UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, CIÊNCIAS E LETRAS DE RIBEIRÃO PRETO
DEPARTAMENTO DE COMPUTAÇÃO E MATEMÁTICA

JOSÉ ANDERY CARNEIRO

Algoritmo para segmentação dentária em radiografias panorâmicas

Enhanced tooth segmentation algorithm for panoramic radiographs

Ribeirão Preto–SP

2023

JOSÉ ANDERY CARNEIRO

Algoritmo para segmentação dentária em radiografias panorâmicas

Enhanced tooth segmentation algorithm for panoramic radiographs

Versão Corrigida

Versão original encontra-se na FFCLRP/USP.

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP) da Universidade de São Paulo (USP), como parte das exigências para a obtenção do título de Mestre em Ciências.

Área de Concentração: Computação Aplicada.

Orientador: Profa. Dra. Alessandra Alaniz Macedo

Ribeirão Preto–SP

2023

José Andery Carneiro

Algoritmo para segmentação dentária em radiografias panorâmicas. Ribeirão Preto–SP, 2023.

88p. : il.; 30 cm.

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras
de Ribeirão Preto da USP, como parte das exigências para
a obtenção do título de Mestre em Ciências.
Área: Computação Aplicada.

Orientador: Profa. Dra. Alessandra Alaniz Macedo

1. Diagnóstico bucal. 2. Sistemas de visão computacional. 3.Aprendizado profundo. 4. Radiografias panorâmicas. 5. Segmentação de dentes.

Este trabalho é dedicado aos meus familiares e amigos,
que tanto me apoiaram durante esta etapa da vida.
Em especial para Camila, Raquel e Kleber.

Agradecimentos

Agradeço sinceramente a todas as pessoas que tornaram possível a realização deste trabalho, uma jornada que se revelou desafiadora e recompensadora. Gostaria de expressar minha gratidão à minha orientadora, Profª Drª Alessandra Alaniz Macedo, pelos seus insights valiosos ao longo de todo o processo de pesquisa. Quero estender meus agradecimentos aos colegas do grupo de pesquisa InReDD (Interdisciplinary Research group in Digital Dentistry), em especial ao Mestre Breno Augusto Guerra Zancan. A colaboração e e a amizade foram essenciais para o desenvolvimento desta dissertação. Além disso, sou grato pelo suporte financeiro fornecido pela USP e pela CAPES. À minha família e amigos, que lá de Santa Rita, compartilharam comigo os sonhos e desafios desta jornada acadêmica, meu agradecimento é profundo. Seu apoio constante, encorajamento e compreensão foram pilares essenciais ao longo desse percurso. Cada uma dessas contribuições foi fundamental para a concretização deste trabalho, e por isso, expresso meu mais sincero agradecimento a todos que, de alguma forma, foram parte desta conquista.

“O correr da vida embrulha tudo,
a vida é assim: esquenta e esfria,
aperta e daí afrouxa,
sossega e depois desinquieta.
O que ela quer da gente é coragem.
(Guimarães Rosa, Grande Sertão: Veredas)

Algoritmo para segmentação dentária em radiografias panorâmicas

Resumo

A saúde bucal abrange uma ampla gama de condições, incluindo cáries dentárias, doenças periodontais, perda de dentes e câncer oral. Manter uma boa saúde bucal requer tanto a prevenção quanto o tratamento dessas condições. A detecção oportuna é crucial para evitar sua progressão. Embora as inspeções clínicas sejam eficazes em muitos casos, elas enfrentam limitações na identificação de problemas ocultos ou de difícil acesso. A radiografia dentária desempenha nestes casos um papel vital na garantia de diagnósticos precisos. Para aprimorar a velocidade e a precisão da análise de radiografias, os profissionais de saúde bucal estão cada vez mais adotando soluções que utilizam de Visão Computacional, com ênfase em Aprendizado Profundo para o processamento de imagens. Essas soluções deram origem a diversas ferramentas de diagnóstico, que vão desde a identificação de cáries até o auxílio em tratamentos de canal. Um passo inicial comum para essas ferramentas envolve a detecção dos dentes presentes nas imagens radiográficas. Para aprimorar essa fase crítica, apresentamos um sistema modular de segmentação de dentes. Esse sistema é composto por dois componentes-chave: (i) detecção da região bucal e (ii) segmentação de cada dente dentro da cavidade bucal identificada. Utilizamos a rede RetinaNet para a detecção da boca e a rede Cascade Mask R-CNN para a identificação dos dentes. Treinamos esses modelos com um conjunto de dados anotado por profissionais experientes, que inclui 935 radiografias panorâmicas com caixas delimitadoras da boca e, dentre elas, mais 605 com polígonos contornando os dentes, totalizando 14.582 dentes anotados. As tarefas propostas nesta pesquisa estão interligadas, com a saída de uma etapa sendo a entrada para a próxima. Nosso sistema obteve resultados excepcionais, com a detecção da boca alcançando 92,446 mAP e 0,982 F1-score, e a segmentação de instância dos dentes atingindo 79,222 mAP e 0,9894 F1-score, superando os benchmarks estabelecidos por estudos similares. Nossa ferramenta modular permite futuras expansões, integrando diversas novas funcionalidades, como a numeração dos dentes ou análise de cáries. Além de servir como auxílio diagnóstico, oferecendo suporte aos dentistas como uma segunda opinião, nosso sistema tem o potencial de agilizar a geração de relatórios epidemiológicos para grandes amostras populacionais. Ele também encontra relevância na medicina forense, uma área especializada dedicada à identificação de indivíduos com base em suas características orais e dentárias.

Palavras-chave: Diagnóstico bucal. Sistemas de visão computacional. Aprendizado profundo. Radiografias panorâmicas. Segmentação de dentes.

Enhanced tooth segmentation algorithm for panoramic radiographs

Abstract

Oral health encompasses a broad range of conditions, including dental caries, periodontal disease, tooth loss, and oral cancer. Maintaining optimal oral health requires both prevention and treatment of these conditions. Timely detection is crucial to prevent their progression. While clinical inspections are effective in many cases, they face limitations in identifying hidden or hard-to-reach issues. Dental radiography plays a vital role in ensuring accurate diagnoses. To enhance the speed and precision of radiograph analysis, oral health professionals are increasingly embracing advancements in Computer Vision, particularly leveraging Deep Learning for image processing. These techniques have given rise to various diagnostic tools, ranging from identifying cavities to classifying root canal treatments. A common initial step for these tools involves the detection of teeth in radiographic images. To enhance this critical phase, we introduce a modular system for teeth instance segmentation. This system comprises two key components: (i) dentomaxilo region detection (including mandible, maxilla and teeth) and (ii) segmentation of individual teeth within the identified dentomaxilo area. We employed RetinaNet for dentomaxilo region detection and Cascade Mask R-CNN for tooth identification. We trained these models using a dataset annotated by experienced professionals, which includes 935 panoramic radiographs with bounding boxes delimiting the dentomaxilo area and, among them, an additional 605 with tooth polygons, totaling 14,582 annotated teeth. These tasks are interconnected, with the output of one phase feeding into the next. Our system achieved good results, with dentomaxilo region detection scoring 92.446 mAP and 0.982 F1-score, and tooth segmentation attaining 79.222 mAP and 0.989 F1-score, surpassing benchmarks set by comparable studies. Our modular tool allows for future expansions, with the potential to integrate diverse new functionalities, such as tooth numbering or caries identification. Beyond serving as a diagnostic aid, offering support to dentists as a secondary opinion, our system has the potential to expedite the generation of epidemiological reports for large population samples.

Keywords: Oral diagnosis. Computer vision systems. Deep learning. Panoramic radiography. Teeth segmentation.

List of Figures

List of Tables

List of abbreviations and acronyms

AI	Artificial Intelligence
CNN	Convolutional Neural Network
COCO	Common Objects in Context
DL	Deep Learning
FC	Fully Connected
FCN	Fully Connected Network
FDI	Federation Dentaire Internationale
GT	Ground Truth
LVIS	Large Vocabulary Instance Segmentation
mAP	Mean Average Precision
ML	Machine Learning
NMS	Non-Maximum Suppression
PAN	Panoramic Radiograph
RoI	Region of Interest
RPN	Region Proposal Network

Contents

Introduction

Oral health is a multifaceted concept that serves as a indicator of an individual's overall well-being and quality of life. This concept encompasses a wide range of conditions and diseases, including, but not limited to, tooth decay, periodontal disease, tooth loss, and oral cancer (WHO, 2022). These conditions raise significant public health concerns due to their potential to substantially diminish individuals' quality of life, leading to discomfort, pain, and in some cases, chronic systemic infections[1].

Clinical examination, supplemented by the use of dental probes and hand mirrors, can identify advanced dental cavities resulting from caries. However, certain hidden or inaccessible lesions may necessitate radiographs examinations for accurate diagnosis (LIAN et al., 2021). In addition to detecting dental caries, which primarily affect the enamel and dentin, dental radiographs play a crucial role in uncovering various issues within mineralized tissues (KUMAR; BHADAURIA; SINGH, 2021; WANG et al., 2016). Without the aid of radiographic images, dentists would remain unable to detect these dental issues until they reach advanced and even irreversible stages. Consequently, radiographic images become invaluable sources of information for clinical diagnosis, treatment strategy formulation, and surgical interventions. They enable the discovery of hidden dental structures, as well as the identification of malignant or benign masses, bone loss, cavities, fractures in both bone and teeth, bone lesions, and other anomalies (WANG et al., 2016). Furthermore, dental radiographs play a pivotal role in epidemiological studies, being a source of information on large population samples (MURAMATSU et al., 2020) and forensic medicine, a specialized field dedicated to the identification of individuals through their oral and dental characteristics, which are esteemed as the most enduring anatomical features within the human body (OKTAY, 2018).

Oral radiographs are typically classified into two main categories: intraoral and extraoral (WANG et al., 2016; KUMAR; KHAMBETE; PRIYA, 2011). Three examples of commonly used intraoral radiographs are: interproximal (or bitewing), periapical and occlusal radiographs, both of which require the placement of imaging equipment inside the patient's oral cavity. On the other hand, extraoral radiography features panoramic radio-

[1] Oral health, World Health Organization (WHO). Available at https://www.who.int/health-topics/oral-health

graphy (PAN) as a technique with a high frequency of use. Unlike intraoral radiographs, PAN captures images from outside the patient's mouth, offering a comprehensive view of the mandible, maxilla, specific facial bones, and the entirety of the dentition in a single image, resulting in reduced patient discomfort and lower radiation exposure. (SILVA; OLIVEIRA; PITHON, 2018).

However, manual radiograph analysis is susceptible to misinterpretations and errors, especially when performed by less experienced professionals. Radiographs typically appear as grayscale images with potential issues such as noise, artifacts, low contrast, and uneven lighting. Furthermore, these images frequently feature overlapping structures, further complicating the analysis, as documented in (KUMAR; BHADAURIA; SINGH, 2021; MALLYA; ERNEST, 2018). Dentists routinely encounter a substantial volume of radiographs in their daily practice, aiming to assist in individual diagnoses or observational clinical studies(CHEN et al., 2019). Beyond the time-intensive process of analysing numerous radiographs, these professionals must meticulously document their observations in printed or digital records (ESTAI et al., 2022). Given this considerable workload, factors like inexperience, stress, and fatigue among dental practitioners can impede the accurate interpretation of their patients' oral conditions. All these obstacles can result in inconveniences for patients, including incorrect treatments, exacerbated issues, wasted time, and financial burdens, among other complications. Nevertheless, these challenges can be mitigated by the application of intelligent tools tailored to assist in radiograph analysis (CHEN et al., 2019).

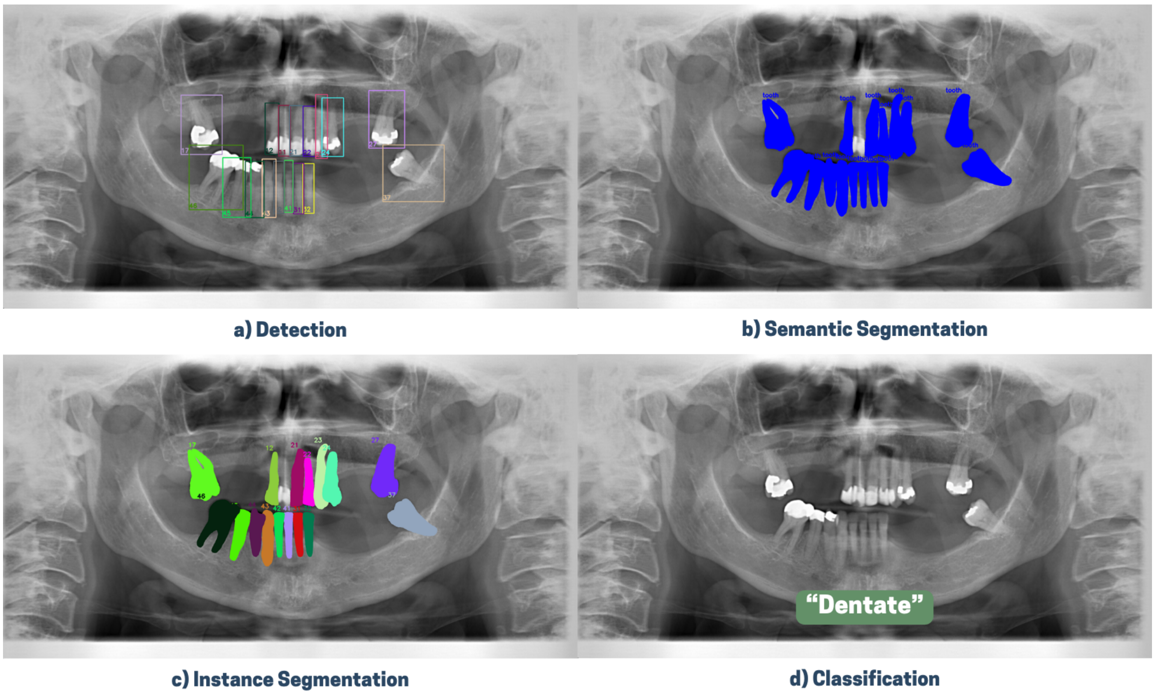
Intelligent tools commonly refer to Artificial Intelligence (AI), a broad category encompassing various computational methods trained to recognize patterns and deliver optimal responses to input data (BENKE; BENKE, 2018). Computer Vision (CV) delves deeply into the application of AI methods to extract invaluable insights from visual inputs, such as digital images and videos (FERNANDES; DóREA; ROSA, 2020; PRINCE, 2012). Machine Learning (ML) is a branch of AI that allows computers to learn without being directly programmed. As per the definition of ML given by (MITCHELL, 1997), "a computer program learns from experience E in the context of a specific class of tasks T and a performance measure P, if its performance on tasks within T, as assessed by P, demonstrates improvement with the accumulation of experience E". Deep Learning (DL), a sub-field of ML, empowers the manipulation of complex data structures, including images (SCHWENDICKE; SAMEK; KROIS, 2020), by representing them in a more abstract manner through non-linear layers (LECUN; BENGIO; HINTON, 2015). The foremost advantage of DL over conventional ML methods lies in its capacity to autonomously extract hierarchical and contextual features, contributing to a deeper understanding of complex visual data (MAHDI; YAGI; KOBASHI, 2020; BAYRAKTAR; AYAN, 2022). Convolutional Neural Networks (CNNs), a type of DL architecture inspired by the biological nervous system (OKTAY, 2017; CHEN et al., 2019; TUZOFF et al., 2019), play a significant role in the research dedicated to the advancement of AI tools for image interpretation. CNNs automate the pattern recognition process, furnishing a robust platform for CV technicians and researchers.

