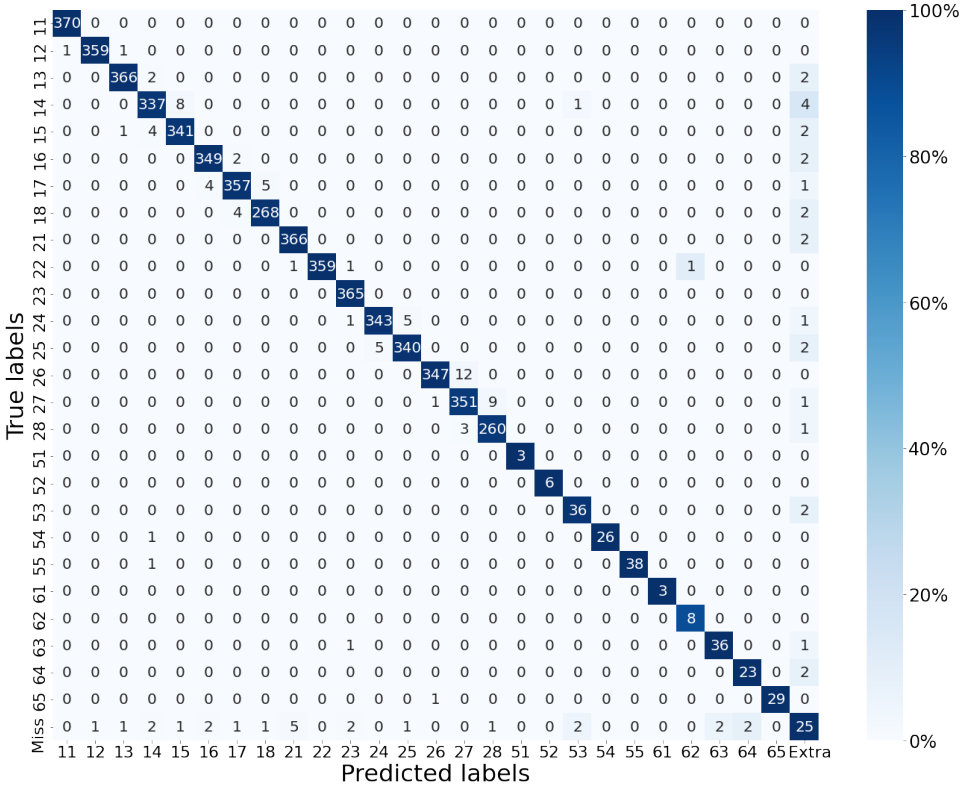


(a) HTC 1's upper teeth confusion matrix.

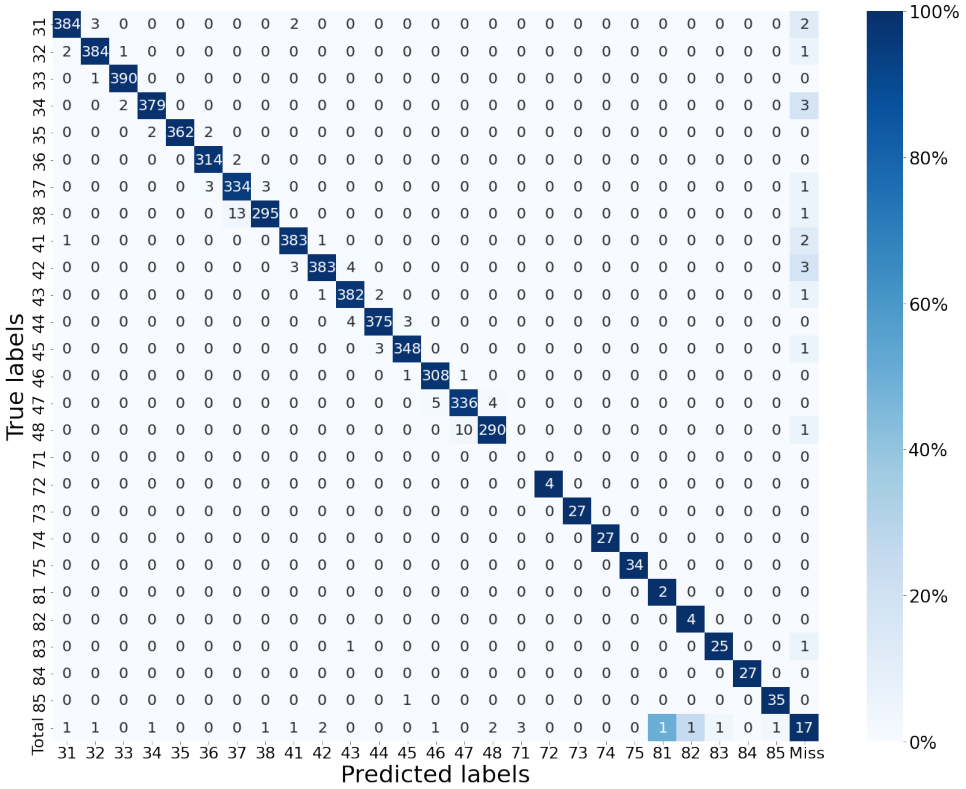


(b) HTC 1's lower teeth confusion matrix.

Figure 4.5: HTC 1's upper and lower teeth confusion matrices for a 0.5 IoU detection threshold. The last lines are for the false negatives per tooth type, while the last columns are for the false positives.



(a) HTC 4's upper teeth confusion matrix.



(b) HTC 4's lower teeth confusion matrix.

Figure 4.6: HTC 4's upper and lower teeth confusion matrices for a 0.5 IoU detection threshold. The last lines are for the false negatives per tooth type, while the last columns are for the false positives.

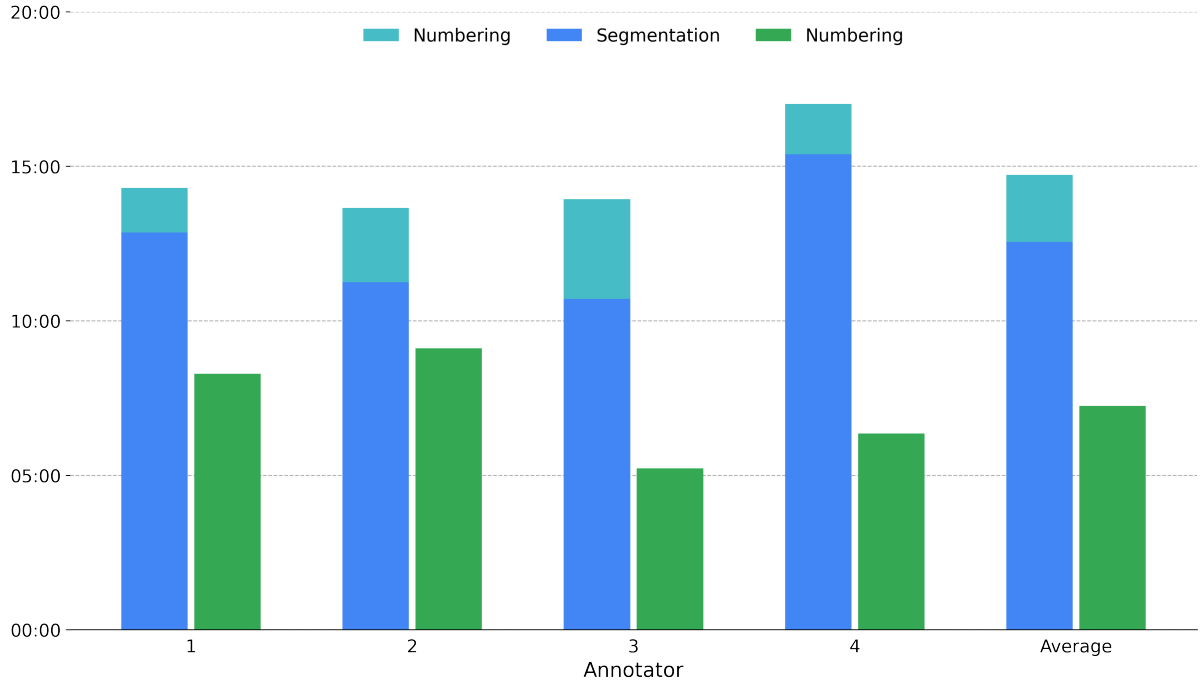


Figure 4.7: Comparison of each annotator time for HITL labeling verification time against manual labeling time, the latter split into segmentation and tooth numbering. The HITL labeling lasts considerably less (51%) than manual labeling.

manual labeling, they split their time into segmentation and tooth numbering. The latter is the time to type and assign the tooth class.

From Figure 4.7, one can perceive that the labeling verification procedure was significantly faster than manual labeling. Labeling radiographs manually lasted on average 14 minutes and 43 seconds per radiograph, while labeling using the HITL concept took 14 minutes and 43 seconds, a 51% time reduction. The annotation verification was faster than manual segmentation, even if we disregarded the numbering procedure. In that case, the HITL approach reduced the labeling time by 42% compared to manual labeling, which took 12 minutes and 33 seconds on average. If we considered the 51% time reduction, the HITL procedure saved more than 390 continuous working hours.

4.3.6 HITL bottlenecks

We investigated possible bottlenecks that could have significantly slowed down the HITL verification procedure. Already in the first iteration, in which we established the verification protocol, our annotators mentioned several times the presence of serrated segmentation that comprises most of the correction time. This serrated pattern came from the 28×28 low-resolution masks and appeared especially on the tooth crowns, but it was also frequent on the other parts of large and complex-shaped tooth instances, such as molars. The annotators with no deep learning background considered these incongruous masks somewhat surprising, as the crowns are usually well-defined and easier for humans

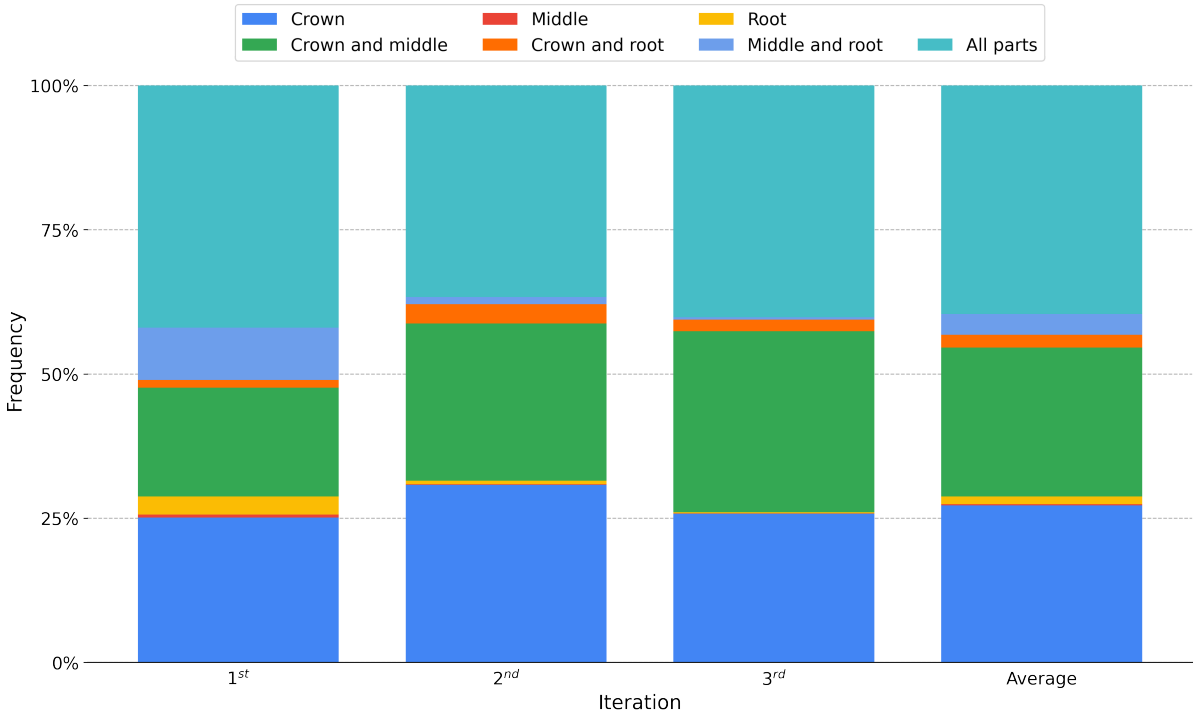


Figure 4.8: Frequency of corrections according to tooth part. The crown segmentation was adjusted in more than 85% of the corrected instances in all HITL iterations.

to segment. For segmentation models, the tooth crowns are fine-detailed objects with acute borders, which hampers the segmentation task.