Within this context, the integration of DL techniques, mainly CNNs, into decision-support tools can offer substantial advantages for the analysis of patient images (PARK; PARK, 2018). DL's ability to extract relevant information via pattern recognition, regardless of its location, accelerates the data analysis process, leading to a substantial reduction in the time needed for these tasks compared to manual analysis without computational tools (LEE et al., 2018). The adoption of DL in diagnostic imaging presents other benefits, such as solving problems of subjectivity in individual exams, thereby enhancing the overall efficiency of healthcare services. Moreover, it contributes to cost reduction by automating repetitive tasks (SCHWENDICKE; SAMEK; KROIS, 2020). In a wider context, intelligent tools can take on a fundamental role in: (i) Elevating the precision of diagnoses and facilitating the planning of dental treatments; (ii) Reducing the time invested in radiographic image analyses and mitigating the risks of misdiagnosis which can often be influenced by stress, fatigue, or inexperience; (iii) Automating essential tasks, including report generation and the management of dental records (MAHDI; YAGI; KOBASHI, 2020; JADER et al., 2018; LEE et al., 2022; MURAMATSU et al., 2020).

CV applications employed in image processing cover a wide spectrum of tasks, ranging from classification, object region detection, object semantic segmentation, and object instance segmentation (SINGH; RAZA, 2022; MOHAMMAD-RAHIMI et al., 2022; SINGH; SEHGAL, 2021; JADER et al., 2018). In the medical domain, these applications are used in tasks such as automatically identifying and classifying pulmonary nodules (HOSNY et al., 2018), interpreting mammograms for cancer screening (SHIMIZU; NAKAYAMA, 2020), detecting liver diseases (ZHOU et al., 2019), and discerning melanomas and malignant carcinomas (SCHMIDT-ERFURTH et al., 2018). Similarly, in the field of dentistry, DL techniques are increasingly assuming an important role as useful tools to assist dental professionals in their decision-making processes (PARK; PARK, 2018). These applications encompass a diverse array of activities, including tooth numbering (SILVA et al., 2020; PINHEIRO et al., 2021; ZHANG et al., 2018; CHEN et al., 2019), automating the completion of dental records (MURAMATSU et al., 2020), diagnosing caries (GEETHA; APRAMEYA; HINDUJA, 2020; SINGH; SEHGAL, 2017), detecting dental restorations (ABDALLA-ASLAN et al., 2020), and the development of predictive models capable of establishing the relationship between the frequency and quality of brushing and instances of toothache (KIM; LIM; RHEE, 2009).

In the CV applications for oral radiographs, particularly those aimed at emphasizing dental characteristics, tooth detection or segmentation frequently stands out as the initial step. This process is of immense value to researchers in the field, enabling them to undertake a spectrum of tasks, including diagnosis, tooth numbering, dental age estimation, among others (LIN; HUANG; HUANG, 2013). Object detection encompasses the prediction of coordinates that define the smallest rectangle capable of enclosing the object, with these coordinates denoting the object's position, width, and height within the image. In a complementary manner, image segmentation involves dividing the image into discrete areas by grouping pixels that belong to the same object or region. Figure 1 depicts examples of these tasks. In a), an example of tooth detection in PANs is showcased, with bounding boxes delimiting each tooth separately. In b) and c), two examples of tooth segmentation are presented, grouping pixels belonging to the identified object in masks. The distinction between semantic segmentation, depicted in b), and instance segmentation, depicted in c), lies in the latter's ability to separate different instances of the same object, while the former creates a single mask containing all pixels of the identified objects in the image. Lastly, an example of classification is presented in d), where the entire image is analyzed by the application and assigned a label based on its characteristics. In this example, the label was "dentate," as the image contains a PAN with teeth.

Figure 1 – Examples of common computer vision tasks in the context of dentistry



a) Detection

b) Semantic Segmentation

c) Instance Segmentation

d) Classification

Source: Author's collection

In the context presented, this research introduces an innovative modular approach designed for the automatic segmentation of teeth in PANs. Our system performs two steps to achieve this objective: (i) the preprocessing step, inspired by (MURAMATSU et al., 2020), which entails detecting the dentomaxilo region (considering mandible, maxilla and teeth) to exclude surrounding areas that do not contribute to the network analysis, and (ii) the instance segmentation of the teeth using only the dentomaxilo region. We conducted experiments with various DL network architectures for each task, with the goal of optimizing mean Average Precision (mAP) in dentomaxilo region detection and teeth segmentation. The focus on PANs in this research is justified by the fact that this specific medical image configuration encompasses information about all teeth in a single image. As a valuable contribution, our aim is to enhance the precision of patients' dental diagnoses and their treatment planning, ultimately reducing errors. In addition to being a diagnostic aid tool, the resulting system holds the potential to automate tasks such as report generation and the completion of dental records in various contexts.

This work is part of a multidisciplinary research group at the University of São Paulo (USP) called InReDD (Interdisciplinary Research group in Digital Dentistry)[2]. As previously mentioned, tooth detection and segmentation are critical stages in various applications. This study focuses on the detection of the dentomaxilo area and the segmentation of teeth in PANs. Other lines of research within the InReDD group aim to use the outcomes of this investigation as a preprocessing step for tasks that involve tooth numbering according to the Federation Dentaire Internationale (FDI) numbering system, as well as the classification of teeth based on criteria such as decayed, restored, and other relevant labels. The structure of this document is as follows: In Chapter 1, the theoretical foundations supporting the project are elaborated in more detail; In Chapter 2, the systematic mapping conducted on the state of the art is presented; Chapter 3 provides a description of the materials and methods employed in our proposal; Chapter 4 presents the results obtained, discusses and compares them with related work; and Chapter 5 concludes with final remarks.

[2] InReDD. Available at https://sites.usp.br/inredd/mestrado/

1

Theoretical Foundation

In the development of our modular tooth instance segmentation system, it was imperative to leverage concepts and definitions from Dentistry, Radiology and Computer Vision (CV). The theoretical background derived from Dentistry and Radiology provide insights into the characteristics of the images generated by radiographic examinations, particularly with regard to panoramic radiography (PAN). These concepts are elucidated in Section 1.1. Consecutively, Sections 1.2, 1.3, 1.4 and 1.5 delve into the topics of CV, offering an exploration of the Neural Networks employed in our system and associated concepts.

1.1 Radiographs

Dental radiological examinations serve various purposes, including endodontic procedures, forensic investigations, and diagnosing conditions like caries (GURSES; OKTAY, 2020). Particularly in diagnostics, radiographs are highly recommended due to their ability to reveal both the internal and external morphology of teeth, providing insights into size, location, and the condition of tissues hidden from the naked eye (RAD et al., 2018).

To generate a radiographic image, patients are exposed to a minimal amount of radiation that crosses the region of the patient's body under investigation, resulting in a two-dimensional composition of overlapping shades, including black, white, and gray (WHAITES; DRAGE, 2013). These shades emerge from the interaction of X-rays with the patient's tissues. X-rays penetrate and cross these tissues and reach the film or sensor of the X-ray device. Different tissues absorb X-rays differently. Radiopaque tissues impede X-ray passage, appearing brighter in the final image. Conversely, radiolucent tissues allow more X-ray penetration, resulting in darker shapes in the image.

Dental radiographs are categorized into two main types: intraoral and extraoral, determined by the placement of the radiographic film or sensor (digital radiography) either within or outside the oral cavity (KUMAR; KHAMBETE; PRIYA, 2011). Among intraoral options, three modalities are prominent. Bitewing, also known as interproximal, involves the patient biting down on a positioning device, allowing simultaneous imaging of upper and lower teeth in a single frame. Typically, this radiograph covers a specific region of the mouth, usually the crowns of molar or premolar teeth (WHAITES; DRAGE, 2013). The second intraoral radiography, known as periapical, captures the entire tooth structure, including the crown, root, surrounding alveolar bone, and neighboring regions, but is limited to two or three teeth per image. Lastly, the occlusal radiograph captures the entire dental arch while the patient maintains occlusal pressure on the film or sensor (WANG et al., 2016). Interproximal and periapical radiographs are commonly used for caries diagnosis. However, some patients may struggle to hold the film or sensor inside their oral cavity (WHAITES; DRAGE, 2013), especially those with a gag reflex or significant discomfort with intraoral methods, such as children or individuals with disabilities (CLIFTON; TYNDALL; LUDLOW, 1998). Patient cooperation is crucial for producing artifact-free images (ABDINIAN et al., 2015).

Extraoral radiographs eliminate the need to place films or sensors inside the oral cavity. These images encompass all teeth, nasal and facial bones, the chin, and the joints between the jaws and the skull, offering a comprehensive view rather than isolating specific dental regions, as intraoral radiographs (SCHWENDICKE et al., 2019; SILVA; OLIVEIRA; PITHON, 2018). The method for obtaining PANs involves the synchronized rotation of the X-ray source and image receptor around the stationary patient (KAMBUROGLU et al., 2012). These exams are quick, employ minimal radiation doses, cause no patient discomfort, are cost-effective, and relatively straightforward to perform (AKKAYA et al., 2006).

However, factors such as morphological variations in patients' jaws and the positioning of the patient's head during the X-ray procedure can lead to distorted and blurry images (ABDALLA-ASLAN et al., 2020). As highlighted by (OKTAY, 2018), the substantial distortion and overlap of teeth frequently encountered in PANs can pose challenges in segmenting tooth contours, adding complexity to panoramic image analysis. Moreover, (FUKUDA et al., 2019) emphasize the frequent overlapping of anatomical structures in PANs, making interpretation challenging, especially for inexperienced observers. Consequently, diseases might remain undetected due to inaccurate diagnoses. In this context, applying CV methods can enhance precision in analysis and diagnosis.

1.2 Convolutional Neural Network and Deep Learning

A Convolutional Neural Network (CNN) stands out as a specialized variant of artificial neural networks designed for processing and analyzing visual data, such as images and videos. CNNs have demonstrated their effectiveness in a diverse range of computer vision tasks, including image classification, object detection, and image segmentation. CNNs are a subset of Machine Learning (ML) techniques. ML is capable of making highly accurate predictions by leveraging pre-existing data, often referred to as a dataset, which serves as the foundation for learning (MOHRI; ROSTAMIZADEH; TALWALKAR, 2018). This dataset can either be provided by a programmer or gathered through interaction with the specific problem being analyzed.

Deep Learning (DL), a subfield of ML focuses on neural networks with multiple layers (deep neural networks) enabling computer systems to enhance their performance through experience and data. According to (LECUN; BENGIO; HINTON, 2015), training a ML neural network to its full potential necessitates the selection of object features that contribute most effectively to the network's learning, through preprocessing techniques. What sets DL apart is its layers' ability to autonomously explore and choose the most relevant features from input data. Guided by a learning mechanism, DL systems reduce reliance on human experts for the extraction of relevant information for classification or detection, often revealing intricate patterns elusive to human perception. A typical DL system can be depicted as a sequence of simple and non-linear layers, where information undergoes progressive abstraction at each layer, preparing it for the classification process. The upshot is the capacity to master the learning of highly complex functions (LECUN; BENGIO; HINTON, 2015).

In the domain of computer vision, Deep CNNs can automatically learn and extract relevant features from images, starting from low-level features like edges and textures to higher-level features like object parts and object categories. These networks, often referred to as "convnets", are adept at uncovering hierarchical patterns residing within a group of neighboring pixels in an image, achieved through a combination of convolutional layers, pooling layers, and fully connected layers (CHOLLET, 2017).

At the core of this architecture are the convolutional layers. These layers comprise a series of trainable filters, each consisting of (MxM) pixels, named "kernels". These kernels engage in the convolution operation, involving element-wise weighted multiplication between each pixel in the input image and its neighboring counterparts encapsulated within the filter. The weights and biases associated with these filters take responsibility for highlighting and recognizing patterns within the image, including contours and shapes.

After each convolutional layer, an activation function is frequently employed to introduce non-linearity into the model. This function acts as an activator for the convolutional layer, deciding whether the feature map generated by the layer should be forwarded to the subsequent layers. This function helps the network learn complex patterns and relationships in the data. Pooling layers are used to reduce the spatial dimensions (width and height) of the data while preserving important information. These layers are alternated with the convolutional layers, helping in the reduction the computational demands and making the network more robust to variations in the input. After a series of convolutional and pooling layers, CNNs culminate in a flatten layer, succeeded by one or more fully connected layers. The purpose of the flatten layer is to transform the multi-dimensional output from the convolutional layers into a one-dimensional vector, serving as a transition from spatial hierarchies to the fully connected layers. The fully connected layers are similar to those in traditional artificial neural networks. Neurons in a fully connected layer establish connections to all activations in the preceding layer. Similar to the convolutional layers, each connection is associated with learnable weights and biases. In the final fully connected layer, a distinctive activation function known as Softmax, is responsible for making final decisions or predictions based on the learned features.

The training process involves forward and backward passes through the network, adjusting the weights and biases of convolutional and fully connected layers. The primary objective is to minimize the difference between predictions and actual outcomes. To commence training, weights and biases are initialized with random values. The network then processes training images, passing information through convolutional, pooling, flatten, and fully connected layers. At the output, a prediction is generated and compared with the Ground Truth (GT) label for the corresponding training image. This comparison is done through a loss function that quantifies the disparity between the network's predictions and the actual labels. Using this measured error, an algorithm known as backpropagation computes a loss gradient for each weight and bias. These gradients indicate the direction and magnitude of adjustments needed on the weight and bias values to reduce the overall network error. A hyperparameter called learning rate, controls the intensity of corrections applied to these values based on the loss gradient. Each processed image (forward pass) and weight update (backward pass) is called an iteration. This iterative process is repeated numerous times, causing weight and bias updates that guide the network toward minimizing prediction errors. Another significant hyperparameter, known as Batch Size, allows the network to update weight and bias values by considering the average error over multiple iterations rather than after each processed image (forward pass).

Periodically, the network undergoes evaluation on a separate validation set of images/GT to monitor performance on data not encountered during training, preventing overfitting and ensuring adaptability to unseen data. Overfitting arises when a model not only captures the underlying patterns in the training data but also assimilates noise and specific fluctuations. While the model excels on the training data, its ability to generalize to new, unseen data diminishes. A key approach to detect overfitting involves using the validation set. By plotting the model's error or accuracy curves measured on both the training and validation sets on the same graph, we expect that in the early stages of training, the error will decrease, and accuracy will increase. However, as the training process repeats, the model tends to learn more slowly. Overfitting manifests as a widening gap between training and validation performance, observed on the graph as sustained improvement on the training set and a decline on the validation set. For instance, the training error continues to decrease or remain stable, while the validation error starts to increase with each iteration. Recognizing overfitting is pivotal for deciding when to halt training, ideally just before overfitting, where the model attains optimal performance for that specific scenario. After training, the network is tested on a completely independent test set to assess its ability to generalize to unseen data.

A CNN offers two advantages over conventional artificial neural networks: i) the ability to discern patterns independent of their spatial location within images, enabling the detection of a tooth, whether it resides in the mandible or maxilla; and ii) the sharing of weight and bias, effectively reducing the quantity of variables that require training in each convolutional layer (LEE et al., 2018).

1.3 Faster R-CNN

Conventional CNNs are typically designed to classify images based on the patterns they contain. In 2014, (GIRSHICK et al., 2013) developed the R-CNN (Region-Based CNN), one of the pioneering approaches for object detection using DL, integrating a selective search (SS) algorithm with a CNN. The SS algorithm proposes a specified number of rectangular regions in an image, termed Regions of Interest (RoIs), which might contain objects of interest. In this approach, each RoI went through a CNN responsible for generating the region feature map, and these features were subsequently classified using a Support Vector Machine (SVM) for object detection. The computational cost of running the CNN for each RoI was the main issue with this proposal. To address this challenge, the same authors introduced a variation called Fast R-CNN approximately a year later. In Fast R-CNN, the image is processed through a CNN network (called backbone) one single time to generate a feature map for the entire image. Simultaneously, the image underwent SS, and through a new layer called RoI Pooling Layer, the network could extract the set of pixels from the feature map corresponding to each of the RoIs proposed by SS. This idea of sharing the computation of the CNN backbone across all region proposals reduced the complexity and time required for training the network (GIRSHICK, 2015). Finally, (REN et al., 2017) introduced a new model known as Faster R-CNN, which presented an improvement over its predecessors by replacing the SS algorithm with a network called the Region Proposal Network (RPN). The RPN, a fully connected network, predicts region proposals and their objectness scores in a single forward pass.

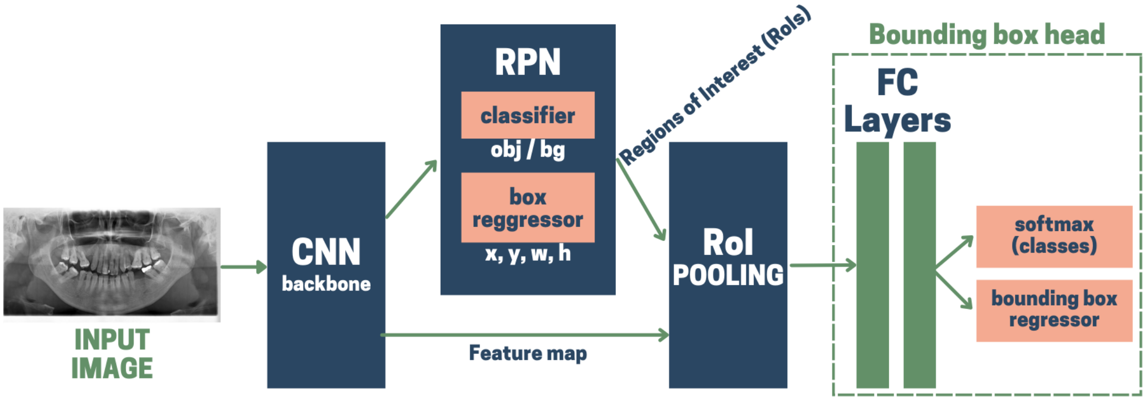
Figure 2 illustrates the Faster R-CNN network. It performs two-stage object detection. Preceding these detection stages, the network employs an initial image preprocessing network known as the backbone, tasked with generating the image feature map. The input image undergoes processing through the backbone network, which can be any image classification CNN with flatten and fully connected (FC) layers removed. The desired output of this preprocessing, highlighted in Figure 2 as the "CNN backbone," is the feature map generated by the convolutional layers.

The first stage of object detection, the RPN comes up with a certain number of possible RoIs where the object might be located. The RPN is a fully convolutional network (FCN) that uses the feature map as input. For each pixel in the feature map, RPN predicts multiple anchor boxes, which are predefined boxes with various aspect ratios and scales. The RPN estimates two lists of values for each anchor box: the first list contains the probability that the anchor box contains an object ("obj") or background ("bg"), shown as "classifier" in Figure 2, and the second list contains the refined coordinates of the proposed region (coordinates "x" and "y" of the upper-left vertex of the region, width "w," and height "h"), shown as "box regressor" in Figure 2. Since multiple proposals may correspond to the same object, the Non-Maximum Suppression algorithm (NMS) is applied to eliminate redundant and overlapping regions, retaining only the most confident RoIs. In the RoI Pooling process, these RoIs are resized to a standard bounding box size and then flattened to a fixed-length feature vector to ensure that proposals of different sizes are represented uniformly and can be fed into a fully connected layer.

In the second stage of detection, the resized vectors are processed by the "bounding box head" of the network, composed by two FC layers and two sibling layers: One branch ("softmax") is another FC layer with softmax activation function, responsible for classifying the object among the labels or as background for object classification, which assigns a class label to the RoI. The other branch ("bounding box regressor") refines the coordinates of the RoI bounding box to align it more accurately with the actual object boundaries.

The learning process of Faster R-CNN involves a series of forward and backward passes on training images. During the forward pass, RoIs are computed using the RPN, and these proposals are further refined with the bounding box head. Subsequently, the error in the generated predictions is calculated. The Faster R-CNN's loss function combines different components in a weighted manner: the prediction error of potential RoI proposals generated by the RPN, the error in regressing the bounding box coordinates, and the error in classifying the object contained within the bounding box. Following the computation of this error, the weights and biases of the RPN and bounding box head are adjusted through the backpropagation algorithm. While it is possible to train the backbone network in this process, it is common for this network to be leveraged from other tasks through the transfer learning process, as explained in Section 1.6.

Figure 2 – Faster R-CNN architecture



Source: Author's collection

1.4 RetinaNet

Similar to the Faster R-CNN network, RetinaNet is a DL model designed for object detection in images, drawing labeled bounding boxes around objects of interest (LIN et al., 2017). Introduced in 2017, its objective was to strike a balance between the accuracy of Faster R-CNN, considered a two-stage object detection model, and the speed of YOLO (You Only Look Once), considered an one-stage object detection model (REDMON et al., 2016). The distinction between one-stage and two-stage models lies in the inclusion of a RPN as an intermediary step, often referred to as the "first stage" in generating RoIs. One-stage models bypass this intermediary step, enabling faster predictions. However, until the advent of RetinaNet, they were generally considered less accurate (LIN et al., 2017).

Figure 3 provides an overview of the RetinaNet model. It starts with the preprocessing backbone network, which can be a ResNet, ResNeXt, or similar CNN architecture, removing the flatten and FC layers. The backbone's purpose aligns with that of Faster R-CNN: extracting high-level features from the input image to generate the feature map. These backbone networks are typically deep, incorporating multiple convolution layers consecutively, thereby generating a sequence of feature maps that progressively capture more complex features. This sequence of feature maps is visually represented in Figure 3 as an ascending pyramid of layers enclosed within the box labeled "CNN backbone."

RetinaNet introduces an innovation by integrating the Feature Pyramid Network (FPN) atop the backbone. FPN plays a pivotal role in detecting objects of diverse scales within an image. It achieves this by constructing a pyramid of feature maps at multiple scales, empowering the model to detect both small and large objects. For each layer of the feature map pyramid, RetinaNet generates several potential object location candidates. It does so by using the Anchor Box concept, similar to Faster R-CNN, where predefined rectangles with varying sizes and aspect ratios pass over all pixels of the feature map in pursuit of objects of interest. However, the distinction is that, while Faster R-CNN's RPN proposes RoIs based on a single feature map, FPN uses anchor boxes to predict object locations across distinct feature maps of varying dimensions. Subsequently, the Bounding Box Head analyses each object location predicted with the anchor boxes. This stage encompass (i) determining whether an object is present within a specific anchor box and assigning a corresponding class label and (ii) fine-tuning the anchor box coordinates to achieve a more precise alignment with the detected object's boundaries.

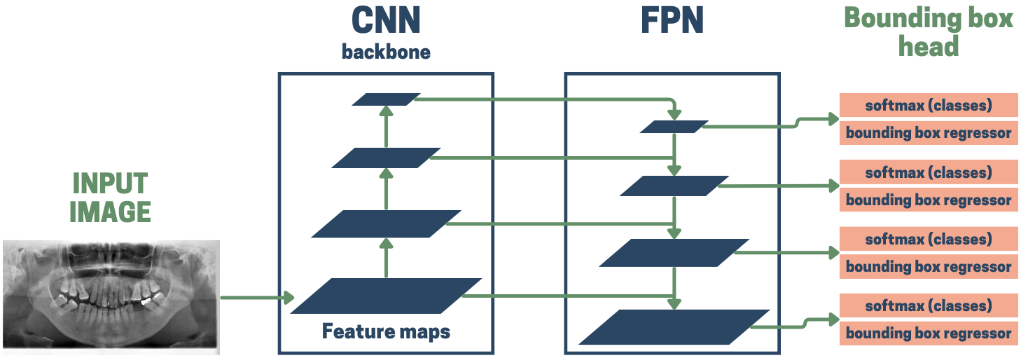
The learning process of RetinaNet involves a series of forward and backward passes on training images. During the forward pass, feature maps are extracted, and the bounding box regression and object classification are performed through the bounding box head. The subsequent step entails calculating the loss by comparing the predicted outputs with the GT annotations. This loss is focused on the bounding box head and comprises two components: classification loss and regression loss. A significant contribution of RetinaNet is the introduction of Focal Loss, which serves as the loss function in the classification part. Focal Loss addresses the challenge of class imbalance in object detection datasets. It down-weights well-classified examples during training, allowing the model to concentrate more on challenging examples. This approach significantly enhances the detection of rare or complex classes (LIN et al., 2017). Based on this combined error, the weights and biases of the bounding box head are adjusted using the backpropagation algorithm. Similar to Faster R-CNN, the backbone network can be reused through the transfer learning process, as elucidated in Section 1.6.

1.5 Mask R-CNN

Mask R-CNN (HE et al., 2017) is an extension of the Faster R-CNN architecture introduced in 2017, incorporating a novel “mask head” branch, parallel to the two other described for Faster as “bounding box head”. This addition expands the network's capabilities, enabling it to predict pixel-level object masks in addition to object bounding boxes.

Figure 4 presents the Mask R-CNN. The backbone CNN serves the same role as in Faster R-CNN and RetinaNet, processing the input image to generate the feature

Figure 3 – RetinaNet architecture



Source: Author's collection

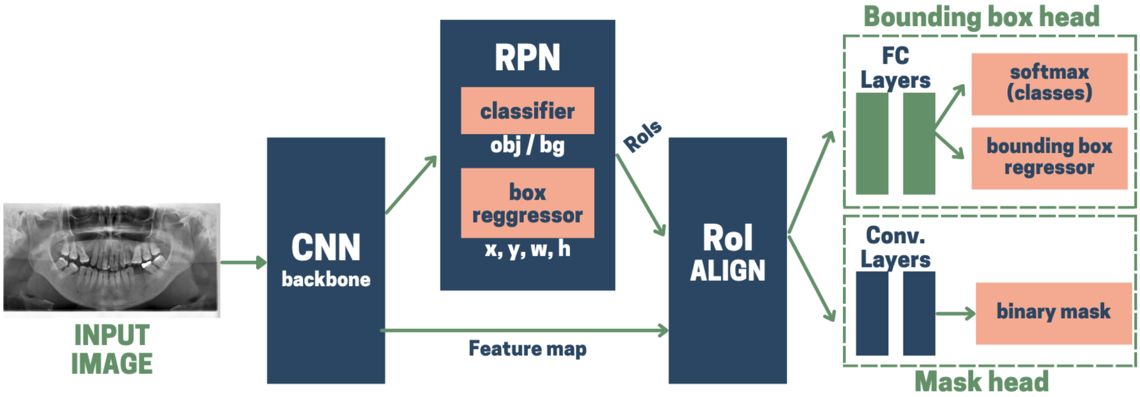
map. This CNN can be a VGG-16, ResNet, DenseNet, among others, and can utilize pre-trained weights and biases. Another component shared with Faster R-CNN is the RPN, which aims to propose potential RoIs where objects may be located. These proposals are generated using predefined anchors boxes and the feature map, subsequently scored based on their objectiveness and refined in position. Similar to Faster R-CNN, the NMS algorithm is employed to eliminate overlapping and low-confidence regions, selecting a configurable number of top candidates.

One difference between Faster R-CNN and Mask R-CNN is the replacement of the RoI Pooling technique with RoI Align. While the core objective remains consistent (generating vectors of uniform size from the top proposals provided by the RPN and NMS) RoI Align avoids the misalignment issue encountered when converting pixel coordinates to feature map coordinates (HE et al., 2017). This change significantly enhances mask prediction accuracy. Subsequently, these vectors are delivered to both the bounding box head (which remains identical to that of Faster R-CNN) and the mask head, a fully convolutional network responsible for segmenting the object of interest. The Mask Head assigns labels to each pixel within the bounding box, determining whether it belongs to the object or the background.

As Mask R-CNN is an extension of Faster R-CNN, its learning process closely mirrors that of its predecessor. As described in the preceding paragraphs, beyond object detection, Mask R-CNN creates pixel-wise segmentation masks for each identified object. Consequently, the loss function employed for error computation encompasses components from RoI proposals, classification, bounding box regression, and introduces a novel element related to segmentation mask prediction. Following the calculation of the error, the backpropagation algorithm adjusts the weights and biases within both the RPN and the Bounding Box Head, as well as the Mask Head.

Cascade R-CNN (CAI; VASCONCELOS, 2018), introduced in 2018, is an expan-

Figure 4 – Mask R-CNN architecture



Source: Author's collection

sion of the Faster and Mask R-CNN models, aiming to enhance object detection and segmentation accuracy. This innovative approach creates a cascade architecture by integrating multiple detectors, with the objective of iteratively refining bounding boxes and, consequently, segmentation masks at each stage of detection. In this cascade model, the initial bounding box head refines and assigns class labels to the bounding boxes proposed by the RPN, using the features from the backbone network. In the next stages this iterative process is meticulously repeated, with each successive bounding box head contributing to the refinement of bounding box coordinates and class labels inherited from the previous bounding box heads. These stages are designed to target the reduction of false positives and the enhancement of object detection accuracy (CAI; VASCONCELOS, 2018). Each individual bounding box head uses a specific Intersection over Union (IoU) threshold for bounding box generation and uses the outputs of previous detectors to refine its own weights. Furthermore, Mask Heads are implemented in each Bounding Box Head, similarly to the Mask R-CNN model.