In order to understand how much impact these jagged contours have in the HITL, we quantified the correction fractions related to tooth parts: Crown, middle, root, or a combination of them, including correction on all tooth parts. We automatically split each tooth instance into these three parts in the vertical axis for this analysis and measure the frequency of the modifications in each part, disregarding the size of the changes.

Figure 4.8 summarizes the obtained results, showing the frequency of parts where the correction took place at each iteration. One can perceive that, in all iterations, adjustments in the crown segmentation have been made in more the 85% of the corrected instances. These adjustments were highly frequent, mainly due to the serrated patterns and heavily slowed down the verification process. Possible solutions to reduce this issue when neglecting these tiny errors is not an option include increasing the segmentation mask resolution, employing a two-stage instance segmentation approach, or using a more specialized method, such as the PointRend module.

4.3.7 Qualitative analysis

The quantitative analyses guided our qualitative analyses. We focused on the best and worst results according to the primary metric, comparing the ground truth with the network predictions and the verified labels. Figures 4.9 (a) and (b) illustrate the best and

the worst HTC 4's results, respectively, according to the segmentation mAP on the test data set. Figure 4.9 (a) corresponds to a well-focused, crisp and clear radiograph from a 32-teeth healthy mouth, characteristics common to the best results. From the zoomed area, we see that the annotation correction led to a final label closer to the ground truth and also less noisy.

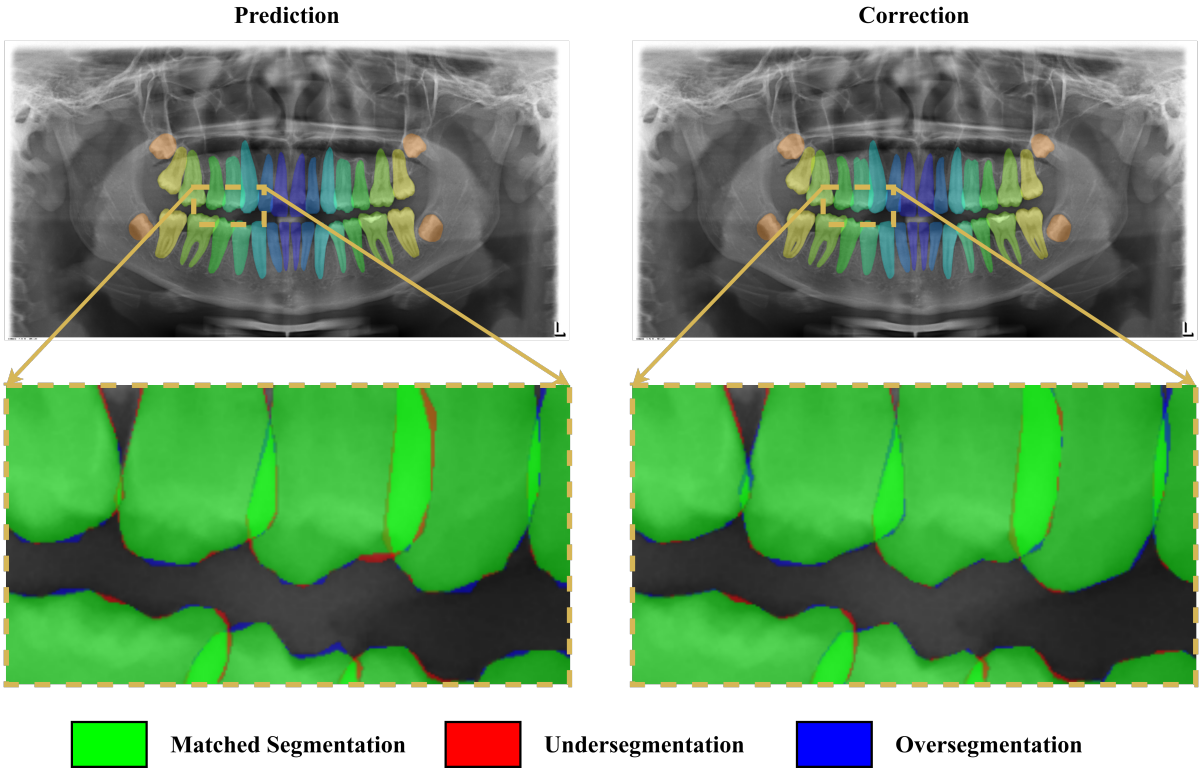
The worst result, illustrated in Figure 4.9 (b), came from a slightly blurry image from an unhealthy mouth, a common pattern in the radiographs of the worst results. However, in this case, the network performance was reasonably good, and the low metric was due to main factors. First and most important, the annotator wrongly labeled the teeth 32, 33, 34 and 35 respectively as teeth 31, 32, 33 and 34, probably due to a sequence of typos, which reduced the segmentation mAP significantly. Second, the presence of radiolucent material prostheses and restoration encumbered the segmentation task for both model and annotator. The zoomed area shows that model undersegmented those spots, which were adjusted by the annotator, but still missed some areas. The other annotation corrections smoothed the noisy borders and reduced the difference from the ground truth labels. We additionally illustrate in Figure 4.9 (c) a sample result on a mixed dentition mouth. These radiographs are challenging for models and human annotators due to overlapping. In this particular image, there are also occlusions between posterior teeth, hardening the task. However, the model prediction proved to be adequate, even before the labeling verification. The zoomed area shows that the corrections reduced the gap to the manual ground truth labels, but there were some divergences for root segmentation of teeth 54 and 55.

4.4 SUBMISSION PLATFORM, EVALUATION PROTOCOLS, AND BASELINES

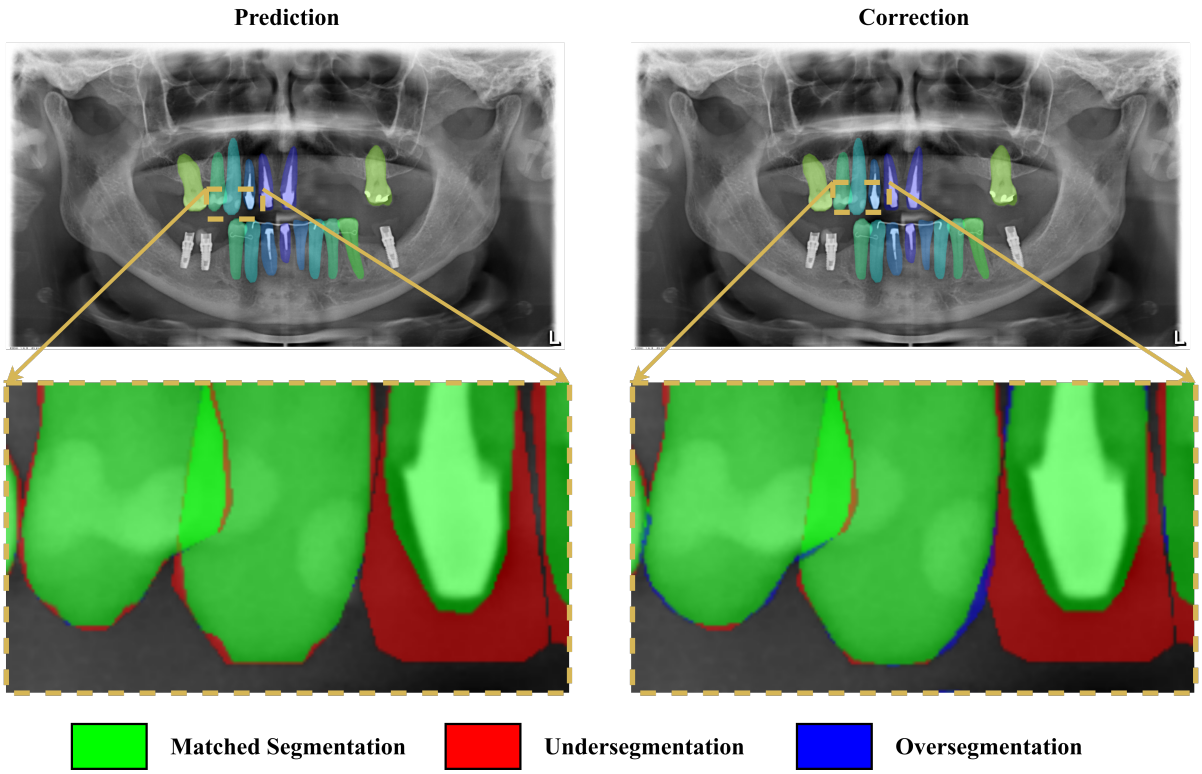
Our data set comprises 4000 labeled radiographs (850 manually labeled and 3150 HITL labeled), from which 2000 are used for solution assessments in the **OdontoAI platform**[1]. The platform consists of a website where researchers can submit their predictions in a standardized fashion, enabling a fair benchmarking of the proposed methods. We provide 2000 radiographs along with their labels (650 manually labeled and 1350 HITL labeled radiographs) for model training and validation. The remaining 2000 images (1800 labeled in the HITL scheme and 200 manually labeled) do not have their labels publicly available and consist of the platform test set. We also provide in the platform precise instructions on how to submit solutions and open-source codes of the used metrics and for creating the submission files.

We configured three benchmarks for the OdontoAI platform, comprising classical computer vision tasks useful for analyzing dental panoramic radiographs. The tasks are tooth **instance segmentation**, **semantic segmentation**, and **numbering**, which we detail in the following sections together with the selected metrics. As baselines, we included the results of neural networks trained on the 2000 publicly available labeled radiographs architectures using the architectures presented in our instance segmentation benchmark (Section 4.2.2).

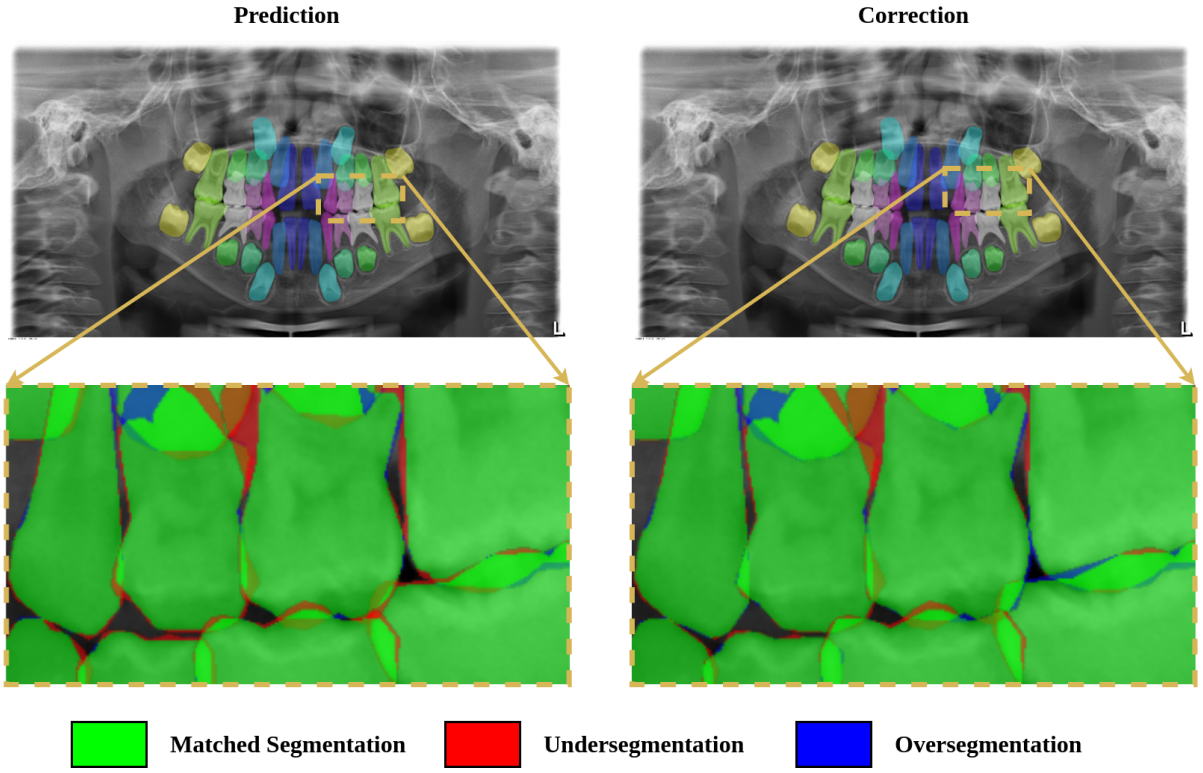
[1]The platform link will be available upon the article's acceptance and publication.