1.6 Final Remarks

Interproximal and periapical radiographs constitute essential intraoral examinations in dental diagnostics. However, they can cause a higher level of patient discomfort compared to PANs, necessitating patient tolerance and cooperation (ABDINIAN et al., 2015). In contrast, PANs provide a comprehensive view of the entire dental arch within a single examination. This procedure is characterized by its simplicity, speed, minimal radiation exposure, and enhanced patient comfort (AKKAYA et al., 2006). The field of computer vision has witnessed significant advancements in recent years, especially in the application of DL models. These models have grown in complexity and have shown remarkable accuracy in object detection and segmentation tasks. Consequently, with access to an appropriate database and suitable annotations, it is possible to train DL models to accurately segment teeth in PANs, thereby creating a valuable tool for diagnostic support, epidemiological research among others.

In this research, advanced DL detection networks, including Faster R-CNN and RetinaNet, as well as the instance segmentation networks Mask R-CNN and its adapted version, Cascade Mask R-CNN, were employed. To avoid implementing each of these networks from scratch, the models provided by the open-source tool Detectron2, developed by the Facebook Artificial Intelligence Research Group (FAIR)[1], were leveraged. Training a model from scratch for a specific detection or segmentation task requires a substantial volume of annotated data and considerable computational resources. To mitigate these challenges, the transfer learning technique was adopted. In transfer learning, a model previously trained on one task is adapted or fine-tuned for a second task.

This fine-tuning process involves reusing the learned weight and bias values from the pre-trained model for a new task, utilizing a different dataset. In this research we utilized models pre-trained by FAIR for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC)[2]. These pre-trained models had been trained on two datasets: the Common Objects in Context (COCO) instance segmentation dataset[3], containing over 330 thousand images and 1.5 million annotated objects spanning more than 80 categories, and the Large Vocabulary Instance Segmentation (LVIS) dataset[4], encompassing over 164 thousand images and 2 million annotated objects across more than 1200 categories. To adapt the weights and biases of these models for our specific task, the fine-tuning technique was employed. This involved not retraining the weights and biases of the backbone networks (freezing). With our dataset, we only trained the weights and biases of the RPN and FPN (when applicable), as well as the bounding box heads and mask heads of the networks.

[1] Detectron2. Available at https://github.com/facebookresearch/detectron2

[2] Imagenet LSVR Challenge. Available at https://www.image-net.org/challenges/LSVRC/index.php

[3] COCO dataset. Available at https://cocodataset.org/#home

[4] LVIS dataset. Available at https://www.lvisdataset.org

2 Related Work

Systematic Mapping (SM) consists of the identification and classification of primary studies, aimed at gathering information into a particular subject matter, with the goal of identifying best practices and common trends. SM adopts a broad research question and refrains from employing meta-analysis and narrative synthesis techniques (KITCHENHAM; CHARTERS, 2007).

As previously stated, this study forms a component of the InReDD group's initiatives and lays the groundwork for ongoing research endeavors within the same group. Notably, it intersects with another research on tooth numbering employing the Federation Dentaire Internationale (FDI) system, directly influencing the proposed teeth instance segmentation. This project is being pursued by Breno Augusto Zancan, a master's student of the Programa de Pós Graduação em Computação Aplicada (PPG-CA) and who is also a member of the InReDD group.

Given the close connection between these research, a collaborative SM was executed to investigate the context of both areas (teeth detection/segmentation and tooth numbering). The objective was to collect information and conduct an analysis of the current state of the art within this areas. The task of dentomaxilo region detection (considering mandible, maxilla and teeth), which is a component of this research, was omitted from the SM. A preliminary literature review we conducted prior to this SM revealed that dentomaxilo region detection is an infrequently explored area in the research community, with no dedicated studies available.

The manual developed by (KITCHENHAM; CHARTERS, 2007) served as a guideline in planning this SM. For the systematic organization and compilation of the acquired information, we opted for the online tool Parsifal [1]. The methodologies employed, the results obtained, and the discussions generated from this research have been formally submitted as an article titled "Deep Learning to Detect and Classify Teeth, Dental Caries, and Restorations: A Systematic Mapping" to the Journal Dentomaxillofacial Radiology.

[1] Parsifal. Available at https://parsif.al/about/

Currently, the article is undergoing the peer review process.

The purpose of the SM is to provide theoretical support for this research. Its development is briefly outlined in this chapter, following these phases: Section 2.1 presents the SM planning and details its conduction with the search string formulated during the planning, and Section 2.2 presents and analyzes the mapping results.

2.1 Systematic Mapping Planning and Conduction

In the planning phase, we established the SM scope. The principal aim of this SM was to systematically analyze papers that implemented Deep Learning (DL) methodologies for detecting or segmenting teeth, caries, and restorations, and also the classification of teeth in dental radiographs. To structure our investigation, we divided this objective into five specific goals (SGs) and corresponding research questions (RQs):

- [**SG1**]: The identification and comparison of the source and annotation protocol, including evaluation of consistency (eg, number and experience of annotators) and adequacy (eg, number of samples) of the databases.
    - **RQ1**: What types, sources and numbers (training, validation and testing) of radiographs are used by the retrieved studies? Has the use of these images been approved by an ethics committee? How many specialists annotate images and prepare the study dataset? How good are they in terms of levels of knowledge and experience? Is there information about how the datasets are stored during the annotation period? What is the availability of the datasets? Public or private repositories?
- [**SG2**]: The identification of the methods used to classify teeth, caries and restorations.
    - **RQ2**: What techniques are used to classify decayed and restored teeth on radiographs?
- [**SG3**]: The identification of the methods used for numbering or identifying teeth.
    - **RQ3**: What techniques number or identify teeth on radiographs?
- [**SG4**]: The identification of the methods used to segment or detect teeth, caries and restorations.

  - **RQ4**: What techniques segment or detect teeth, caries and restorations on radiographs?

- [**SG5 Outcome**]: Comparisons of the metrics that evaluate the methods resulting of SG2,SG3 and SG4.

  - **RQ5**: What evaluation criteria measure the quality of the results? What are the best results considering each technique?

To construct the search strings, we create a collection of keywords designed to tackle the RQs. RQ 1 and 5 focus respectively on datasets and results evaluation. In contrast, RQs 2, 3, and 4 correspond to distinct tasks. Consequently, distinct search strings were created for RQ 2, 3, and 4, each incorporating task-specific keywords. Lastly, a fourth search string was generated by amalgamating the three previous ones, with the objective of identifying papers that simultaneously address all three tasks of RQ 2, 3, and 4. We also created a set of keywords named "Context Keywords" (CK) to restrict the research scope. All search strings include CK, which comprises the following terms: *("dentistry" OR "digital imaging" OR "odontology" OR "radiology") AND ("tooth" OR "carie" OR "decay" OR "dent" OR "dentition" OR "filling" OR "restoration" OR "teeth") AND ("deep learning" OR "mask" OR "neural network" OR "u-net") AND ("x-ray" OR "bitewing" OR "panoramic" OR "periapical" OR "radiograph").* The strings were created as follows:

- String 1 (S1): CK *AND ("classification" OR "labeling")*

- String 2 (S2): CK *AND ("numbering" OR "type identification")*

- String 3 (S3): CK *AND ("segmentation" OR "demarcation" OR "detachment" OR "mask" OR "partition)*

- String 4 (S4): *CK AND S1 AND S2 AND S3*

The databases utilized for this SM were the ACM Digital Library[2] and the IEEE Digital Library[3], both focused on computing. Additionally, we search on PubMed[4], focused on medical and dental research, and Scopus[5], a renowned digital library in both the fields of computing and dentistry. The search strings were applied to these digital libraries between April 22 and April 28, 2022, and were updated between June 8 and June 18, 2023.

[2] ACM Digital Library. Available at http://portal.acm.org

[3] IEEE Digital Library. Available at http://ieeexplore.ieee.org

[4] PubMed. Available at https://pubmed.ncbi.nlm.nih.gov

[5] Scopus. Available at http://www.scopus.com

The selection criteria were defined during the planning phase. Inclusion criteria (IC) permitted the approval of works that directly addressed the research questions. For our context, the sole inclusion criterion was (IC): the paper aligns with RQ2, 3, or 4. Exclusion criteria (EC) were established to remove incomplete, inaccessible, or irrelevant articles or publications other than articles. The adopted exclusion criteria were as follows: EC1: The full paper is not available on the Web or on the *Portal de Periódicos da Capes* (Available at https://www.periodicos.capes.gov.br); EC2: The article is a book chapter; EC3: It is not a primary study; EC4: The goal of the paper does not collaborate with the research questions; EC5: The article does not use dental radiographs; EC6: The article was published before 2012 (the use of DL in images became popular with the publication of (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), which took place after that date (MOHAMMAD-RAHIMI et al., 2022).

During the conduction phase of the SM, the search strings were executed across the chosen digital libraries, and the found articles were submitted to a two-step filtering process. In the initial step, duplicate articles between databases and those meeting the ECs were excluded, with only the titles and abstracts considered. In the subsequent round of filtering, the remaining articles were fully analysed and selected based on both the ECs and IC. In total, 394 articles were initially found. After the first filtering stage, 125 articles remained, and after the second round of filtering, 63 articles were approved. Furthermore, a manual search was conducted within the references of the 63 approved articles to identify related articles that had not been initially discovered. This supplementary search yielded an additional 6 studies, resulting in a total set of 69 articles for the SM.

2.2 Systematic Mapping Results

During our SM, we categorized the 69 approved articles into eight distinct categories based on their primary objectives. Here are the categories along with the respective number of articles in each:

- C1 (1 article): Tooth detection;
- C2 (14 articles): Tooth classification into decayed tooth or restored tooth classes;
- C3 (10 articles): Tooth semantic segmentation;
- C4 (4 articles): Tooth instance segmentation;
- C5 (21 articles): Tooth classification into type or FDI numbering;
- C6 (8 articles): Detection of caries or restoration;

- C7 (3 articles): Classification of caries stages or restoration types;
- C8 (8 articles): Others.

For this research, the 4 articles in category C4 stand out, which have the same objective as ours. Below, we provide concise summaries of each of these articles.

The first article focuses on performing instance segmentation of teeth in panoramic radiographs (PANs) using the Mask R-CNN network pretrained without altering its default parameter values. To train the network, the authors adapted a subset of the UFBA-UESC Dental Images dataset (SILVA; OLIVEIRA; PITHON, 2018), which was annotated for the semantic segmentation of teeth. They manually divided the region containing teeth into separate entities. Utilizing 276 PANs, the authors achieved precision, recall, and F1-score values of 0.94, 0.84, and 0.88, respectively (JADER et al., 2018).

The second article shares the same objective of segmenting tooth instances in PANs. It utilizes 50 annotated images from their specific dataset (obtained from Yonsei University Dental Hospital) and employs data augmentation techniques to train the Mask R-CNN network. Rather than containing all tooth annotations within a single image, the authors generate an image/annotation pair for each tooth. To illustrate, if a radiograph comprises 28 teeth, this data augmentation technique results in the generation of 28 training images. This approach yields precision, recall, and F1-score values of 0.858, 0.893, and 0.875, respectively (LEE et al., 2020).

The third article employs a distinct strategy for tooth instance segmentation. The initial step involves detecting teeth in PANs using the DeepLabV3 object detection network. Subsequently, they apply an fully connected network (FCN) to each bounding box encompassing the detected teeth, thereby creating masks for individual teeth. The FCN used in this process is the same segmentation network employed in the Mask R-CNN. Using 153 PANs from the M3BE database (DOI: 10.1111/ocr.12297), the authors achieve precision, recall, and F1-score values of 0.969, 0.983, and 0.975, respectively (LEITE et al., 2021).

The fourth and final article introduces the concept of Federated Learning (FL), described by the authors as a method that "enables collaborative training of Artificial Intelligence (AI) models from multiple data sources without directly sharing data" (SCHNEIDER et al., 2023). To achieve individual tooth segmentation, the authors utilize 4,177 PANs from nine diverse centers worldwide, including universities and clinics. They employ the UNet++ network for this purpose. The key idea is that each center independently trains the same model, and subsequently, the parameter sets found in each center are aggregated into a single final model. Although they report the F1-score achieved by the models at each center, the paper does not explicitly mention the metric achieved by the final model, with aggregated parameters (SCHNEIDER et al., 2023).

Furthermore, a number of other articles included in the SM utilize tooth instance segmentation as an intermediary step to accomplish their primary objectives. Notably, two categories, C2, “Tooth classification into decayed tooth or restored tooth classes” , and C5, “Tooth classification into type or FDI numbering” , capture this trend. From this subset of articles, five have been chosen for detailed discussion. These selections were made based on their use as sources of inspiration for our research (CHEN et al., 2019; MURAMATSU et al., 2020; ESTAI et al., 2022) and their application of metrics similar to those employed in our study (SILVA et al., 2020; PINHEIRO et al., 2021). Here is an overview of these articles:

In the category C2, an application was conducted in (CHEN et al., 2019) for the automatic detection and numbering (using the FDI system) of teeth, including the identification of missing teeth. The authors employed the Faster R-CNN architecture in conjunction with a custom Deep Neural Network, trained using 1250 periapical radiographic images. To evaluate the performance of the developed tooth detection system, three expert dentists with varying levels of experience were invited to annotate the test dataset. The results of the proposed system closely matched those of the least experienced dentist. Tooth detection achieved a precision of 0.988 and a recall of 0.985 (CHEN et al., 2019).

In the category C5, (MURAMATSU et al., 2020) introduces a semi-automatic technique for dentomaxilo region detection. The authors aimed to precisely delineate the area of interest in PANs, detecting and classifying teeth into categories such as incisors, canines, premolars, and molars. Additionally, they sought to classify tooth condition, distinguishing between healthy teeth and those with non-metallic, light metallic, or complete metallic restoration. Employing the Canny edge detector and contrast filters, with various manual adjustments and post-processing refinements, they isolated the dentomaxilo region by locating the mandible line and the highest point of the hard palate. For the tasks of tooth detection and classification, they harnessed the power of the GoogleNet and ResNet networks. With a training set comprising 100 PANs, the authors achieved a sensitivity of 96.4% for detection, and accuracy rates of 93.2% and 98.0% for tooth type and tooth condition classification, respectively. The authors did not furnish metrics pertaining to the detection of the dentomaxilo region.

Also in C5, the authors (ESTAI et al., 2022) set out to detect the dentomaxilo region and, from within this area, identify and enumerate teeth with the FDI nomenclature. To annotate the dentomaxilo region, they utilized the space corresponding to the union of tooth bounding boxes as a teeth region annotation (similar to the dentomaxilo region). Employing a U-Net, they segmented this defined region. Subsequently, a Faster R-CNN was deployed to detect and enumerate each tooth within this segmented area. Leveraging a dataset comprising 591 PANs, the authors achieved recall and precision metrics for tooth detection (0.9919 and 0.9936) and tooth numbering (0.9803 and 0.98). However, it's noteworthy that the authors did not provide metrics related to the detection of the dentomaxilo region.

Still in C5 category, the authors (SILVA et al., 2020) conducted instance segmentation and tooth numbering on PANs using DL techniques. They compared the performance of the Mask R-CNN, Hybrid Task Cascade (HTC) (LIU et al., 2018a), Split-Attention Network (ResNeSt) (ZHANG et al., 2020), and Path Aggregation Network (PANet) (LIU et al., 2018b) networks on a dataset consisting of 543 images. The PANet network delivered the most promising results, achieving a mean Average Precision (mAP) of 71.9 for segmentation and 74.0 for numbering, using the FDI system.

Another C5 paper (PINHEIRO et al., 2021) conducted a comparative analysis between two networks, Mask R-CNN and Mask R-CNN with the PointRend module, for the segmentation and numbering (FDI) of permanent and deciduous teeth in PANs. In a dataset containing 874 images with individually segmented teeth, Mask R-CNN with PointRend achieved results of 77.3 mAP for segmentation and 75.3 mAP for numbering.

Below, we have compiled some general trends and discussions observed after analyzing all 69 articles that constitute the SM.

- Upon examining the publication dates of the 69 articles, a significant increase in interest in the subject in recent years becomes evident. This is reflected in the growth rate of publications between 2017 and 2022, with the latter year witnessing the highest number of articles published (20 papers). This trend suggests that 2023 and the subsequent years may experience even more extensive research and publications focused on the application of DL in Dentistry.

- (Related to Research Question - RQ - 1) The majority of these articles (66.667%) utilize PANs. In terms of image origin, most papers employ proprietary datasets, often without divulging them (75.362%). Additionally, not all articles provide explicit details regarding ethical approvals for data utilization or the annotation processes, including the amount and experience of the annotators. This lack of detailed information in scientific research publications increases the risk of biases, raises doubts about result reliability, and impedes the replication of experiments and progress in the state of the art.

- (Related to RQs 2, 3 and 4) Concerning the DL networks applied, the segmentation networks U-Net (27.536%) and Mask R-CNN (18.841%) have the highest frequency of use. In contrast, specialized detection networks such as Faster R-CNN and YOLO make fewer appearances. This observation suggests that despite segmentation being a more complex task than detection, researchers are opting for segmentation due to its capacity to provide richer spatial information, which can be useful in future applications.

- (Related to RQ 5) The prevailing performance metrics adopted, in general, for result reporting are, in order, accuracy, recall, and precision. Only 23.188% of the analyzed papers adhere to some form of guideline, such as STARD (Standards for Reporting Diagnostic accuracy Studies), CLAIM (Checklist for Artificial Intelligence in Medical Imaging), or Checklist AI (Checklist for Artificial Intelligence in Dental Research), to standardize the reporting of methodologies and results. While this percentage may appear relatively low, there appears to be a growing trend in the adoption of guidelines from 2019 to 2022.

2.3 Final Remarks

This chapter has outlined the planning, execution, and outcomes of a Systematic Mapping designed to explore the current state of the art for teeth detection/segmentation and tooth numbering. There has been a consistent growth in the number of publications in this field in recent years. The majority of articles identified in the SM employ PANs, utilize proprietary datasets, and offer limited information about the annotation process. Mask R-CNN and UNet, which focus on segmentation, are the most frequently used networks, while accuracy, recall, and precision are the preferred evaluation metrics.

Out of all the studies reviewed, only four were focused on tooth instance segmentation, with others integrating this task as an intermediate component of their research. The most notable results achieved were a precision of 0.969 (LEITE et al., 2021) and an mAP of 77.3 (PINHEIRO et al., 2021). There remains potential for further enhancements in the precision and mAP of tooth segmentation models by implementing innovative strategies. This ongoing progress promises to deliver a valuable tool for diagnostic support, epidemiological research, and various other applications.

In this chapter, we elucidated the execution methodology of the SM, presented its RQs and key trends and discussions observed after answering these RQs. Our comprehensive article on this SM, titled “Deep Learning to Detect and Classify Teeth, Dental Caries, and Restorations: A Systematic Mapping”, is presently undergoing peer review for the Journal of Dentomaxillofacial Radiology. It contains our detailed findings for each of these RQs. The trends and discussions presented here are the ones that contributed to the strategic decisions for this research:

- We opted to work with panoramic radiographs (PANs), driven by their prevalence in this mapping (66.667%). In addition, we chose to work with PANs because they contain all teeth in the dentomaxilo region in a single image, coupled with the advantages of cost-effectiveness and minimal radiation exposure.

- We chose to test the performance of the Mask R-CNN network because, among other reasons, it is the second most common DL network found in the mapping (18.841%). We did not apply the U-Net network because it is focused on semantic segmentation. Other reasons for choosing the Mask R-CNN include its open-source implementation and good performance in international challenges focused on object instance segmentation.
- We adopted the recall and precision metrics in our work, two of the most commonly found metrics. Additionally, we incorporated the Average Precision metric for its adoption in assessing image detection and segmentation networks (ZOU et al., 2023), along with the F1-score metric, which strikes a balance between recall and precision.

We offer a concise overview of the four articles classified as C4, focusing on Tooth Instance Segmentation, in our SM. Moreover, we include summaries of other articles used either as sources of inspiration or for comparative analysis alongside our results. For a comprehensive exploration of articles across all categories of the SM, please refer to our full-length article(CARNEIRO et al., 2023).

3 Materials and Methods

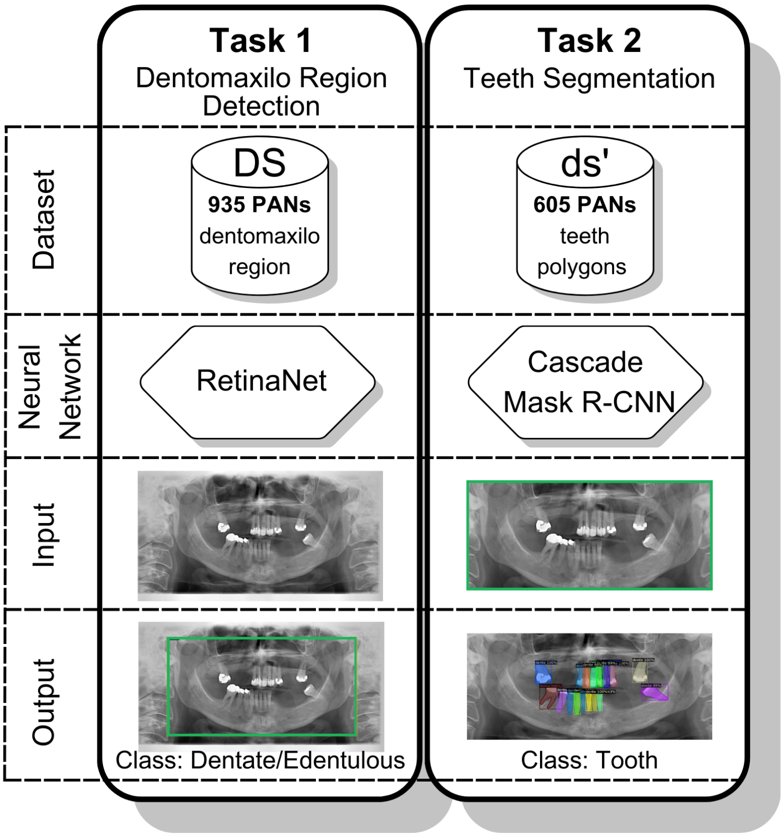
Our objective is the precise segmentation of teeth in panoramic radiograph (PANs), where dentomaxilo region detection (including mandible, maxilla and teeth) serves as an initial step to eliminate surrounding areas from the image. We aim to achieve the highest performance in mean Average Precision (mAP) for both detection and segmentation tasks. To attain this objective, we have designed and developed a modular system that combines two distinct Deep Learning (DL) networks. The utilization of two networks in combination to accomplish a goal is a prevalent approach in the literature. While a single network may suffice for the proposed task, recent studies have demonstrated significant potential by breaking down the overarching objective into smaller tasks and employing specialized networks for each (MURAMATSU et al., 2020; LEITE et al., 2021; ESTAI et al., 2022).

Figure 5 illustrates the project's design, with dentomaxilo region detection and teeth segmentation presented as two separate tasks. For each task, we specify the dataset used, the applied network, and provide an example of input and output for that network.

As presented in Section 1.1, PANs capture all teeth within a single image, including the mandible, maxilla, and parts of facial bones (SILVA; OLIVEIRA; PITHON, 2018), and that is why we chose to work with this type of dental image, a choice observed in most related works, as presented in Section 2.3. Given our focus on teeth segmentation, any regions outside the dentomaxilo area are irrelevant. Hence, *Task 1* aims to detect the dentomaxilo region (utilizing a bounding box) and remove all PAN portions beyond this region of interest. This task was influenced by the findings in (ESTAI et al., 2022), which showcase a notable reduction in false positives during tooth detection by eliminating the external PAN regions. To execute this task, we utilized 935 PANs (indicated in Figure 5 as DS) and the RetinaNet network (LIN et al., 2017).

Expanding upon the dentomaxilo area established in the previous task, *Task 2* focuses on the precise identification and delineation of each tooth's edges. This pixel-level contour can be applied to various future applications, including tooth pathology diagnosis. In such scenarios, segmentation may outperforms detection as it allows the exclusion of non-tooth regions, such as adjacent jaw and neighboring teeth. Moreover, segmentation inherently encompasses detection since bounding boxes can be derived from the outermost pixels of the segmented object. We applied the Cascade Mask R-CNN network (CAI; VASCONCELOS, 2018) and employed 605 segmented PANs to accomplish this task (in Figure 5, the dataset used is denoted as ds' since it is a manipulated subset of the original dataset DS).

Figure 5 – Project design

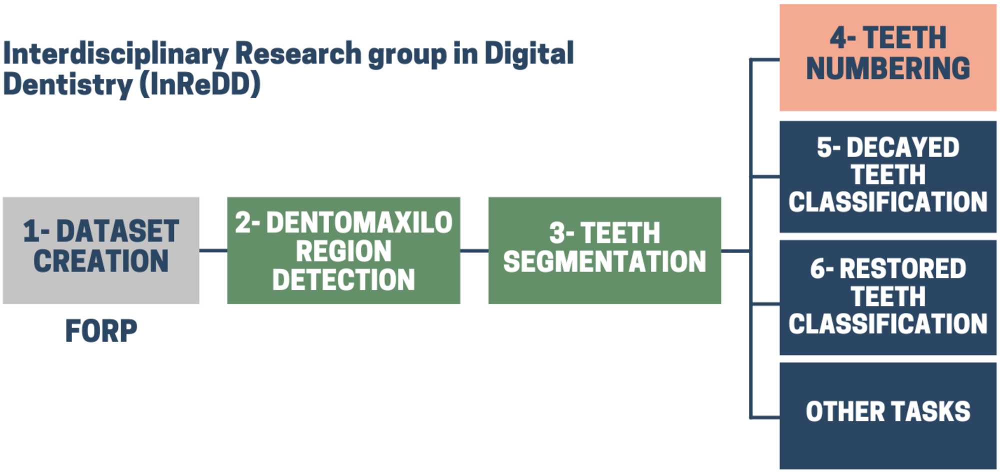


Source: Author's collection

This project is part of a multidisciplinary research initiative at the University of São Paulo (USP) known as InReDD (Interdisciplinary Research Group in Digital Dentistry). The group is divided into two cores: one focused on computer science within the Department of Computer Science and Mathematics (DCM) at the Faculty of Philosophy, Sciences, and Letters in Ribeirão Preto (FFCLRP - USP), and the other dedicated to dentistry and radiology within the Department of Dental Materials and Prosthodontics (DMDP) and the Department of Stomatology, Public Health, and Legal Dentistry (DESCOL) at the Faculty of Dentistry in Ribeirão Preto (FORP - USP). The group includes professors, postdoctoral researchers, doctoral and master's students, as well as undergraduate research students.

InReDD's objective is to develop a comprehensive system encompassing segmentation, classification, and numbering tasks. This system is designed to: a) detect the dentomaxilo area, b) segment teeth in radiographic images, c) numbering the detected teeth, d) identify missing teeth, e) identifying carious teeth, and f) identifying restored teeth in PANs. Figure 6 presents a visual representation of the project's architecture. The specific project outlined in this document focuses on the initial tasks of (a) dentomaxilo region detection and (b) teeth segmentation (depicted as stages 2 and 3 in Figure 6). Another component of this solution involves numbering teeth according to the nomenclature established by the Federation Dentaire Internationale (FDI), represented as stage 4. This stage is the focus of the research conducted by the master's student Breno Augusto Guerra Zancan (ZANCAN, 2023). These stages together aim to develop a tool capable of generating initial reports on teeth present in PANs. Utilizing the outcomes of this tool, various specialized applications can be developed. Examples include the detection of carious teeth (Stage 5) and restored teeth (Stage 6), among numerous other potential applications. The research dataset employed in this project was meticulously created by FORP dental and radiology professionals and is identified as stage 1 in Figure 6.

Figure 6 – InReDD project diagram


Interdisciplinary Research group in Digital Dentistry (InReDD)
1- DATASET CREATION
FORP
2- DENTOMAXILO REGION DETECTION
3- TEETH SEGMENTATION
4- TEETH NUMBERING
5- DECAYED TEETH CLASSIFICATION
6- RESTORED TEETH CLASSIFICATION
OTHER TASKS

Source: Author's collection

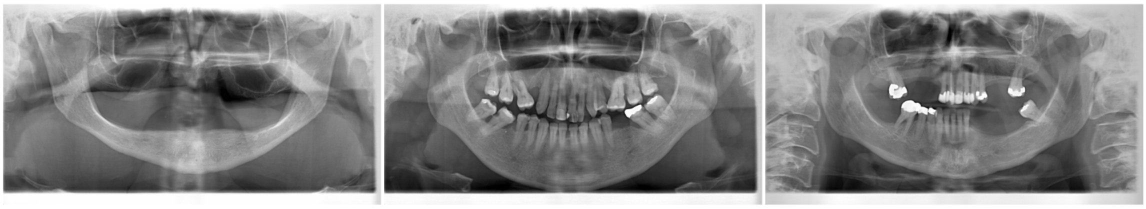
The following sections will provide a detailed description of the dataset used, the definition of the implemented DL networks, the experimental procedures, and the employed metrics to measure the performance of the experiments.

3.1 Dataset

Our system's initial dataset comprises 935 PANs compiled by the dentists and radiologists of InReDD. This dataset underwent approval by the local Research Ethics Committee (Plataforma Brasil, CAAE: 51238021.2.0000.5419) prior to its utilization. Beyond its capacity to capture all teeth in a single image, our choice to work with PANs was influenced by their prevalence in analogous studies, as outlined in Section 2.3. The process of generating the dataset, including image selection and annotation, has been extensively documented by the dental professionals within the InReDD group in a submitted article that is currently undergoing peer review at Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology (COSTA et al., 2023). In the subsequent paragraphs, we will outline key aspects of the dataset's composition and the annotation protocol employed.