(a) HTC 4's best result, which happened on a well-focused, crisp and clear radiograph from a 32-teeth healthy mouth, characteristics common to the best results.



(b) HTC 4's worst result, which happened on a slightly blurry radiograph from an unhealthy mouth with radiolucent material prostheses.



(c) Sample of HTC 4's results on a mixed dentition radiograph.

Figure 4.9: HTC 4's best and worst results according to the segmentation mAP on the test set, and an additional result sample on a mixed dentition radiograph. The illustrations compare the predictions before and after the corrections by the annotators. The zoomed areas highlight the matched segmentation, undersegmentation, and oversegmentation with the ground truth labels, evincing that the corrections led the final labels to be less noisy and closer to the ground truth.

4.4.1 Instance segmentation task

The instance segmentation task is a straightforward application of our data set. This task is challenging and comprehensive, as it combines instance detection and segmentation. Many researchers investigate instance segmentation due to its usefulness, but lack of data may be an issue. Our data set solves this problem.

We chose mAP as the main metric to evaluate instance segmentation. A rigid metric is necessary, as our experiments showed that the AP50 and AP75 are rather loose metrics to the task. The adopted metric, mAP, is not only stricter but also more comprehensive, being suitable for the instance segmentation task benchmark of the OdontoAI platform. AP50 and AP75 are included as secondary metrics as well the equivalent metrics for detection with bounding boxes. Table 4.10 illustrates a sample of the benchmark ranking available in our platform, with the attained results by the baselines.

Table 4.10: A sample of the OdontoAI platform benchmark ranking for the instance segmentation task with baselines.

Rank	Architecture	Detection			Segmentation		
		AP50	AP75	mAP	AP50	AP75	mAP
1	HTC	**0.924**	0.964	**0.821**	**0.941**	0.964	**0.821**
2	DetectoRS	0.920	**0.967**	0.803	0.933	**0.967**	0.809
3	Cascade R-CNN	0.893	0.951	0.786	0.920	0.952	0.790
4	Cascade R-CNN with DCN	0.875	0.930	0.773	0.886	0.931	0.770
5	Mask R-CNN	0.868	0.935	0.749	0.893	0.936	0.760
6	ResNeSt Cascade R-CNN	0.866	0.903	0.764	0.849	0.903	0.652
7	ResNeSt Mask R-CNN	**0.825**	**0.879**	**0.726**	**0.828**	**0.880**	**0.637**

Table 4.11: A sample of the OdontoAI platform benchmark ranking for the semantic segmentation task with baselines.

Rank	Architecture	Accuracy (%)	Specificity (%)	Precision (%)	Recall (%)	F1-score (%)	IoU (%)
1	HTC	**98.8**	**99.5**	**98.2**	**96.2**	**97.2**	**94.5**
2	DetectoRS	98.7	99.4	97.8	96.1	96.9	94.1
3	Cascade R-CNN	98.7	99.4	97.7	96.1	96.9	94.0
4	Cascade R-CNN with DCN	98.7	**99.5**	98.0	95.6	96.8	93.8
5	Mask R-CNN	98.6	99.3	97.4	95.9	96.6	93.5
6	ResNeSt Cascade R-CNN	**97.0**	98.6	94.7	**91.2**	92.9	86.7
7	ResNeSt Mask R-CNN	**97.0**	**98.5**	**94.3**	91.3	**92.8**	**86.5**

4.4.2 Semantic segmentation task

Semantic segmentation is also a basilar task in computer vision, being the reason why we included it in our platform's benchmarks. The tasks consist of segmenting classes precisely as possible, disregarding object instances. In our benchmark, there is only one class (tooth), and the researchers should propose methods to distinguish it from the background. Due to this dichotomic nature, we employed the usual metrics for binary segmentation: accuracy, specificity, precision, recall, f1-score, and IoU. The latter is the main metric, and it is equivalent to the binary mIoU, a commonly used metric in semantic segmentation benchmarks.

Table 4.11 shows a sample of the benchmark ranking at the OdontoAI platform for the semantic segmentation task. The baseline metrics were computed after converting the instance segmentation predictions of the networks into segmentation masks.

4.4.3 Numbering task

Finally, we included the task of "numbering," which is almost equivalent to the multi-label classification computer vision task. It slightly differs from the multi-label classification task because one or more supernumerary teeth may appear. In the numbering task of our benchmark, the goal is to predict the present teeth in the panoramic radiograph. Although this task may not be advantageous as a preprocessing step for analyzing panoramic radiographs, it naturally appears in practical applications such as form fillings

Table 4.12: A sample of the OdontoAI platform benchmark ranking for the numbering task with baselines.

Rank	Architecture	Exact Match (%)	Micro Accuracy (%)	Micro Precision (%)	Micro Recall (%)	Hamming Loss
1	HTC	**67.9**	**98.6**	**98.8**	**98.5**	**0.0143**
2	DetectoRS	66.2	98.4	98.7	98.3	0.0164
3	Cascade R-CNN with DCN	65.7	98.4	98.7	98.3	0.0161
4	Cascade R-CNN	62.8	98.2	98.5	98.2	0.0177
5	ResNeSt Cascade R-CNN	60.7	97.9	98.4	97.8	0.0206
6	Mask R-CNN	59.0	98.0	98.5	97.8	0.0197
7	ResNeSt Mask R-CNN	**56.3**	**97.7**	**98.4**	**97.3**	**0.0231**

and automatic report generation. While reports customary document the patient's permanent missing teeth, the OdontoAI platform's numbering task expects a list of present teeth. We chose this conventional because deciduous and supernumerary teeth may occur.

Our experiments showed that it is easy to identify the present teeth correctly. Through a general instance segmentation (HTC 4 neural network), the numbering task resulted in only 241 errors among false positives, false negatives, and misclassifications. Therefore, we chose a rather rigorous metric main metric, "exact match," in which a true positive is only taken into account when all tooth numbering predictions are correct. Other than the main metric, the OdontoAI platform includes other metrics, such as micro accuracy, micro precision, micro recall and Hamming loss. The Hamming loss averages the fraction of the incorrect predictions for each label. We illustrate in Table 4.12 a sample of numbering benchmark ranking found at the OdontoAI platform.

4.4.4 Training Instance Segmentation Network for Tooth Crop Generation

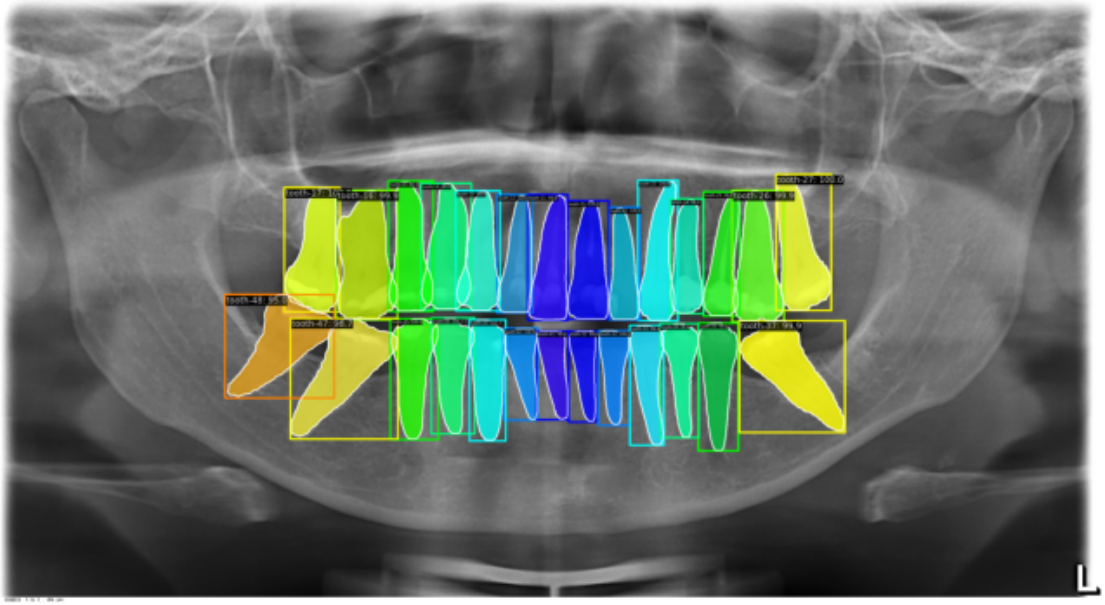
The experiments with the HITL approach confirmed that we could train sufficiently accurate instance segmentation networks. As a result, we proceeded with generating the tooth crops necessary for our framework to function. The main goal here was to train an instance segmentation neural network, specifically the HTC, as it was the winner architecture in our coducted bencmark, to detect and number the teeth to generate the crops subsequently. Following the previous setup, the trained HTC used a ResNeXt neural network as its backbone with 101 layers and a cardinality of 64 (XIE et al., 2017). The initial weights of this network were derived from the training on the ImageNet dataset to leverage the transfer of the learning technique later. The training data comprised 4,000 images from the O^2PR dataset. No data was allocated for testing, emphasizing that the goal was not efficiency measurement but tooth crop generation for the subsequent phases.

Data augmentation was purely horizontal flips, carefully changing the labels of the teeth from the right side to the left and vice versa. To optimize network performance, the radiographs were cropped from their prevailing 2,440 × 1,292 pixels to 1,876 × 1,036, removing 159 pixels from the top, which resulted in more focused teeth. The batch size was 1 (one), and the optimizer was stochastic gradient descent, with a learning rate of 0.0015, momentum of 0.9, and no weight decay. The threshold value for tooth detection was 0.5. The network was trained for 20 epochs in an NVIDIA GeForce GTX TITAN X. After training the neural network; it was applied to the 12,824 unlabeled images from the RPR and TRPR datasets. This allowed one to create tooth crops for all the radiographs

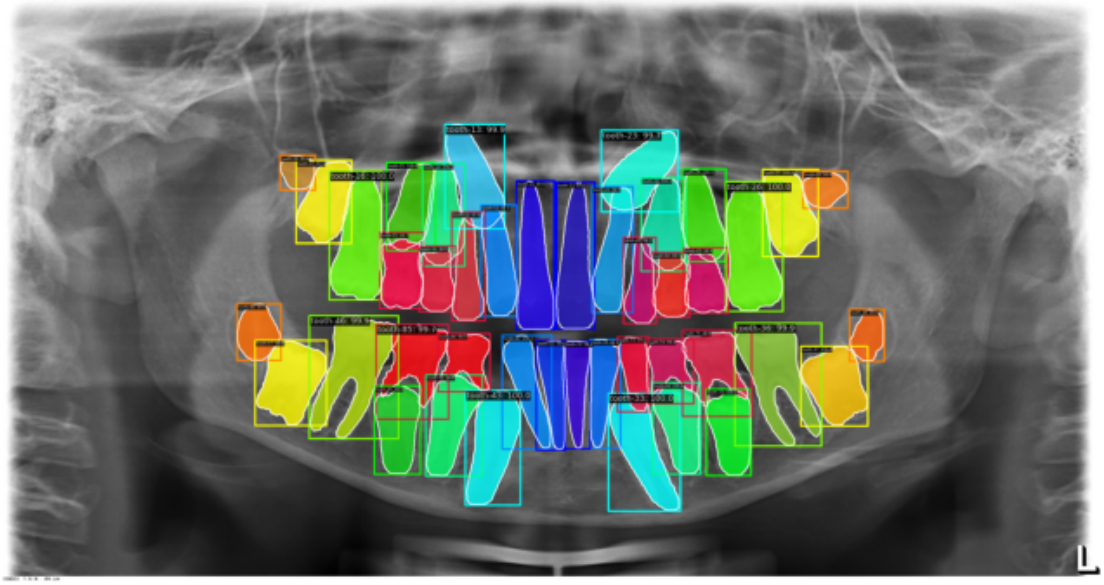
in the used images. Fig. 4.10 shows two samples of the instance segmentation results, in an adult's and a child's mouth, using the color code introduced in Fig. 2.1, demonstrating its strong performance.

4.5 CLOSURE

This chapter detailed the construction process of the datasets employed in this study, which included two distinct datasets: full-size labeled panoramic radiographs and tooth-centered labeled crops. The adoption of the Human-in-the-Loop (HILT) approach was particularly effective, significantly accelerating the labeling process for the full-size dataset. Instance segmentation neural networks were trained on these datasets, yielding excellent performance. These promising results provided the necessary confidence to leverage the trained networks to generate the tooth crop dataset based on network predictions.



(a) Instance segmentation results of a panoramic radiograph of an adult's mouth.



(b) Instance segmentation results of a panoramic radiograph of a child's mouth.

Figure 4.10: Qualitative results of the trained instance segmentation neural network, using the color code introduced in Fig. 2.1.

Chapter 5

CLASSIFICATION OF DENTAL CONDITIONS

5.1 INTRODUCTION

The crop generation discussed in the previous Chapter reduced the instance segmentation problem to a classification task. This change in perspective clearly simplifies the problem, as classification tasks are easier than detection tasks. Under these circumstances, we were able to use Vision Transformer (ViT), a state-of-the-art classification network. In our setup, another advantage emerges: we could leverage the unlabeled data from the RPR and O^2PR datasets by utilizing Masked Autoencoders (MAE) directly in conjunction with the ViT. We discuss our experimental analysis in the following.

5.2 EXPERIMENTAL ANALYSIS

We conducted a thorough experimental analysis to evaluate the performance of the different setups. The goal was to assess the impact of MAE-based pretraining and varying crop sizes on classification accuracy. For each setup, we measured the network's classification performance across a range of dental conditions, tracking the MCC metric on both the validation and test sets. We also set baseline networks without pretraining providing a reference point for comparison.

5.2.1 Neural network pretraining

The MAE technique was exploited to pretrain neural networks for subsequent transfer learning to final classification networks for each dental condition. All the available data of tooth crops was used for pretraining the ViTs, reserving some images for validation and testing purposes (see Table 3.4). The experiment encompassed three scenarios: the first employed a baseline network devoid of pretraining, the second involved networks pretrained on the ImageNet dataset, and the third used custom-generated tooth crops. Each scenario was executed twice, accommodating both crop configurations: 224x224 crops (less context) and 380x380 crops (more context), culminating in six distinct experimental setups:

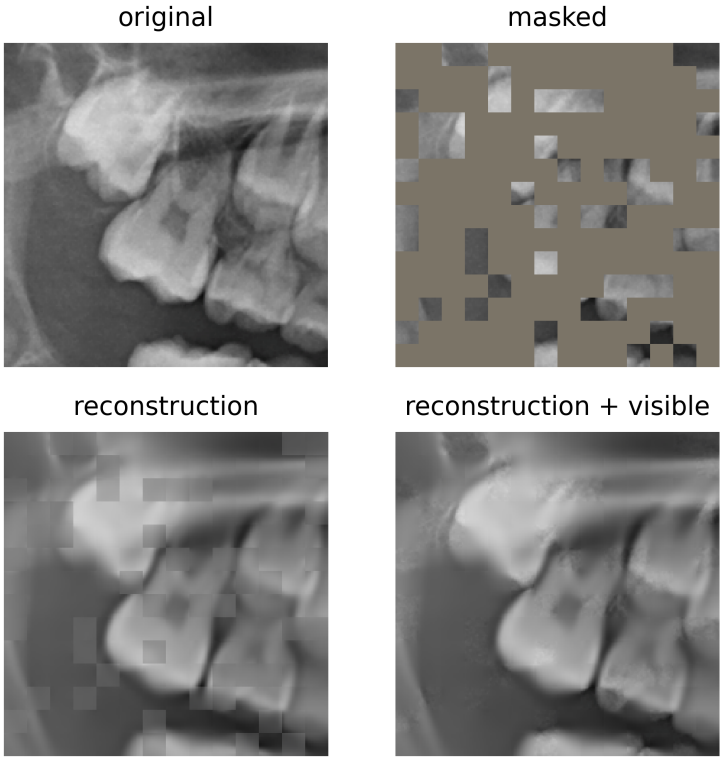


Figure 5.1: Reconstruction sample from a pretrained neural network using MAE as a pretraining strategy. The reconstruction showcases the efficacy of MAE in enhancing the network's reconstructive capabilities.

- Less context crops without pretraining
- Less context crops pretrained on ImageNet dataset
- Less context crops pretrained on Crops dataset
- More context crops without pretraining
- More context crops pretrained on ImageNet dataset
- More context crops pretrained on Crops dataset

In pretraining scenarios, data augmentation techniques used horizontal flip and random resized crop, with scales ranging from 0.2 to 1. The batch size was 512, and the optimizer was AdamW with a learning rate of 9.5×10^{-4}, betas of 0.9 and 0.95, and no weight decay. The network was trained for 800 epochs with a linear warm-up in the first 40 epochs. The hardware used for training was eight NVIDIA A100 of 80 GB. The depicted sample in Fig. 5.1 showcases the considerable qualitative success of the reconstruction outcomes from the pretraining configuration using tooth crops.

5.2.2 Label extraction

In this phase, OpenAI's LLM GPT-4 was used to streamline and expedite the extraction of noun phrases from textual reports of the TRPR dataset. For the current study, the

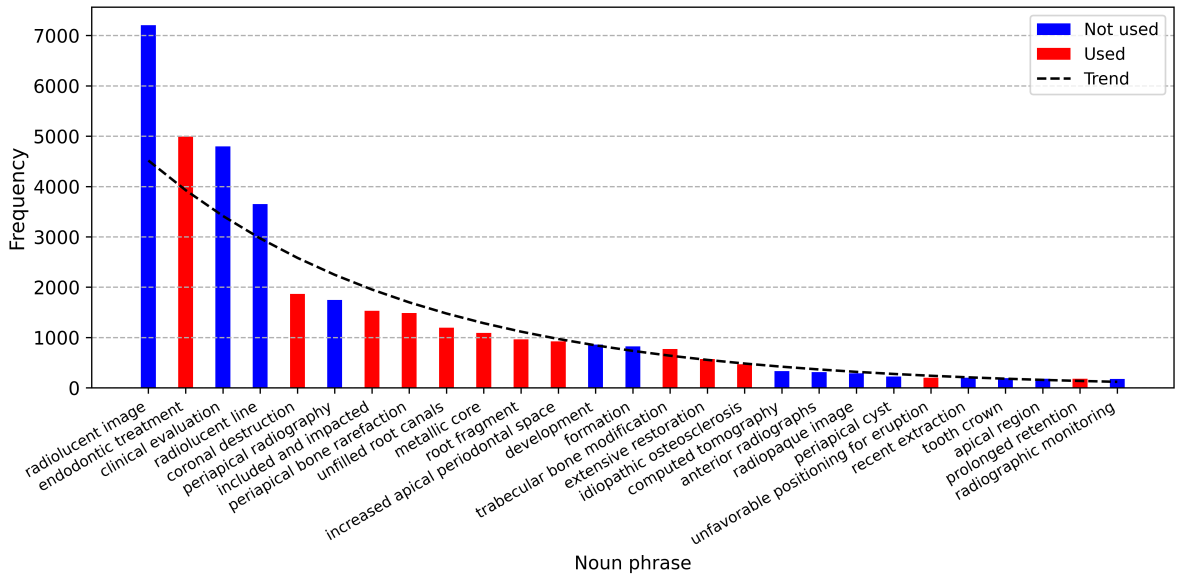


Figure 5.2: Bar chart of the 27 most common noun phrases, showing their frequency and trends, and illustrating their long tail distribution. Noun phrases identified as dental conditions are highlighted with red bars, while those not selected are represented with blue bars.

frequency of all noun phrases was gauged, and only those with occurrences higher than 150 were considered. This threshold was chosen arbitrarily, believing it to represent the minimum necessary for a network to learn effectively. Afterward, similar phrases, such as "unfilled root canal" with its plural form "unfilled root canals", were manually grouped. As not all noun phrases are dental conditions, the selection was refined through manual filtering. For example, "endodontic treatment" is a dental condition, whereas "clinical assessment" is not and, therefore, was excluded from the analysis. Fig. 5.2 displays a bar chart depicting the 27 most common noun phrases, their frequency, and trends, evincing their long tail distribution. Noun phrases identified as dental conditions are marked with red bars, whereas those not selected are shown with blue bars.

In the end, the descriptions of the selected dental conditions, each assigned a unique numerical index and ordered by their frequency of occurrence (indicated in parentheses), were:

1. Endodontic treatment (4,994) - a procedure that treats infections inside the tooth, typically involving the removal of the pulp and nerves, followed by the filling and sealing of the pulp chamber and root canals.

2. Coronal destruction (1,866) - Damage or decay to the crown portion of the tooth.

3. Included and impacted (1,532) - Teeth trapped within the jawbone or gums and cannot erupt naturally.

4. Periapical bone rarefaction (1,486) - A reduction or loss of bone density around the apex of a tooth root, often due to inflammation or infection.

5. Unfilled root canals (1,194) - Root canals that have not been filled or sealed after an endodontic procedure.

6. Metallic core (1,091) - A metal post used to support a restoration or crown, especially in a tooth undergoing endodontic treatment.

7. Root fragment (964) - A piece or portion of a tooth root left behind, typically after tooth extraction or breakage.

8. Increased apical periodontal space (922) - Enlargement of the space around the tooth root's apex, which may indicate an inflammatory response.

9. Trabecular bone modification (773) - Changes in the spongy part of the bone, which can be indicative of disease or other conditions.

10. Extensive restoration (573) - Large dental fillings or excessive material used in a dental restoration.

11. Idiopathic osteosclerosis (470) - A localized increase in bone density without a known cause.

12. Unfavorable positioning for eruption (200) - The positioning of a tooth that hinders its natural eruption process.

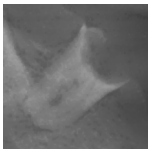
13. Prolonged retention (181) - The extended presence of a tooth or dental element beyond its normal duration, often referring to baby teeth that don't fall out on time.

Fig. 5.3 displays examples of each condition. Upon determining the conditions to be evaluated, the adopted linkage process, which associates every tooth mentioned in a sentence with all the dental conditions stated in that sentence, was applied as described in Section 3.4.3.

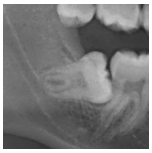
5.2.3 Classification neural network training

The TRPR dataset was split into train (70%), validation (15%), and test (15%) subsets for training and evaluation (see Table 3.3). The tooth crops were 224×224 without resizing (less context), or the 224×224 resized from 380×380 (more context) crops. Data augmentation techniques used were horizontal flip 50% of the time, 10 degrees random rotation, and color jitter with parameters 0.2 for brightness, 0.2 contrast, and 0.2 saturation. The positive classes were oversampled by a factor of 10 due to their insufficient representation. The batch size was 64, and the optimizer was AdamW with a base learning rate of 10^{-3}, betas of 0.9 and 0.95, and no weight decay. The network was trained for 50 epochs with no linear warm-up. The hardware used for training was eight

(a) Endodontic treatment.