The dataset consists of PANs obtained from adult subjects aged 18 and above. These radiographs were randomly selected by a radiologist from the clinical image repository at FORP-USP. Figure 7 showcases three random samples of these PANs, one edentulous and two dentate. These radiographs were originally acquired for routine patient care and have the patient's consent for using the data in research. Sourced from the same clinical setting, these PANs were captured using the Veraviewepocs device by J. Morita in Japan. Different exposure settings were employed, tailored to each patient's unique characteristics. These collected PANs exhibit diagnostic quality, characterized by their adequate sharpness and contrast. The dataset excludes radiographs featuring deciduous or mixed dentition, supernumerary teeth, bone fractures, bone loss, images of recent surgical interventions, orthodontic appliances, dental implants or any other metalic appliance in bones, fixed dental prosthesis and bone lesions of any type within the maxillofacial region. Radiographs displaying low contrast, limited sharpness, or motion artifacts were likewise omitted from the dataset. These exclusions were implemented to create an dataset free from outliers and unwanted noise.

Figure 7 – Examples of PANs from the dataset



Source: Author's collection

Following this selection process, the chosen images underwent anonymization and were stored in a PACS system, namely LyriaPacs - I-Medsys - Innovative Medical Informatics software [1]. The radiographs were stored in their original JPEG format, maintaining their dimensions of 2903 x 1536 pixels and 300 dpi resolution. No adjustments, such as brightness, contrast, cropping, or resizing, were applied during this phase.

The dataset underwent annotation in two distinct cycles, conducted as follows:

[1] LyriaPacs. Available at http://lyria.i-medsys.com/lyriaViewer-web/

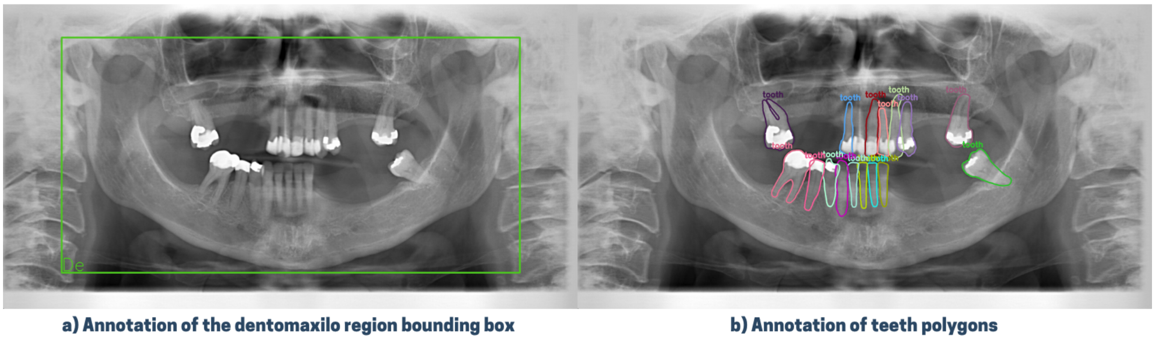
Initially, all 935 PANs received bounding box annotations. This task was accomplished by three radiologists, each having a decade of experience. One radiologist individually labeled the dentomaxilo region (comprising maxilla, mandible and teeth) in the radiograph. The bounding box borders for the dentomaxilo region were limited at right and left by the most prominent anatomic landmark, (at the angle of mandible or the mandibular condyle) and similarly at bottom (the lowest region of the mandibular shymphisis) and top (at upper region of the mandibular condyle). Subsequently, a second radiologist independently validated the annotations. In discrepancy cases, a third radiologist was consulted to arbitrate. Dentomaxilo regions were categorized as either dentate, edentulous (no teeth), edentulous mandible (no teeth in the mandible), or edentulous maxillae (no teeth in the maxillae). Figure 8 a) illustrates an example of a PAN with annotations generated in this initial cycle. In this example, the dentomaxilo region is classified as "dentate", indicated by the letters "De" on the image. The outcome of this initial cycle yielded a dataset with 935 bounding boxes outlining the dentomaxilo area, a resource used in *Task 1* (as depicted in Figure 5).

Subsequently, two radiologists individually segmented the teeth within a random subset of 605 PANs, excluding those with edentulous mouths (120 images). The remaining 210 PANs categorized as dentate, edentulous mandible or edentulous maxillae in the first annotation cycle were not segmented due to insufficient time before the project deadline. Employing the open-source annotation software, LabelMe [2], these radiologists delineated the tooth contours pixel-wise, following the labels generated during the first annotation cycle. In this segment of 605 images, the radiologists successfully labeled a total of 14,582 teeth. Figure 8 b) presents an example of a PAN with annotations generated during this second cycle. In this instance, all teeth were outlined with polygons and classified as "tooth", as indicated in the image. In instances of tooth overlap (such as when the crowns of two teeth slightly intersect due to misalignment and/or the X-ray projection orientation) annotators generated polygons that overlap while preserving the potential complete shape of each tooth, adopting approaches from the dataset provided by (PINHEIRO et al., 2021). Consequently, a given pixel in the image may be associated with more than one object. Figure 9 provides a fragment of the annotated example showcased in Figure 8. This close-up highlights the creation of overlapping polygons, aiming to capture the actual shape of each tooth. This second cycle created 605 segmented PANs, applied on the implementation of *Task 2* (as shown in Figure 5).

It is important to state that the annotation protocol was collaboratively developed by professionals from the fields of Radiology, Dentistry, and Computing within the InReDD group, following discussions and calibration procedures. This protocol was crafted to ensure the annotators' alignment and to generate datasets that are robust,

[2] LabelMe. Available at http://labelme.csail.mit.edu

Figure 8 – Examples of annotations visually represented on PANs

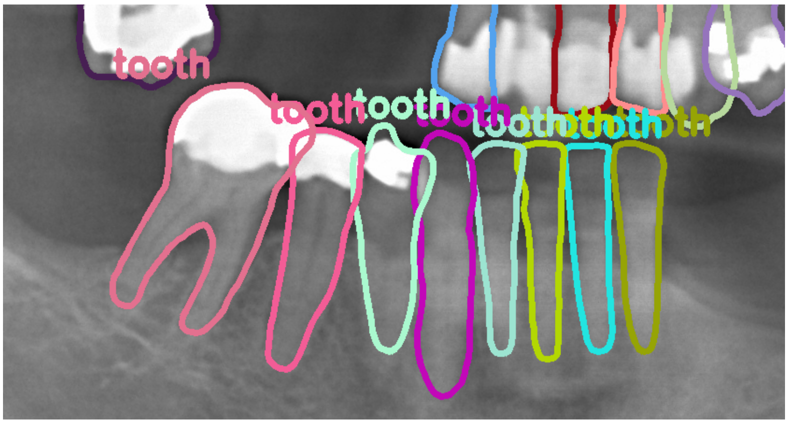


a) Annotation of the dentomaxilo region bounding box

b) Annotation of teeth polygons

Source: Author's collection

Figure 9 – Examples of overlapping annotations



Source: Author's collection

manipulable, operationally feasible, and user-friendly.

3.2 Dentomaxilo Region Detection

PAN captures all teeth, the jaw, and a portion of facial bones in a single image (SILVA; OLIVEIRA; PITHON, 2018). However, for precise teeth segmentation, only the dentomaxilo region holds relevance. Therefore, the first task focuses on detecting and cropping the dentomaxilo area within the PAN, discarding the non-essential elements outside this area. This preparatory task allows the next step to learn how to segment teeth with only relevant information, reducing computational costs by decreasing the image size. The research carried out by (ESTAI et al., 2022) validated that excluding this region prevents the network from erroneously detecting teeth in improbable locations (false positives). This study served as a source of inspiration for our work.

All 935 images in the dataset originally have dimensions of 2903x1536 pixels, with 8-bit grayscale intensity. The average size of the bounding boxes delimiting the dentomaxilo area annotated by experts in all these images is 2444x1194. Considering this average size, by cropping only the dentomaxilo region, approximately 35% of the original image size is eliminated, without using resizing techniques.

The equipment used to generate PANs is designed to keep the patient in the same position, often utilizing fixed reference points like the forehead or chin, varying by manufacturer. Although this technique aims to produce relatively standardized images, the patient's size, bone structure, or other characteristics cause slight variations in the position and size of the dentomaxilo area within the image. These variations impede the predetermined or standardized cropping of the dentomaxilo region. Therefore, to perform the dynamic detection of the dentomaxilo area, the application of a specialized DL network for object detection was proposed.

(MURAMATSU et al., 2020) proposes a semi-automatic dentomaxilo region detection technique. The method identifies the mandible line using the Canny edge detector and locates the highest point of the hard palate based on image contrast. By combining these markings, the dentomaxilo area is defined and cropped. However, to detect the mandible line without other contours causing noise in the process, the author uses a fixed mask as a filter, generated from manual expert annotations (MURAMATSU et al., 2012). This method is complex because it uses a sequence of image processing methods, with different parameters and configurations. In light of this, we opted for DL networks, given their minimal preprocessing requirements. The authors (ESTAI et al., 2022) utilize the region corresponding to the union of tooth bounding boxes as a teeth region annotation (similar to the dentomaxilo region). Subsequently, they employ the DL network U-Net for teeth region segmentation and proceed to crop a bounding box aligned with the segmented area. As our goal is to cut a rectangle in the image that contains the teeth region, it was not necessary to use segmentation networks.

To detect dentomaxilo area, we utilized Detectron2[3], an open-source Python library that implements the Faster R-CNN network (REN et al., 2017). Developed by the Facebook Artificial Intelligence Research Group (FAIR) and built using PyTorch and Cuda, Detectron2 provides open access code on GitHub, simplifying the implementation of DL models and the reproducibility of research efforts. Moreover, the library offers pretrained models on Common Objects in Context - COCO - instance segmentation dataset[4], speeding up the network training by using transfer learning technique. To compare the Faster R-CNN network's performance in dentomaxilo region detection, we also incorporated the RetinaNet network (LIN et al., 2017). In contrast to its predecessor, the RetinaNet network functions as a one-stage object detection model. Detectron2 library also includes pretrained models and weights for the RetinaNet network, enabling

[3] Detectron2. Available at https://github.com/facebookresearch/detectron2

[4] COCO dataset. Available at https://cocodataset.org/#home

comprehensive comparative analysis.

3.3 Teeth Segmentation

The primary objective of this project is tooth segmentation. Instead of using the entire PAN, we employ the cropped dentomaxilo regions as the network's input to reduce computational cost and enhance the network's convergence by eliminating unnecessary information.

Object segmentation is a step beyond object detection. It generates a mask that outlines the object within the detected bounding box, at a pixel level. While detection networks attempts to create the smallest bounding box possible to encompass the object, this often includes many background pixels. Segmentation removes this information that doesn't belong to the specific object being looked for.

We observed from the results of our Systematic Mapping (SM), presented in Section 2.2, that the U-Net and Mask R-CNN networks are the most popular among the studies found. Some studies show good segmentation results using U-Net (KOCH et al., 2019; BAYDAR et al., 2023; ARI et al., 2022). However, U-Net is capable of performing semantic segmentation instead of instance segmentation. In other words, the network classifies all pixels in the image and groups objects of the same class into a single mask. To apply it for the InReDD group purpose of subsequently numbering each of the segmented teeth, we would need to implement post-processing steps to separate each tooth from the segmented mask. In contrast, Mask R-CNN performs instance segmentation and is also present in studies with good results (CHANG et al., 2020; LI et al., 2021; VINAYAHALINGAM et al., 2021; RASHID et al., 2022).

As a result, in our tooth segmentation task, we compare the original Mask R-CNN (HE et al., 2017) with a modified version, called Cascade R-CNN (CAI; VASCONCELOS, 2018), both designed for instance segmentation. These networks are both implemented in Detectron2 and are available as open-source code on GitHub, complete with pretrained weights on COCO dataset and also Large Vocabulary Instance Segmentation (LVIS) dataset[5].

3.4 Experiments

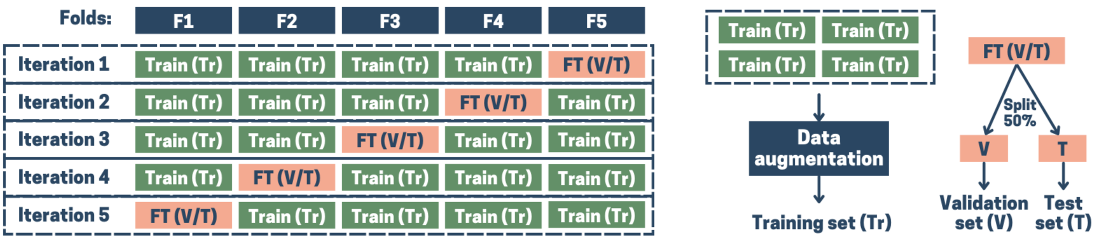
We have designed and executed a comprehensive testing plan to evaluate the mAP of pretrained DL models within our specific context. The primary objective was to iden-

[5] LVIS dataset. Available at https://www.lvisdataset.org

tify the optimal networks that could deliver peak performance given the constraints of our available data. In the forthcoming sections, we will show our dataset organization for training, validation, and test, as well as elucidating the data preprocessing and augmentation methodologies employed. Moreover, we will explain the training and testing procedures and present expected outcomes for each model.

Concerning dataset preparation, we adhered the five-fold cross-validation as the standard practice for measuring the models. To prepare the datasets for both tasks (dentomaxilo region detection and teeth segmentation) we applied the same cross-validation division. For elucidation purposes, we denominated the original image dataset as “E” for each task. Subsequently, “E” was partitioned into five near-equitable subsets, or folds, named “F1, F2, F3, F4, and F5”. This division ensured the absence of duplicate images across the folds. This cross-validation procedure underwent five iterations, where each iteration applied one fold, denoted as “FT”, for the roles of validation (“V”) and testing (“T”). Simultaneously, the remaining four folds were enlisted for training (“Tr”). The “V” and “T” partition signified that 50% of the images from the total “FT” fold were allocated to the validation set “V”, with the remaining 50% constituting the testing set “T”. Figure 10 illustrates the dataset division into five folds, ensuring an equal number of images in each fold, with no repetition across folds. The blue rectangles (F1 to F5) represent these folds. The figure demonstrates an example of the five iterations comprising the five-fold cross-validation. In each iteration, the 4 folds designated for network training are depicted in green and labeled as “Train (Tr).” These 4 folds collectively undergo the data augmentation process described in the subsequent paragraph, forming the training set (Tr). The pink fold, identified as “FT (V/T),” is divided into two sets, one for validation and another for testing, as illustrated on the right side of the image. These two sets (“V” and “T”) do not go through the data augmentation process

Figure 10 – Folds division

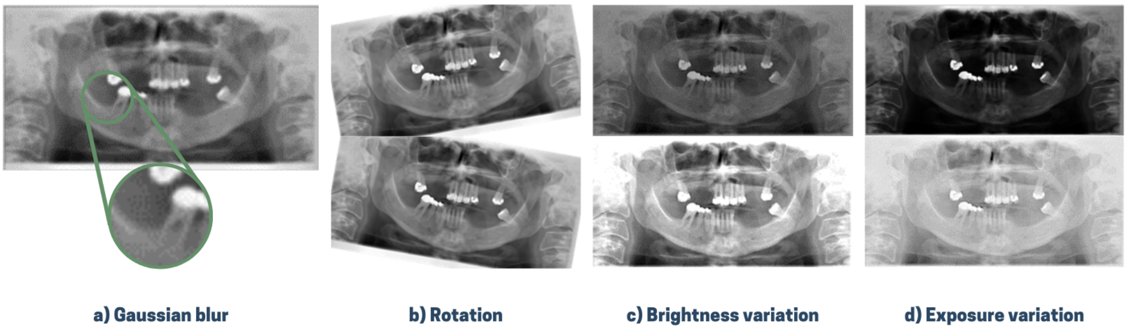


Folds:	F1	F2	F3	F4	F5
Iteration 1	Train (Tr)	Train (Tr)	Train (Tr)	Train (Tr)	FT (V/T)
Iteration 2	Train (Tr)	Train (Tr)	Train (Tr)	FT (V/T)	Train (Tr)
Iteration 3	Train (Tr)	Train (Tr)	FT (V/T)	Train (Tr)	Train (Tr)
Iteration 4	Train (Tr)	FT (V/T)	Train (Tr)	Train (Tr)	Train (Tr)
Iteration 5	FT (V/T)	Train (Tr)	Train (Tr)	Train (Tr)	Train (Tr)

Source: Author’s collection

To augment the amount of PANs for training, we employed a range of data augmentation techniques to all training sets “Tr”. Specifically, two augmented images were generated for each original image. Our data augmentation strategy was developed by adopting ideas from the research of (LEE et al., 2020; CANTU et al., 2020). This strategy encompassed the application of Gaussian blur, rotation and variations in brightness and exposure intensity. Figure 11 presents samples of augmented PANs, generated by applying each of the specified data augmentation techniques. These augmented images were exclusively derived from the original images within the training set (Tr) for each fold. They were applied only during training. No images from the validation (V) and test (T) sets were involved in the data augmentation process for their respective folds, and no augmented images were incorporated into these subsets. The five-folds creation and augmentation process was facilitated through the use of the Roboflow tool[6]. We opted to generate our folds using the external tool Roboflow. As we'll explore in the following sections, we are testing various models for the execution of *Tasks 1 and 2* in our research. External fold generation ensures consistency in using identical training, validation, and test sets for comparing model performance. Beyond employing these augmentation techniques through Roboflow, the Detectron2 library automatically incorporates the horizontal flip technique on training images with a random frequency. Given that all models used in this research were implemented using this library, the horizontal flip technique can also be considered part of our data augmentation strategy. The configuration parameters entered into Roboflow to generate the augmented images were: rotation angle into the range of -3º to +3º, Gaussian blur filter with a maximum intensity of 1 pixel, and adjustments in brightness and exposure intensity ranging from -10% to +10%. In conclusion, we adopted the same data augmentation strategy for both tasks (detecting the dentomaxilo region and segmenting teeth). Despite the distinct nature of these tasks, the images used in both training and subsequent inferences are highly similar. The input for one network is essentially a cropped region of the input from the preceding network.

Figure 11 – Samples of the data augmentation technique



Source: Author's collection

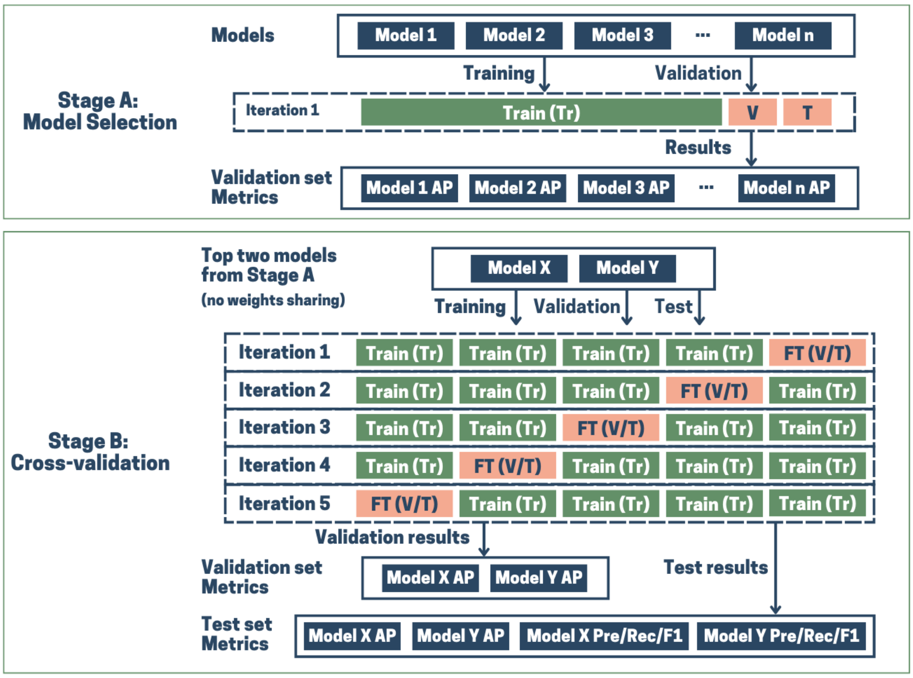
With regard to our regimen of experiments, both tasks (dentomaxilo region detection and teeth segmentation) underwent two sequential training phases. These two phases, identified as *Stage A* and *Stage B*, are described below and visually represented in Figure 12.

[6] Roboflow. Available at roboflow.com

1. *Stage A* “Model selection”: In this initial phase, we conducted experiments with various network configurations available in the Detectron2 library. This investigation focused solely on Iteration 1 to identify the two best models for each task. As illustrated in the upper rectangle of Figure 12, each model was trained on the Iteration 1 training set (Tr), and its performance was assessed using the mAP metric on the corresponding Validation set (V). At this stage, the other 4 cross-validation iterations were not used, as well as the test set (T) from Iteration 1..

2. *Stage B* “Cross-validation”: Having identified the top two models in *Stage A*, we proceeded to the five-fold cross-validation phase. In this stage, we rigorously tested the selected models across all five folds, training each model five times, once in each iteration. The primary objective was to validate their performance, thus confirming their ability to generalize effectively to unseen data.As shown in the bottom rectangle of Figure 12, in each iteration, we trained the model with its respective training set (Tr) and monitored its learning using the validation set (V) and the mAP metric. Following the completion of training, we evaluated the model’s performance in terms of mAP, precision, recall, and F1-score on the test set (T).

Figure 12 – Two stage sequence



Iteration 1	Train (Tr)	Train (Tr)	Train (Tr)	Train (Tr)	FT (V/T)
Iteration 2	Train (Tr)	Train (Tr)	Train (Tr)	FT (V/T)	Train (Tr)
Iteration 3	Train (Tr)	Train (Tr)	FT (V/T)	Train (Tr)	Train (Tr)
Iteration 4	Train (Tr)	FT (V/T)	Train (Tr)	Train (Tr)	Train (Tr)
Iteration 5	FT (V/T)	Train (Tr)	Train (Tr)	Train (Tr)	Train (Tr)

Source: Author’s collection

This systematic experimentation protocol facilitated the precise determination of the best models for our application context. Our utilization of cross-validation lent robustness to our findings, affirming the models’ efficacy in generalizing to previously unseen data.

3.4.1 Dentomaxilo Region Detection Experiments

For *Task 1*, we applied our initial dataset, comprising 935 PANs, and their corresponding bounding boxes delimiting the dentomaxilo area. These bounding boxes included the classes “edentulous” and “dentate”. Annotations initially categorized as “edentulous mandibles” and “edentulous maxillae” were reclassified as “dentate”. An illustrative example of the images used as network input can be observed in Figure 5, within the “Input” section of the first column. No preprocessing methods were applied to these images.

To construct the training, validation and test sets, all 935 images were utilized. The distribution of PANs across the sets was as follows: 80% (748 images) were allocated for training, while 10% (comprising 94 and 93 images, respectively) were assigned to both the validation and testing sets. The image distribution for *Task 1* among the training, validation, and testing sets on each cross-validation iteration is provided in Table 1. The augmentation techniques applied on the training set were rotation within a range of -3º to +3º, the application of Gaussian blur with a maximum intensity of 1 pixel, variations in brightness and exposure intensity spanning from -10% to +10%. The “Train” column in Table 1 outlines the number of augmented images used for training per iteration, with counts of dentate and edentulous images within each set indicated in parentheses.

Table 1 – Datasets for *Task 1*

Iteration	Train	Valid	Test
1	2244 (1938/306)	94 (81/13)	93 (77/16)
2	2244 (1913/331)	94 (85/9)	93 (77/16)
3	2244 (1938/306)	94 (79/15)	93 (76/17)
4	2244 (1919/325)	94 (82/12)	93 (80/13)
5	2244 (1900/344)	94 (86/8)	93 (82/11)

For our training and evaluation process, we harnessed the capabilities of the Faster R-CNN and RetinaNet detection networks available in the Detectron2 library’s Model Zoo module[7]. These models generated bounding box (BB) coordinates that delineated the dentomaxilo region, accompanied by class labels “Ed” (edentulous) or “De” (dentate) assigned to each bounding box. Figure 5, in the first column’s “Output” section, showcases an illustrative example of the outcomes produced by these networks.

To expedite the training process, we adopted transfer learning by leveraging pre-trained weights (trained on COCO dataset) for the backbone stage, provided by Detectron2, for all our models.

In *Stage A*, our primary objective was to explore various network configurations to identify the most effective models. We conducted training and evaluation on eleven distinct configurations of Faster R-CNN and RetinaNet models, exclusively utilizing the

[7] Detectron2 Model Zoo. Available at https://github.com/facebookresearch/detectron2/blob/main/MODEL_ZOO.md

folds of Iteration 1. This approach enabled us to pinpoint the optimal combination of backbone and bounding box head tailored to our specific context. The model configurations were sourced from Detectron2, encompassing three backbone networks (ResNet-50, ResNet-101, and ResNeXt), three distinct bounding box head setups (C4, DC5, and FPN), and two pre-trained weight strategies (1x - pre-trained for approximately 12 epochs on COCO dataset and 3x - about 37 epochs on COCO dataset). Each model went through a training regime consisting of 5,000 iterations, preserving Detectron2's default hyperparameters. Only two hyperparameters had their values changed from the default: Learning Rate, changed to 0.00025 and Batch Size, changed to 2 images per batch.

In *Stage B*, we selected the two most promising models identified in *Stage A* and put them through the five-fold cross-validation methodology, as elucidated earlier. Our top-performing Faster R-CNN model (ResNeXt backbone, FPN bounding box head, and pre-trained 3x weights) and RetinaNet model (ResNet101 backbone, FPN bounding box head, and pre-trained 3x weights) underwent training and evaluation across the five iterations of the cross-validation. We maintained a consistent LR of 0.00025 and a fixed BS of 2, while adhering to other hyperparameters in line with Detectron2's default settings. We also only used the same data augmentation strategy: rotation (-3º/3º), Gaussian blur (1px), brightness (-10%/10%) and exposure (-10%/10%). Throughout the training phase, we evaluated the models against the validation set every 100 iterations, tracking progress in terms of training loss, validation loss, and mAP. These metrics offered insights into the models' performance, indicating their efficacy in detecting and categorizing edentulous and dentate dentomaxilo regions. The Faster R-CNN training was extended to 15,000 iterations, and the RetinaNet model underwent 10,000 iterations of training. These iteration counts were determined through vigilant monitoring of the training and validation loss curves, aimed at preventing potential overfitting pitfalls. We interrupted the training process when we observed the training loss curve consistently decreasing, while the validation loss either plateaued or increased, signaling that the model was fitting the training data too closely and might not generalize well to new data.

3.4.2 Teeth Segmentation Experiments

In *Task 2*, we worked with a dataset comprising dentomaxilo regions extracted from 605 segmented images. These dentomaxilo regions were manually cropped based on expert-annotated bounding boxes, ensuring precise Ground Truth (GT) annotations during model training. We did not use the bounding boxes inferred by the DL models in *Task 1* during the training phase of this task. The training annotations for this task consisted of polygons outlining the contours of individual teeth, each defined by a set of coordinates. Every object within the dataset received the same label “Tooth”. Figure 5, in the “Input” field of the second column, provides an example of the dentomaxilo region crop used as input.

We divided the 605 segmented PANs into the training, validation and test sets, with 484 (80%), 61 (about 10%), and 60 (about 10%) images respectively. Table 2 shows the number of images and the number of teeth per set for each cross-validation iteration, with the second column indicating the number of training images after augmentation. On average, each image contained approximately 24 teeth. The dataset encompassed 4170 incisors, 2182 canines, 3764 premolars, and 4466 molars.

To expand the dataset size, we employed the same data augmentation strategy identified in *Task 1*, comprising the application of the following filters: rotation within the range of -3º to +3º, Gaussian blur with a maximum intensity of 1 pixel, brightness and exposure variations ranging from -10% to +10%. The Roboflow tool was employed to generate two new augmented images for each original image.

Table 2 – Datasets for *Task 2*

Iteration	Train	Valid	Test
1	1452 (35043)	61 (1474)	60 (1423)
2	1452 (34836)	61 (1529)	60 (1437)
3	1452 (34692)	61 (1516)	60 (1498)
4	1452 (35091)	61 (1463)	60 (1418)
5	1452 (35274)	61 (1471)	60 (1349)

For *Task 2*, we utilized the Mask R-CNN and Cascade Mask R-CNN models available in Detectron2. The prediction outputs of these models included a list of detected instances, each defined by four bounding box coordinates, a class label (solely “Tooth” in this case), and a binary matrix representing the object’s segmented mask inside the respective bounding box. Figure 5, under the “Output” field of the second column, illustrates an example of the predicted output, with teeth color-coded to indicate instance segmentation.

In *Stage A*, our primary goal was to experiment with various network configurations to identify the two best models for our application context. We conducted training, validation, and testing of thirteen distinct variations of Mask R-CNN and Cascade Mask R-CNN segmentation networks exclusively on the first iteration of cross-validation. The aim was to determine the most suitable combination of backbone and bounding box head for our specific context. All model variations were sourced from the Detectron2 library’s Model Zoo module[8]. Among the evaluated models, seven shared backbone networks (ResNet-50, ResNet-101, and ResNeXt) and head configurations (C4, DC5, and FPN). Additionally, we introduced specific variations: three models had backbone networks pre-trained using the Large Vocabulary Instance Segmentation (LVIS) public dataset

[8] Detectron2 Model Zoo. Available at https://github.com/facebookresearch/detectron2/blob/main/MODEL_ZOO.md

(version 0.5) for approximately 24 epochs, one model incorporated the Deformable Convolution (DAI et al., 2017) technique into the backbone network, another model employed the Group Normalization (GN) technique (WU; HE, 2018), requiring network retraining from scratch, and one model utilized the Cascade R-CNN network. Each model underwent training for a total of 5000 iterations. In all experiments, we adhered to the default Detectron2 hyperparameters, with the exception of LR (Learning Rate) and BS (Batch Size). For these specific parameters, we opted for the values determined in *Task 1*, namely 0.00025 for LR and a BS of 2 images per batch.