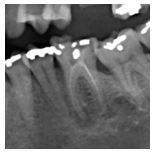


(b) Coronal destruction.



(c) Included and impacted.

(d) Periapical bone rarefaction.

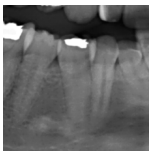


(e) Unfilled root canals.

(f) Metallic core.

(g) Root fragment.

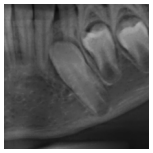
(h) Increased apical periodontal space.



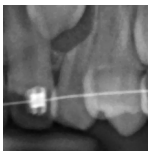
(i) Trabecular bone modification.

(j) Extensive restoration.

(k) Idiopathic osteosclerosis.



(l) Unfavorable positioning for eruption.



(m) Prolonged retention.

Figure 5.3: Dental conditions considered in this study. (They were selected according to their frequency in the textual reports.)

NVIDIA A100 of 80 GB. Finally, the loss was binary cross entropy (BCE) plus MCC loss were calculated:

$$\mathrm{BCE}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N} \left(y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)\right) , \tag{5.1}$$

where N is the number of samples, y is a vector of the target labels, and the $\hat{y}$ is the vector of the predicted probabilities.

The MCC loss is given by $1-\mathrm{MCC}$, where a small value was added on the denominator of Eq. 3.1 to avoid division by 0 (zero).

The final loss is

$$\mathrm{Loss} = \alpha \cdot \mathrm{BCE}(y, \hat{y}) + (1 - \alpha) \cdot (1 - \mathrm{MCC}(y, \hat{y})) . \tag{5.2}$$

Here, $\alpha = 0.5$.

5.2.4 Results and discussions

Table 5.1 showcases the primary numerical MCC results. It presents, for each tooth condition (**Label**), the positive class sample size frequency (**Freq.**), the **Validation** outcomes for each pretraining configuration, the maximum MCC value on validation datasets (**Max Val.**), and the **Test** results. One can conclude from the validation average values that the no-pretraining configurations, indicated in Tables 5.1 and 5.2 by the column **None**, had the worst results both on the less-context and more-context tooth

Table 5.1: Results based on the MCC values from the validation and test sets indicate that pretraining with the ImageNet and Crops dataset was beneficial.

Label	**Freq.**	**Validation**						**Max Val.**	**Test**
		224 × 224 crops (less context)			**380 × 380 crops (more context)**				
		None	**ImageNet**	**Crops**	**None**	**ImageNet**	**Crops**		
1	4,994	0.864	0.903	**0.904**	0.827	0.847	0.846	0.904	0.865
2	1,866	0.421	0.668	**0.714**	0.300	0.663	0.675	0.714	0.658
3	1,532	0.681	0.767	0.740	0.649	0.776	**0.790**	0.790	0.683
4	1,486	0.455	0.589	**0.598**	0.487	0.498	0.523	0.598	0.397
5	1,194	0.264	0.454	0.445	0.455	**0.611**	0.595	0.611	0.436
6	1,091	0.653	0.677	0.695	0.150	0.711	**0.750**	0.750	0.632
7	964	0.318	0.532	0.510	0.167	**0.728**	0.668	0.728	0.583
8	922	0.142	0.275	0.270	0.394	0.399	**0.405**	0.405	0.327
9	773	0.000	0.301	0.309	0.649	**0.506**	0.458	0.649	0.218
10	573	0.000	**0.385**	0.286	0.000	0.284	0.314	0.385	0.252
11	470	0.000	0.182	0.182	0.000	**0.424**	0.414	0.424	0.347
12	200	0.299	0.336	0.420	0.302	0.430	**0.456**	0.456	0.353
13	181	0.240	0.577	**0.666**	0.211	0.386	0.545	0.666	0.426
Average		0.334	0.511	0.519	0.353	0.559	0.572	0.622	0.475

crop scenarios. In contrast, the pretraining from the tooth crop dataset, indicated in the tables by **Crops**, had the best average results in both cases. Pretraining with tooth crop data outperformed **ImageNet** pretraining, on average, by 6.73 percentage points (p.p.) and 1.14 (p.p.) in the less-context and more-context tooth crops, respectively. Despite containing approximately 460,000 images–far fewer than the ImageNet dataset's more than 17 million–pretraining with tooth crop data proved more efficient due to the field-oriented data context. The faster convergence of tooth crop pretraining configurations demonstrates its efficiency. Table 5.2 shows that tooth crop pretraining configurations perform more optimally in fewer epochs than those pretraining with ImageNet.

The test set's results in Table 5.1 were derived from the top-performing network based on the validation sets. Notably, while the test MCC values exhibit considerable variation, they all exceed 0 (zero), indicating performance better than random guessing. Furthermore, according to the positive sample size, the metrics show a noticeable increasing trend. This trend is illustrated in Fig. 5.4, a scatter plot of the data, where the trend was computed using a linear function. The linear function R^2 reached 0.575. R^2 is a statistical measure of how well a mathematical equation represents a set of data. An R^2 between 0.5 and 0.7 indicates a substantial fit, meaning the model reliably explains a significant portion of the variance in the data.

While the size of the positive sample contributes to the MCC trend, it does not account for all of it. A deeper understanding of the challenges in classifying different classes offers more insight into how well the network performs. For instance, some conditions are not in the teeth but around them (*e.g.*, in the gum), requiring more image context. These idiosyncrasies are discussed in the following, indicating in parentheses which configuration performed better, whether with less context or more context in the panoramic.

Table 5.2: Analysis of epoch convergences (values in the table), based on the highest MCC value on the validation sets.

Label	**Freq.**	**Validation**					
		224 × 224 crops (less context)			**380 × 380 crops (more context)**		
		None	**ImageNet**	**Crops**	**None**	**ImageNet**	**Crops**
1	4,994	44	38	11	40	7	7
2	1,866	36	14	23	35	8	13
3	1,532	42	23	26	43	15	30
4	1,486	37	14	13	34	9	10
5	1,194	38	11	11	40	24	11
6	1,091	44	25	21	21	8	11
7	964	32	25	11	34	24	7
8	922	37	3	9	28	6	6
9	773	0	23	6	0	26	17
10	573	0	5	12	18	19	1
11	470	0	16	6	0	6	2
12	200	25	4	19	41	6	3
13	181	38	8	10	27	18	11
Average		29	16	14	28	14	10

1. Endodontic treatment **(less context)**

   An endodontic treatment appears as white (radiopaque) lines in the tooth canals (refer to Fig. 5.3 (a)). Therefore, an image crop close and centered on the teeth eases the task of identifying this condition. This configuration is the case for the 224 × 224 crops. Together with the large amount of positive data, this resulted in a MCC higher than 0.900 on the validation data, while the resize crops from 380 × 380 dimensions reached 0.845 MCC.

2. Coronal destruction **(less context)**

   Coronal destruction appears as darker areas (radiolucencies) because the structure is less dense than a healthy tooth. This decay can be seen as disruptions in the continuous outline of the tooth crown, especially around or underneath existing dental restorations. Therefore, a close, near-the-tooth crop is sufficient for detecting coronal destruction, as depicted in Fig. 5.3 (b). In this case, a maximum of 0.714 was reached from the less-context crops against 0.675 of the more-context one.

3. Included and impacted **(more context)** It refers to a tooth that has not erupted into its expected position in the dental arch due to obstruction by another tooth, bone, or soft tissue (Fig. 5.3 (c)). This phenomenon occurs frequently with wisdom teeth (third molars). Depending on its location, the impacted wisdom tooth can be seen pressing against or tilted towards its neighboring second molar, potentially causing root resorption or displacement. Therefore, a minimal increase in the context of the tooth crop may be beneficial to identify inclusions. The maximum

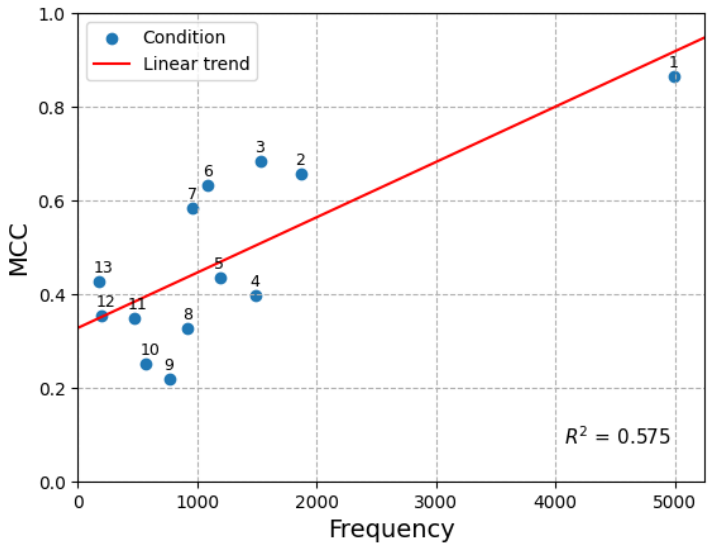


Figure 5.4: Scatter plot of the MCC results on the test set. (The red line shows the MCC increasing trend according to the frequency of the dental condition in the dataset.)

attained results from more-context tooth crops (0.790) were 2.3 (p.p.) higher than the less-context counterpart (0.767).

4. Periapical bone rarefaction **(less context)**

   Periapical bone rarefaction appears as a darker area (radiolucent) around the tooth's root or at its apex (see Fig. 5.3 (d)). This dark spot indicates bone loss or decreased bone density. The borders of this area can be well-defined or more diffuse, depending on the nature and stage of the condition. Under these circumstances, having a well-focused image around the teeth is better, or even necessary, to better diagnose this condition. Here, the less context reached 0.599 against 0.523.

5. Unfilled root canals **(more context)**

   On a panoramic radiograph, unfilled root canals within a tooth appear as relatively dark lines or canals within the lighter, radiopaque outline of the tooth structure (refer Fig. 5.3 (e)). These dark lines represent where the dental pulp once was and should generally be filled with endodontic materials if a root canal treatment has been completed. In those conditions, a well-focused image is better for classification. In the current setup, the networks trained with more-context crops attained an MCC of 0.611 and the less-context crops of 0.454.

6. Metallic core **(more context)**

   On a panoramic radiograph, a "metallic core" within a tooth appears as a highly radiopaque area within the tooth structure, often in the shape of a post or a dense filling (see Fig. 5.3 (f)). It stands out distinctly against the less dense surrounding tooth material and any dental restorations that are not metal-based. However, since

it is not a part of the tooth, the crop center excludes the metallic core "crown". More context is expected to help the classification in light of this situation. A value of 0.75 for MCC in the more-context configuration and 0.695 in the less-context was reached.

7. Root fragment **(more context)**

   On a panoramic radiograph, a "root fragment" appears as a radiopaque structure, resembling a part of a tooth's root, as shown in Fig. 5.3 (g). It is usually isolated, without a crown portion, and may be surrounded by a darker area if inflammation or bone resorption is present. The need to screen the tooth's surroundings makes having more context in the image important. Results of 0.728 and 0.532 were achieved in less-context and more-context scenarios, respectively.

8. Increased apical periodontal space **(more context)**

   On a panoramic radiograph, an increased apical periodontal space, displayed in Fig. 5.3 (h), appears as an enhanced or widened radiolucent area around the tip of the root of a tooth. This dark gap, known as the periodontal ligament space, is usually uniform and thin around the roots of healthy teeth. Bearing this in mind, a closed, well-focused image around the tooth's center may exclude the dental condition, making its diagnosis impossible. Therefore, an image with more context is more beneficial for detecting an increased apical periodontal space. The current study reached a 0.405 MCC in the more context scenario against a 0.275 in the less context (i.e., a performance boost of 47.27%).

9. Trabecular bone modification **(more context)**

   On a panoramic radiograph, a "trabecular bone modification" may appear as changes in the pattern and density of the bone, as shown in Fig. 5.3 (i). Areas with increased density will look whiter, indicating a more solid bone structure, while regions with decreased density will appear darker, suggesting less bone mass. The regular mesh-like pattern of the trabeculae might appear disrupted or altered, which can indicate various dental or bone conditions. These areas appear on the bones surrounding the teeth, not near their center. Therefore, a crop with more context is beneficial for diagnosing trabecular bone modification. In the more-context scenario, we attained 0.506 of MCC; in the less context, we attained 0.309 on the validation datasets.

10. Extensive restoration **(less context)**

    On a panoramic radiograph, "extensive restoration", or "excess restorative material", appears as a filling, crown, or other dental work, that extends beyond the natural contours of the tooth, as displayed in Fig. 5.3 (j). Typically used for fillings or crowns, these materials will stand out as they are denser than the tooth and absorb more X-rays. If overfilled, excess material may also be seen beyond the confines of the tooth's normal borders, such as in the interdental spaces or the pulp chamber. Restorative material frequently occurs on the tooth crown, a distance

from the tooth center. Therefore, cropping too closely may hinder accurate diagnosis. In the current analysis, the images were not excessively cropped, which allowed for a correct diagnosis with less contextual information. Specifically, an MCC of 0.385 was observed in scenarios with less context compared to 0.314 in scenarios with more context.

11. Idiopathic osteosclerosis **(more context)**

    On a dental radiograph, "idiopathic osteosclerosis" appears as a brighter area due to increased bone density (Fig. 5.3 (k)). It is often seen near the roots of teeth but lacks the characteristic dark border of other lesions. Therefore, the view of the tooth's surroundings could be beneficial for detecting it. The results were 0.424 in the more-context scenario and 0.182 in the less-context one.

12. Unfavorable positioning for eruption **(more context)**

    An "unfavorable positioning for eruption" for a tooth appears as a tooth that is misaligned with the normal arch form, often at an abnormal angle or location that suggests it will not erupt into a functional position without intervention (see Fig. 5.3 (m)). This could be a tooth that is tilted, rotated, or horizontally displaced. The context around the tooth is important to verify and confirm if its position is unfavorable for eruption. Indeed, the results in the more-context scenario were 0.456 against 0.420 in the less-context scenario (an increase of 8.57%).

13. Prolonged retention **(less context)**

    "Prolonged retention" of a tooth is indicated by a tooth that remains in the jaw beyond the typical age of exfoliation without evidence of natural shedding or eruption (Fig. 5.3 (n)). It often appears as a tooth with roots that may be resorbed, situated in the jaw without movement, potentially affecting the positioning of adjacent teeth or the eruption of successor permanent teeth. These deciduous teeth are small and do not require a larger context for diagnosis of prolonged retention. A value of 0.666 was reached in the context scenario against 0.545 in the more context one.

5.3 COMPARISON WITH DENTISTRY PROFESSIONALS

The performance evaluation on a large and diverse dataset indicates that the proposed framework has learned, as all MCCs exceeded 0 (zero). However, these values do not provide a desirable comparison to the performance of human professionals when evaluating panoramic radiographs. To make this comparison, annotations made by dentistry professionals were assessed and compared against the results of the classification models' predictions. This was accomplished by inviting five final-year undergraduate students (junior annotators) and five radiologist experts (senior annotators) to label some samples of the same test images used to evaluate the classification models.

In the labeling setup, each participant had to label a cropped panoramic radiograph centered on a specific tooth, similar to the images used for training and evaluating the

models. The images in the setup were the same as the "more-context" 380 × 380 crops, but without resizing to 224 × 224, which was previously necessary to meet the model's input requirements. The dentistry professionals had to identify and mark all visible dental conditions in the area of the central tooth in the crop, based on the provided options (all the conditions considered in this study), or mark none, according to their analysis.

A hurdle that needed to be overcome in this procedure was the number of test images to be annotated. According to the estimates, there were about 32,000 test images (please refer to Table 3.4), which would require more than 300 hours of continuous work for each participant to label, making it impractical. A natural alternative was to sample a subset of the test set while maintaining the positive/negative class proportions. Unfortunately, this option also proved unworkable due to the highly imbalanced datasets. For instance, the positive/negative ratio for condition 13 (prolonged retention) is 0.106. In this case, the professionals would need to annotate approximately 1,180 samples to maintain the proportional ratio, with only one being positive. Annotating such a large number of image crops was beyond reasonable feasibility. The issue was overcome by selecting images through a strategy that ensured a minimum number of positive examples and variability. The adopted strategy consisted of selecting 78 samples (six images per condition) using the following pattern: for each condition, two true positives (TP), two false positives (FP), and two false negatives (FN) were selected based on the original models' predictions. According to the labels extracted from the reports, this approach ensured at least four positive samples (the two TPs and the two FNs) and two negative samples (FP). It also provided potential variability due to the FP and FN typically being borderline cases. This final set of images is designated as **Expert Image Dataset**.

5.3.1 Initial assessment

Table 5.3 shows the results for each professional and the average results for the students and experts, considering the labels from the text reports as the ground truth. The outcomes indicate that the expert group performed moderately better than the students (0.429 vs. 0.455 MCC). The attained MCC by the used models was 0.475 (Table 5.1), which is higher than the scores of both groups, demonstrating strong performance. However, one cannot assert that the models have reached superhuman performance because the scenario is biased in favor of the models. Rather than being trained to detect dental conditions in general, the models were trained to detect conditions as the primary labeler (reports), giving them an advantage over professionals who did not have access to the annotator samples. This issue was mitigated by combining the expert labels, as discussed below.

5.3.2 Definitive assessment with expert consensus

The original training, validation, and testing labels were derived from textual reports. Under these conditions, the models trained and validated on these datasets had an advantage over professionals, as with the proposed solution, when benchmarked. To mitigate this bias in the models' performance, the labels provided by the professionals were leveraged to create a new ground truth.

Table 5.3: Average MCC for each student and expert (the average value for each group is also included.)

Student	**Avg. MCC**	**Expert**	**Avg. MCC**
Student 1	0.479	Expert 1	0.394
Student 2	0.500	Expert 2	0.589
Student 3	0.360	Expert 3	0.387
Student 4	0.435	Expert 4	0.455
Student 5	0.372	Expert 5	0.452
All Students	**0.429**	**All Experts**	**0.455**

Table 5.4: Final average MCC of all conditions for each student and expert, including the average value for each group.

Student	**Avg. MCC**	**Expert**	**Avg. MCC**
Student 1	0.527	Expert 1	0.607
Student 2	0.591	Expert 2	0.689
Student 3	0.490	Expert 3	0.499
Student 4	0.516	Expert 4	0.575
Student 5	0.426	Expert 5	0.574
All Students	**0.510**	**All Experts**	**0.589**

It was assumed that the expert could generate the most accurate labels. Therefore, it was decided that, for the proposed solution and the students, the new ground truth for the Expert Image Dataset would be generated by combining the labels from all experts. This setup not only avoided favoring the proposed solution but also increased robustness and reduced labeling noise, as a majority vote of the annotators created the new labels. For the experts, a leave-one-out layout was built where, for each of the five rounds, the ground truth was computed from the labels of four experts, and the remaining specialist was evaluated against this new ground truth.

Under this new layout, the results for the models and the average results for the students and experts were 50.8%, 51.0%, and 58.9%, respectively. The detailed results for each participant are included in Table 5.4. It is possible to draw two conclusions from these numbers. Firstly, the MCC values are considerably higher than those attained on the reports' ground truth. This increase can be attributed to the more robust ground truths that were less noisy, as the combination of the labels of several experts generated them. Furthermore, the similarity between the MCCs attained by the proposed solution and the students (50.82% vs 51.00%) led one to conclude that the proposed solution reached the level of a junior professional.

Fig. 5.5 depicts a bar chart that further investigates the results, breaking them down by classes and predictors (models, professionals, students, and experts). One can observe that the proposed solution, when compared to the professionals, demonstrates significantly higher performance in Condition 5 (unfilled root canals) and considerably worse performance in Condition 12 (unfavorable positioning for eruption) and 13 (prolonged

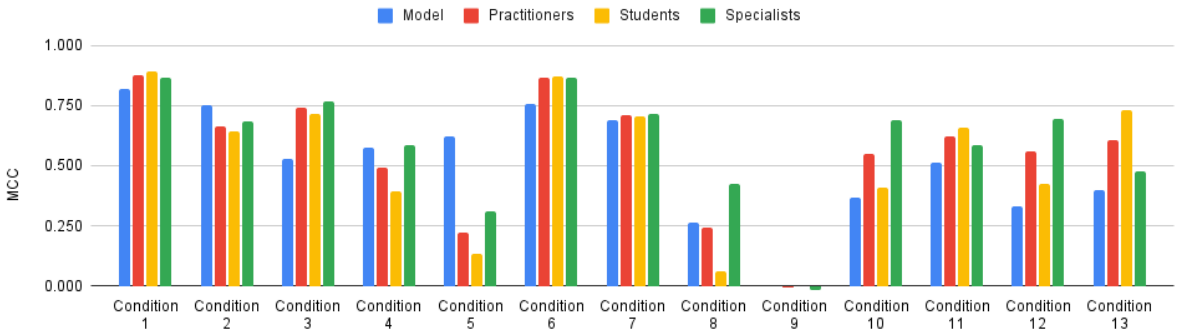


Figure 5.5: Bar chart that breaks down the definitive assessment results according to condition and professional group. The results on Condition 9, trabecular bone modification were closer to zero.

Table 5.5: Frequency of positive samples, Fleiss' Kappa and model's MCC on the definite evaluation set for each condition dental condition.

Condition	1	2	3	4	5	6	7	8	9	10	11	12	13
Frequency	24	10	4	7	6	10	5	5	6	5	6	4	4
Kappa	0.776	0.506	0.601	0.479	0.234	0.750	0.700	0.256	0.045	0.389	0.485	0.336	0.468
MCC	0.819	0.749	0.526	0.577	0.620	0.755	0.688	0.264	0.000	0.369	0.514	0.397	0.330

retention), the ones of less positive samples. However, what stands out the most is the almost null MCC values for condition 9 (Trabecular bone modification) class. One can hypothesize that this result stemmed from a lack of agreement among this class's experts. It can be expected that the higher the agreement between the labelers, the higher the MCC of the model is attained. Consequently, it was decided to conduct a statistical agreement analysis on the ground truth labels.

5.3.3 Statistical agreement analysis

The statistical agreement analysis aims to evaluate the consistency among the ground truth labelers, who, here, are the experts. Fleiss' Kappa is a statistical measure used to assess the reliability of agreement between multiple raters for categorical items (FLEISS, 1971). Fleiss' Kappa was employed to evaluate the consistency of diagnostic decisions made by multiple experts on dental conditions

Table 5.5 contains the frequency of positive samples for each condition, the attained MCC values of the models, and the computed Fleiss' kappa values in the Expert Image Dataset. An inspection of the table's data shows that the hypothesis made holds true: the lack of agreement between the experts on condition 9 (Kappa of 0.045) was connected to the poor performance of all groups (model, students, and experts) on the same condition (MCC of 0 (zero)), indicating that the models struggled to learn from inconsistent labels. Table 5.5 also indicates substantial agreement among the labelers for conditions 1 and 6 (Kappas of 0.776 and 0.750), which are the conditions where the models attained their best results. These results suggest a correlation between the models' performance and the level of agreement among the labelers.

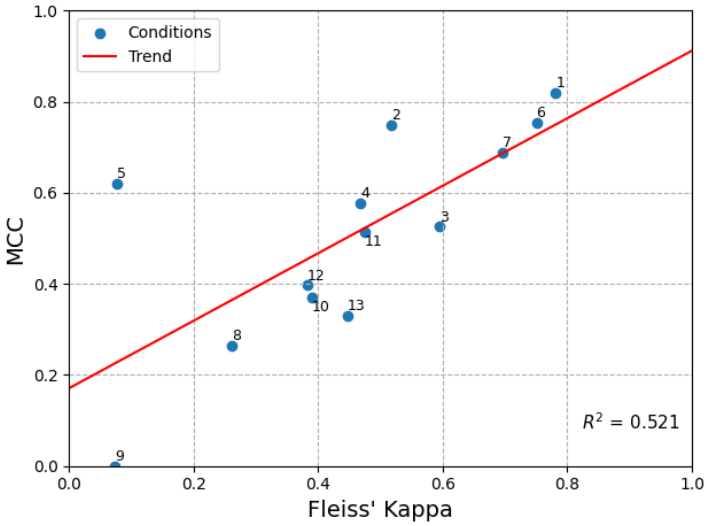
Fig. 5.6 (a) is a scatter plot of the attained MCC results against the computed kappas. The value of R^2 reached 0.521. Similar to the correlation between the Frequency of Positive Samples and MCC of Fig. 5.4, this correlation was not expected to be flawless. Instead, it was aimed to demonstrate a trend—specifically, an increasing one—where a higher kappa corresponds to a higher MCC of the proposed solution.

The correlation between the independent variables (kappa values and positive sample frequency) and the dependent variable (attained MCC) was further investigated. Fig. 5.6 illustrates the final result. With these two independent variables, R^2 reaches 0.769, a value considered a good fit.

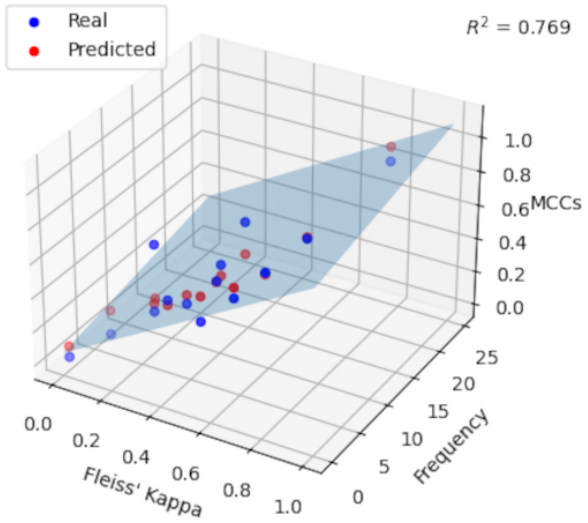
In summary, the performed statistical agreement analysis supports the hypothesis that higher inter-rater agreement leads to better model performance, as measured by MCC. The number of positive samples also has an increasing impact on the values of MCC results. The results underscore the importance of achieving consensus among labelers to improve the reliability of ground-truth data and, consequently, the performance of predictive models.

5.4 CLOSURE

This chapter outlines the procedures for conducting the dental classification experiments and their evaluation. The process consisted of neural network pretraining, label extraction, selection of thirteen conditions, neural network training, comparison with dental professionals, and a final assessment by expert consensus. The extensive experiments and evaluations allowed us to draw several conclusions. A key finding was that our system reached the proficiency level of a junior practitioner, leading us to conclude that our framework was successful.



(a) Scatter plot of MCC results for each condition against Fleiss' Kappa, showing an increasing trend.



(b) Scatter plot of the MCC results for each condition against Fleiss' Kappa and the frequency of positive samples in the dataset. (The blue dots represent the actual values, and the red dots represent the predicted values according to the fitted linear function.)

Figure 5.6: Plots showing the linear trends of MCC results based on Fleiss' Kappa and the frequency of positive samples for each condition.

Chapter 6

CONCLUSION

6.1 STRENGHS AND CONCLUDING REMARKS

Computer vision studies, including those on panoramic radiographs, have relied primarily on supervised learning, but this approach is becoming impractical due to its heavy dependence on labeled data. The time required for data annotation can exceed 80% of the project's duration, hindering its scalability. This limitation underscores the need to explore other learning paradigms, such as semi-supervised and self-supervised learning.

In this work, we used semi-supervised learning to construct a large labeled data set of dental panoramic radiographs: the OdontoAI Open Panoramic Radiographs (O^2PR). The O^2PR comprises 4000 images, in which the teeth were segmented and numbered, and is four times larger than the previously most extensive data set on the matter available in the literature (PANETTA et al., 2021). The magnitude of the O^2PR data set was reached by using the Human-In-The-Loop (HITL) concept to speed up the labeling process. Our results indicated about 51% of labeling time reduction, even instructing our annotators to attend to tiny segmentation errors. We estimate having saved at least 390 continuous working hours. In practice, this number is even higher, as manual labeling is more human-demanding. The HITL annotation verification process is less burdensome, as confirming the labels through visual inspection (rather than correcting them through mouse clicks and the point drag-and-drop feature) corresponds to a significant fraction of the verification process.

The performance of the trained networks on distinct data (validation data, HITL data, and, most important, on a separated manually labeled test dataset) were consistent, showing an increasing trend in the considered metrics over the HITL iterations. HTC 4's segmentation mAP was +5.4 percentage points higher than the HTC 1's on the test data set. For comparison, the performance gain from the standard Mask R-CNN choice to the HTC, the winner architecture of our benchmark (Section 4.2.2), was +4.4 in terms of segmentation mAP. The work by Pinheiro et al. (2021) boosted the segmentation mAP of the Mask R-CNN architecture in +2 percentage points by replacing the original FCN segmentation head for a PointRend module. These results reinforces a common, though frequently ignored, knowledge in the deep learning field: it is often better to gather data

than expending much time in refining a model. For the purpose of enlarging the labeled data sets, the HITL concept is very beneficial.

The less refined segmentation, mainly over the tooth crowns, was the major bottleneck for faster labeling. The segmentation of deep learning solutions slightly differs from the human annotators, overall on the object's fine-detailed borders. If the application allows neglecting these errors, the labeling speed up can increase substantially. Therefore, we conclude that the HITL benefits for instance segmentation applications might vary significantly according to the application due to today's state of deep learning. Currently, the HITL use is more beneficial to applications that do not demand greater segmentation accuracy than those that demand. In this work, we did not neglect the tiny segmentation errors, as we want our data set to be general-purpose.

The use of HITL reduces the human burden on the labeling data process. However, it does not completely shift the learning paradigm from supervised learning as it requires human intervation and, in the end, all the available data will be used for training the machine learning models. This approach differs entirely from self-supervised learning, as all the data employed can be unlabeled, which was fundamental to our investigations into the classification of dental conditions. Our investigations into dental classification began with the construction of a dataset, including the preprocessing of textual reports. In the second stage, we created tooth crops using human annotations and pseudolabels generated by instance segmentation neural networks. However, we lacked condition labels for more than 50% of the teeth, as they did not have corresponding reports. Without employing Masked Autoencoders (MAE), a self-supervised technique, these data would have been wasted, underscoring the importance of adopting different paradigms.

We developed a robust framework leveraging the aforementioned paradigms and validated it across 13 dental conditions through a comprehensive experimental analysis to assess the performance of different setups. The Matthews Correlation Coefficient (MCC) was used as the evaluation metric, measuring whether the model's performance surpassed random guessing—an outcome achieved for all conditions. The experiments were conducted with crops of varying context, demonstrating that context significantly impacts the final result.

The noun phrase extraction through GPT-4 Large Language Model proved to be an effective strategy by not only expediting the labeling process but also identifying the primary classes with minimal human intervention. Our strategy was thoroughly analyzed and rigorously tested. The proposed solution's lower performance in certain conditions led to an investigation of the impact of inter-rater reliability. It was discovered that 52.1% of model performance, as measured by MCC, correlated linearly with Fleiss' kappa. This relationship highlights the critical role of expert consensus, as higher kappa values were associated with higher MCC values. Furthermore, the combination of kappa with the frequency of positive examples results in $R^2 = 0.769$, suggesting that more extensive and consistently labeled datasets could significantly boost performance.

In conclusion, this work covered all the stages of a machine learning project: problem definition, data collection, exploratory data analysis, data labeling, dataset construction (including techniques to accelerate this step), data preprocessing, model selection, neural network pre-training, model training, hyperparameter tuning, ablation studies, evaluation

with metrics, and analysis with human comparison. The findings emphasize the need for comprehensive datasets and consistent annotations to improve model accuracy. Moreover, exploring alternative learning paradigms can help overcome the limitations of supervised learning, paving the way for more robust and reliable dental diagnostics.

6.2 SHORTCOMINGS

We can point out some limitations of our work. For instance, despite its relative scale, the dataset remains limited for rare conditions, which affected model performance on underrepresented categories. The importance of datasets in machine learning cannot be overstated, particularly in medical imaging, where classes are often highly imbalanced. Although the current study used the largest dataset in the literature, of the selected conditions had only 181 samples out of over 200,000 images, representing just 0.11% of the total cropped tooth images. This highlights the need for even larger datasets to improve generalizability.

A notable limitation of this study is the reliance on radiographic data obtained from a single machine model. While this ensured consistency in image quality and parameters, it may have introduced a bias, limiting the generalizability of the findings to radiographs produced by other equipment or configurations. Dental radiographs can vary significantly across different machines due to variations in resolution, exposure settings, and sensor technologies, which might affect the performance of machine learning models. As a result, the developed models could face challenges when applied to radiographs from diverse sources, potentially requiring further fine-tuning or retraining to adapt to new imaging environments.

Another limitation of this study is that the textual report from the training data was labeled by only two radiologists, and the labeling process relied on a heuristic approach. While the radiologists provided expert insights, the limited number of annotators may introduce bias in the labels, possibly reducing the model's classification performance and generalizability. Furthermore, the use of heuristics to guide label extraction, though practical, may not fully capture certain dental conditions, which could further impact the model's ability to perform accurately in diverse clinical scenarios.

The challenge of evenly sampling the dataset from the broader population is another limitation. To address this, we utilized data collected directly from a dental clinic, allowing us to replicate patient distribution patterns and closely mirror the local population. However, focusing on a local population excludes ethnicities from regions farther away, possibly limiting the generalizability of the findings. Therefore, we must acknowledge that our results are biased toward the ethnicities represented in the patient population from which the data were collected.

6.3 APPLICATIONS

The contributions of this study offer several practical applications in both research and clinical settings. The O^2PR dataset, with its large-scale annotated radiographs, serves as a valuable resource for advancing computer vision applications in dental imaging. Given

the importance of teeth in radiographic analysis, researchers can leverage this dataset to benchmark new algorithms for tasks such as tooth segmentation, numbering, and condition classification, thereby promoting further exploration into the automation of dental diagnostics. Additionally, the use of HITL demonstrates the feasibility of creating new instance segmentation datasets efficiently, paving the way for larger and more diverse datasets. This methodology can be extended to other medical imaging fields, where instance segmentation plays a crucial role but manual annotation remains time-consuming and labor-intensive.

In clinical practice, the models developed in this study provide the foundation for automated tools capable of assisting radiologists in diagnosing dental conditions from panoramic radiographs. The use of semi-supervised learning, combined with self-supervised paradigms, allows these models to adapt more effectively to limited or incomplete data, addressing a common challenge in medical imaging. Moreover, the strategy of integrating GPT-4 for automatic label extraction offers a new avenue for enhancing labeling efficiency in medical datasets. This approach can be applied to other domains involving large-scale text and image data, facilitating the development of robust machine learning systems with minimal human intervention.

Finally, the framework and methodologies developed in this research can be extended to other areas within the healthcare sector, fostering innovation in automated diagnostics. This paves the way for future applications of AI in dentistry, offering the potential for earlier detection of conditions, streamlined workflows, and improved patient outcomes.

6.4 FUTURE WORK

We believe that our work opens numerous avenues for future research. One promising direction is the expansion of the instance segmentation dataset, both in size and scope, by including additional objects of interest such as implants, prostheses, and jaw structures. Increasing the diversity of annotated objects, as well as incorporating patients from different ethnicities, will enable the development of more comprehensive models and better reflect real-world clinical scenarios. Additionally, the inclusion of new public datasets of dental panoramic radiographs from various devices, even if limited in size, would be invaluable for evaluating the generalization ability of these techniques across diverse imaging conditions.

Future research should also focus on refining segmentation and tooth numbering techniques by considering global anatomical context and geometric relationships between dental structures. Aditionally, the results showed that crops containing more context yielded better outcomes, even though these sizes were chosen empirically. Future research should aim to systematically determine crop sizes and increase the number of positive samples to enhance performance.

Finally, it is worth noting that this work utilized a version of GPT-4 with a simple prompt. Future research should focus on refining prompts tailored for different large language models (LLMs). Additionally, developing advanced prompts capable of directly extracting dental conditions without relying on heuristics would be highly beneficial, enhancing both efficiency and accuracy in diagnostic tasks.

BIBLIOGRAPHY

AMASYA, H. et al. Development and validation of an artificial intelligence software for periodontal bone loss in panoramic imaging. *International Journal of Imaging Systems and Technology*, Wiley Online Library, v. 34, n. 1, p. e22973, 2024.

BONFANTI-GRIS, M. et al. Evaluation of an artificial intelligence web-based software to detect and classify dental structures and treatments in panoramic radiographs. *Journal of Dentistry*, Elsevier, v. 126, p. 104301, 2022.

BROOKS, J. *COCO Annotator*. 2019. <https://github.com/jsbroks/coco-annotator/>.

CAI, Z.; VASCONCELOS, N. Cascade r-cnn: Delving into high quality object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2018. p. 6154–6162.

CAI, Z.; VASCONCELOS, N. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, 2019.

CHANG, H.-J. et al. Deep learning hybrid method to automatically diagnose periodontal bone loss and stage periodontitis. *Scientific reports*, Nature Publishing Group UK London, v. 10, n. 1, p. 7531, 2020.

CHEN, H. et al. Dental disease detection on periapical radiographs based on deep convolutional neural networks. *International Journal of Computer Assisted Radiology and Surgery*, Springer, v. 16, p. 649–661, 2021.

CHEN, K. et al. Hybrid task cascade for instance segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2019. p. 4974–4983.

CHEN, K. et al. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

CHEN, L.-C. et al. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

CHEN, Q. et al. Mslpnet: multi-scale location perception network for dental panoramic x-ray image segmentation. *Neural Computing and Applications*, Springer, p. 1–15, 2021.

CHUNG, M. et al. Individual tooth detection and identification from dental panoramic x-ray images via point-wise localization and distance regularization. *arXiv preprint arXiv:2004.05543*, 2020.

CHUNG, M. et al. Individual tooth detection and identification from dental panoramic x-ray images via point-wise localization and distance regularization. *Artificial Intelligence in Medicine*, Elsevier, v. 111, p. 101996, 2021.

CORDTS, M. et al. The cityscapes dataset for semantic urban scene understanding. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 3213–3223.

CUI, W. et al. Toothpix: Pixel-level tooth segmentation in panoramic x-ray images based on generative adversarial networks. In: IEEE. *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. [S.l.], 2021. p. 1346–1350.

DAI, J. et al. Deformable convolutional networks. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2017. p. 764–773.

DOUGLAS, D. H.; PEUCKER, T. K. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization*, University of Toronto Press, v. 10, n. 2, p. 112–122, 1973.

EKERT, T. et al. Deep learning for the radiographic detection of apical lesions. *Journal of endodontics*, Elsevier, v. 45, n. 7, p. 917–922, 2019.

FLEISS, J. L. Measuring nominal scale agreement among many raters. *Psychological bulletin*, American Psychological Association, v. 76, n. 5, p. 378, 1971.

FUKUDA, M. et al. Evaluation of an artificial intelligence system for detecting vertical root fracture on panoramic radiography. *Oral Radiology*, Springer, v. 36, p. 337–343, 2020.

GAO, L. et al. Ai-aided diagnosis of oral x-ray images of periapical films based on deep learning. *Displays*, Elsevier, v. 82, p. 102649, 2024.

HE, K. et al. Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2022. p. 16000–16009.

HE, K. et al. Mask r-cnn. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2017. p. 2961–2969.

HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 770–778.

HSU, T.-M.; WANG, Y.-C. Deepopg: Improving orthopantomogram finding summarization with weak supervision. *arXiv preprint arXiv:2103.08290*, 2021.

JADER, G. et al. Deep instance segmentation of teeth in panoramic x-ray images. In: IEEE. *Conference on Graphics, Patterns and Images*. [S.l.], 2018. p. 400–407.

JING, B.; XIE, P.; XING, E. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*, 2017.

KHAN, H. A. et al. Automated feature detection in dental periapical radiographs by using deep learning. *Oral surgery, oral medicine, oral pathology and oral radiology*, Elsevier, v. 131, n. 6, p. 711–720, 2021.

KHAN, S. et al. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, ACM New York, NY, v. 54, n. 10s, p. 1–41, 2022.

KIRILLOV, A. et al. Pointrend: Image segmentation as rendering. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2020. p. 9799–9808.

KOCH, T. et al. Accurate segmentation of dental panoramic radiographs with u-nets. In: IEEE. *International Symposium on Biomedical Imaging*. [S.l.], 2019. p. 15–19.

KROIS, J.; SCHNEIDER, L.; SCHWENDICKE, F. Impact of image context on deep learning for classification of teeth on radiographs. *Journal of clinical medicine*, Multidisciplinary Digital Publishing Institute, v. 10, n. 8, p. 1635, 2021.

KWON, O. et al. Automatic diagnosis for cysts and tumors of both jaws on panoramic radiographs using a deep convolution neural network. *Dentomaxillofacial Radiology*, The British Institute of Radiology., v. 49, n. 8, p. 20200185, 2020.

LANGLAIS, R. P.; MILLER, C. *Exercises in Oral Radiology and Interpretation-E-Book: Exercises in Oral Radiology and Interpretation-E-Book*. [S.l.]: Elsevier Health Sciences, 2016.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *nature*, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015.

LEE, J.-H.; KIM, D.-H.; JEONG, S.-N. Diagnosis of cystic lesions using panoramic and cone beam computed tomographic images based on deep learning neural network. *Oral diseases*, Wiley Online Library, v. 26, n. 1, p. 152–158, 2020.

LEITE, A. F. et al. Artificial intelligence-driven novel tool for tooth detection and segmentation on panoramic radiographs. *Clinical oral investigations*, Springer, v. 25, n. 4, p. 2257–2267, 2021.

LIN, T.-Y. et al. Feature pyramid networks for object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2017. p. 2117–2125.

LIU, F. et al. Recognition of digital dental x-ray images using a convolutional neural network. *Journal of Digital Imaging*, Springer, v. 36, n. 1, p. 73–79, 2023.

LIU, S. et al. Path aggregation network for instance segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2018. p. 8759–8768.

LONG, J.; SHELHAMER, E.; DARRELL, T. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 3431–3440.

MATTHEWS, B. W. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, Elsevier, v. 405, n. 2, p. 442–451, 1975.

MENZE, M.; GEIGER, A. Object scene flow for autonomous vehicles. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2015. p. 3061–3070.

OLIVEIRA, H. N.; FERREIRA, E.; SANTOS, J. A. D. Truly generalizable radiograph segmentation with conditional domain adaptation. *IEEE Access*, IEEE, v. 8, p. 84037–84062, 2020.

PANETTA, K. et al. Tufts dental database: A multimodal panoramic x-ray dataset for benchmarking diagnostic systems. *IEEE Journal of Biomedical and Health Informatics*, IEEE, 2021.

PINHEIRO, L. et al. Numbering permanent and deciduous teeth via deep instance segmentation in panoramic x-rays. In: SPIE. *Symposium on Medical Information Processing and Analysis (SIPAIM)*. [S.l.], 2021. p. 95 – 104.

QIAO, S.; CHEN, L.-C.; YUILLE, A. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2021. p. 10213–10224.

RANJBAR, F. R.; ZAMANIFAR, A. Autonomous dental treatment planning on panoramic x-ray using deep learning based object detection algorithm. *Multimedia Tools and Applications*, Springer, p. 1–35, 2023.

REN, S. et al. Faster r-cnn: Towards real-time object detection with region proposal networks. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2015. p. 91–99.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: SPRINGER. *International Conference on Medical image computing and computer-assisted intervention*. [S.l.], 2015. p. 234–241.

SCHWENDICKE, F. a.; SAMEK, W.; KROIS, J. Artificial intelligence in dentistry: chances and challenges. *Journal of dental research*, SAGE Publications Sage CA: Los Angeles, CA, v. 99, n. 7, p. 769–774, 2020.

SILVA, B. et al. Dental image analysis: Where deep learning meets dentistry. In: *Convolutional Neural Networks for Medical Image Processing Applications*. [S.l.]: CRC Press, 2022. p. 170–195.

SILVA, B. et al. A study on tooth segmentation and numbering using end-to-end deep neural networks. In: IEEE. *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. [S.l.], 2020. p. 164–171.

SILVA, B. P. M. et al. Boosting research on dental panoramic radiographs: a challenging data set, baselines, and a task central online platform for benchmark. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, Taylor & Francis, p. 1–21, 2023.

SILVA, G.; OLIVEIRA, L.; PITHON, M. Automatic segmenting teeth in x-ray images: Trends, a novel data set, benchmarking and future perspectives. *Expert Systems with Applications*, v. 107, p. 15–31, 2018.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

TAO, A.; BARKER, J.; SARATHY, S. *DetectNet: Deep Neural Network for Object Detection in DIGITS*. 2016. Disponível em: <https://developer.nvidia.com/blog/detectnet-deep-neural-network-object-detection-digits/>.

TASSOKER, M.; ÖZIÇ, M. Ü.; YUCE, F. Performance evaluation of a deep learning model for automatic detection and localization of idiopathic osteosclerosis on dental panoramic radiographs. *Scientific Reports*, Nature Publishing Group UK London, v. 14, n. 1, p. 4437, 2024.

TONETTI, M. S. et al. Impact of the global burden of periodontal diseases on health, nutrition and wellbeing of mankind: A call for global action. *Journal of clinical periodontology*, Wiley Online Library, v. 44, n. 5, p. 456–462, 2017.

TUZOFF, D. et al. Tooth detection and numbering in panoramic radiographs using convolutional neural networks. *Dentomaxillofacial Radiology*, v. 48, n. 4, 2019.

VINAYAHALINGAM, S. et al. Automated chart filing on panoramic radiographs using deep learning. *Journal of Dentistry*, Elsevier, v. 115, p. 103864, 2021.

WANG, A. et al. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

WHITE, S. C.; PHAROAH, M. J. *Oral radiology-E-Book: Principles and interpretation*. [S.l.]: Elsevier Health Sciences, 2014.

WHO, W. H. O. *Global oral health status report*. 2022. Disponível em: <https://www.who.int/team/noncommunicable-diseases/global-status-report-on-oral-health-2022>.

Wikipedia contributors. *Panoramic radiograph — Wikipedia, The Free Encyclopedia*. 2024. [Online; accessed 11-October-2024]. Disponível em: <https://en.wikipedia.org/w/index.php?title=Panoramic_radiograph&oldid=1249523277>.

WU, X. et al. A survey of human-in-the-loop for machine learning. *arXiv preprint arXiv:2108.00941*, 2021.

XIE, S. et al. Aggregated residual transformations for deep neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2017. p. 1492–1500.

YÜKSEL, A. E. et al. Dental enumeration and multiple treatment detection on panoramic x-rays using deep learning. *Scientific reports*, Nature Publishing Group UK London, v. 11, n. 1, p. 12342, 2021.

ZHANG, H. et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.

ZHAO, Y. et al. Tsasnet: Tooth segmentation on dental panoramic x-ray images by two-stage attention segmentation network. *Knowledge-Based Systems*, Elsevier, v. 206, p. 106338, 2020.