In *Stage B*, the most precise Mask R-CNN and Cascade Mask R-CNN models identified during the testing phase were further evaluated using a five-fold cross-validation approach. The top-performing Mask R-CNN model, featuring a ResNet50 backbone, FPN bounding box head, and retraining from scratch with GN, was trained for 5000 iterations. Meanwhile, the Cascade Mask R-CNN model, with a ResNet50 backbone, FPN bounding box head, and pre-trained weights, underwent 10000 iterations of training. Both models adhered to the default Detectron2 hyperparameters, with a fixed LR of 0.00025 and a BS set to 2. Similar to *Task 1*, these models were evaluated every 100 iterations, monitoring the training and validation loss curves for indications of overfitting. The models' learning progress was also observed through mAP growth. Additionally, transfer learning was applied to all models, focusing on the backbone pre-training stage, utilizing weights from training the models on the COCO or LVIS dataset.

3.5 Evaluation metrics

To evaluate the performance of our detection and segmentation networks, we conducted a comparison between the objects (dentomaxilo area and tooth) detected or segmented by the models, called "predictions" and the annotations provided by experts for all test set images, called "Ground Truth (GT)". We measured the results using four main performance metrics: precision; recall; F1-score and mean Average Precision (mAP). For the detection task mAP comprises two classes ("Dentate" and "Edentulous") and for the segmentation task, focused solely on the "Tooth" class.

The metrics recall, precision, and F1-score were selected given that these metrics were among the most frequently encountered in our mapping, as presented in Section 2.3. The mAP metric was selected as our primary evaluation criteria due to this extensive adoption in assessing image detection and segmentation networks (ZOU et al., 2023). Numerous object detection algorithms, including Faster R-CNN (REN et al., 2017), and YOLO (REDMON et al., 2016), utilize mAP as a key metric for evaluating their models. mAP is also a standard evaluation measure in various benchmark challenges, including

COCO[9], LVIS[10] and others. mAP is an object-level metric and takes into account the balance between the precision and recall of the model across various confidence levels, calculated through the area under the precision-recall curve. In addition to the assessment of prediction correctness, mAP also considers the quality of object localization and segmentation in a pixel-level, measured by Intersection over Union (IoU). The mAP metric considers predictions over a range of IoU levels to ensure that the final evaluation is not distorted by an arbitrarily selected IoU level.

In line with our experimental approach conducted in two stages (as detailed in Section 3.4), we employed the mAP metric to assess model performance during *Stage A* for both *Task 1* and *Task 2*, specifically on the validation set. During *Stage B*, mAP evaluation extended to both the validation and test sets. Furthermore, the metrics precision, recall, and F1-score were applied to the test set of *Stage B*. As *Stage A* focuses on model selection for cross-validation, precision, recall, and F1-score were not utilized in this stage.

To compute the four metrics employed in this study, it is essential to apply the definitions of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN), taking into account the model's predictions and the GT:

- True Positives (TP): The model correctly predicted a label that aligns with the GT;
- False Positives (FP): The model predicted a label that is not part of the GT;
- True Negatives (TN): The model did not predict the label, and it is also not part of the GT;
- False Negatives (FN): The model failed to predict a label that is, however, present in the GT.

Precision serves as a gauge for the accuracy of a model's predictions. It is calculated as the ratio of TP to the total number of positive predictions (TP and FP). A high precision value indicates that when the model predicts an object, it is likely to be correct. Recall, alternatively known as sensitivity or the true positive rate, assesses a model's capability to capture all instances outlined in the GT. It is computed as the ratio of TP to the total number of actual instances (TP and FN). The F1-score represents the harmonic mean of precision and recall.

Precision (P), recall (R) and F1-score (F1) are defined in Equations (3.1), (3.2) and (3.3).

$$P = \frac{TP}{TP + FP} \tag{3.1}$$

[9] COCO dataset. Available at https://cocodataset.org/#home

[10] LVIS dataset. Available at https://www.lvisdataset.org/challenge_2021

$$R = \frac{TP}{TP + FN} \tag{3.2}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{3.3}$$

To determine whether a prediction is considered TP, FP, or FN, we first evaluate the degree of overlap between each predicted bounding box or polygon (Pred) and the GT objects through the Intersection over Union (IoU) metric, defined by Equation (3.4):

$$IoU = \frac{AreaPred \cup AreaGT}{AreaPred \cap AreaGT} \tag{3.4}$$

The IoU value acts as the criterion (threshold) for determining whether a prediction is qualified as a true positive (TP). False positives (FP) denote inaccurate predictions, while false negatives (FN) signify instances within the GT that the model did not predict. To compute precision, recall, and F1-score in *Stage B*, a fixed threshold of 85% IoU was employed. A prediction was designated as a TP if it exhibited an overlap of 85% or more with its pixels (bounding box or polygon) compared to the GT. Instances with IoU below 85% were categorized as FP, while GT elements without any associated predictions were counted as FN.

For computing mAP, applied in *Stages A and B*, the procedure involves the following steps:

1. Input all images from the assessed set (validation or test set) into the model to obtain predictions;
2. Select a class for measurement;
3. Set an IoU value as a threshold;
4. Examine all model predictions on all images, comparing them with the GT. Classify these predictions as TP or FP based on the defined IoU;
5. Count the GT elements not detected as FN;
6. Construct an ordered list of all predictions, sorted in descending order based on the model's confidence level for each prediction (confidence level is a standard output of the models, indicating the certainty of a given prediction on a scale from 0 to 1);
7. Calculate precision and recall for the assessed set for each prediction, following the ordered list;
8. Plot precision and recall values on a line graph, using the precision-recall pairs as coordinates (precision-recall curve);

9. Compute AP, which is calculated as the average of the P values based on the area under the precision-recall curve, with R spanning a finite and discrete range of 11 values from R = 0.0 to 1.0, incremented by 0.1. Employing interpolation, calculate P(R) for each R value as the maximum P value within the interval [R', R], where R' denoted the preceding R value in the assessment. The formula for AP is defined in Equation (3.5).

$$AP = \frac{1}{11} \sum_{R=0}^{1} P(R) \tag{3.5}$$

AP represented the Average Precision for each class. In models with multiple classes, such as *Task 1* involving two classes (edentulous and dentate), mAP was calculated as the average of each class's AP. As highlighted earlier in this section, the mAP metric considers predictions across various IoU levels. Consequently, determining the model's final mAP value involves iteratively calculating AP, varying the IoU threshold. Following the standard used by the models (Faster R-CNN and YOLO) and challenges (COCO and LVIS) mentioned above, in this study, the model final mAP was computed as the arithmetic mean of mAP values, varying IoU between 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, and 0.95. Note that, for *Task 1*, this process was replicated for each class (edentulous and dentate). mAP was calculated for all *Stage A* models in the validation set and *Stage B* models in the validation and test sets.

4

Results and Discussion

In this section, we present the results achieved in each task of our research after training and validation. Furthermore, we conduct an in-depth analysis and discussion of the achieved results.

These results are the outcome of the process involving data collection and preparation, model training, and subsequent evaluation, as outlined in Chapter 3. Our research has successfully achieved its objective, which is to propose a modular system driven by Deep Learning (DL), designed for the automatic segmentation of teeth in Panoramic Radiographs (PANs). This system begins with a preprocessing step, referred to as *Task 1*, involving the detection of the dentomaxilo region (including mandible, maxilla and teeth) to exclude surrounding areas that do not contribute to the network analysis. The final step, known as *Task 2*, focuses on instance segmentation of the teeth using only the dentomaxilo area.

In the context of *Task 1*, Dentomaxilo Region Detection, we delineated bounding boxes and distinguished between "dentate" and "edentulous" dentomaxilo regions, deploying the Faster R-CNN and RetinaNet networks. We conducted experiments, comprising eleven distinct permutations of Faster R-CNN and RetinaNet models, exclusively on the first cross-validation iteration. The results are detailed in Table 3. The first column (Mod) characterizes each model, with designations starting with the letter D to denote detection. The second column (Network) outlines the network employed, with eight models utilizing Faster R-CNN and three models harnessing the capabilities of RetinaNet. The third column (Configuration) exhibits the configuration of each model, encompassing the backbone network, bounding box head, and the pre-trained weights routine. The fourth column (s/it) specifies the average time taken per iteration during the training process for each model, measured in seconds. Finally, the fifth (ValD mAP) column exhibits the mAP values calculated in the validation sets (inferred during the last phase of the validation process after training completion).

Among the observed models, the model D4 emerged as the faster in terms of training efficiency, averaging 0.503 seconds per iteration. It was closely followed by models D1, D9, and D10. It's noteworthy that deeper backbone networks necessitate more extensive training duration. Models implementing the ResNet50 architecture exhibited superior training speeds compared to their ResNet101 counterparts. The most time-consuming model was D8, utilizing the ResNeXt network, with an average iteration time of 2.481 seconds. Models embracing the DC5 configuration also demonstrated relatively slower training speeds.

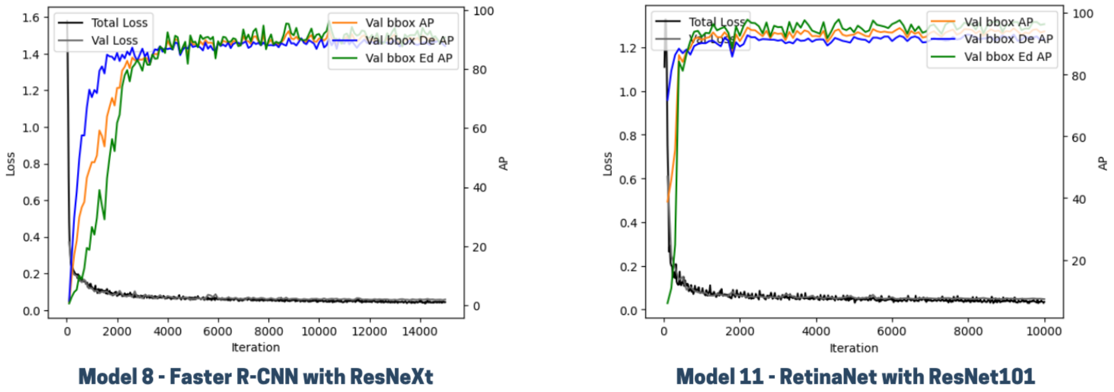
Furthermore, models D9, D10, and D11, encompassing the entire spectrum of RetinaNet networks, consistently outperformed their Faster R-CNN counterparts in terms of mAP. However, when preparing for the five-fold cross-validation in *Stage B*, we selected models D8 and D11, representing one each from Faster R-CNN and RetinaNet, respectively. This selection was made based on the highest mAP values recorded in the validation set, standing at 85.157 for D8 and an 91.908 for D11.

Table 3 – *Stage A* Model selection for *Task 1*

Mod	Network	Configuration	s / it	ValD mAP
D1	Faster	ResNet50 FPN 1x	0.526	67.984
D2	Faster	ResNet50 C4 3x	0.612	64.895
D3	Faster	ResNet50 DC5 3x	0.795	76.589
D4	Faster	ResNet50 FPN 3x	0.503	71.265
D5	Faster	ResNet101 C4 3x	0.822	66.781
D6	Faster	ResNet101 DC5 3x	1,087	64.658
D7	Faster	ResNet101 FPN 3x	0.712	65.785
D8	Faster	ResNeXt FPN 3x	2.481	85.157
D9	RetinaNet	ResNet50 FPN 1x	0.513	91.141
D10	RetinaNet	ResNet50 FPN 3x	0.520	91.716
D11	RetinaNet	ResNet101 FPN 3x	0.740	91.908

In *Stage B*, we conducted experiments employing the D8 and D11 models, utilizing the five-fold cross-validation approach. As previously outlined, D8 (Faster R-CNN) underwent an training regimen of 15,000 iterations, while D11 (RetinaNet) was subjected to 10,000 iterations, chosen through vigilant oversight to prevent overfitting. Figure 13 provides a visual representation of the training dynamics for models D8 and D11. The charts illustrate the models' loss values measured on the training set at each iteration ("Total Loss") and on the validation set every 100 iterations ("Val Loss"). The loss scale is depicted on the Y-axis to the left of the graph. Simultaneously, on the right side of the Y-axis, the evolution of mAP on the validation set is showcased every 100 training iterations. The mAP for each class ("Val bbox De AP" and "Val bbox Ed AP") and the average mAP ("Val bbox AP") are presented. Examining the mAP curves reveals a progressively slower improvement as the iterations progress. Furthermore, by observing the loss curves, an initial deviation between the training and validation losses becomes apparent, with the validation loss on an ascending trajectory and the training loss on a descending one. In light of this disparity and considering the modest enhancement in precision (mAP), the decision is made to conclude the training process.

Figure 13 – Training Progress Charts for Models D8 and D11 in *Stage B* of *Task 1*



Model 8 - Faster R-CNN with ResNeXt

Model 11 - RetinaNet with ResNet101

Source: Author's collection

The results of the *Stage B* experiments are presented in Table 4. Each data within the table represents the arithmetic mean of the outcomes obtained across the five folds. The table not only provides insights into the average time per iteration (s/it) and the mean Average Precision (mAP) values assessed in both the validation (ValD mAP) and test (TestD mAP) sets but also presents the average values for precision (Pre), recall (Rec), and F1-score (F1) across the five folds for the two models. Since *Task 1* involves a multiclass detection assignment, distinguishing between "dentate" and "edentulous" cases, we calculated the arithmetic mean of precision and recall across both classes, allowing us to derive the precision and recall metrics for each model.

Figure 14 displays 3 samples of dentomaxilo region detection generated by models D8 and D11 on the test set.

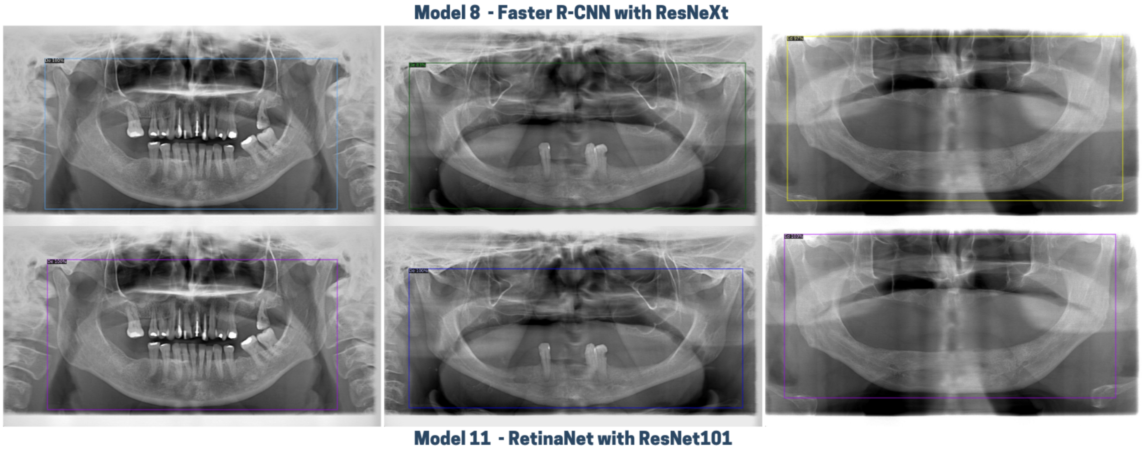
Table 4 – *Stage B* Results for *Task 1*

Mod	s / it	ValD mAP	TestD mAP	Pre	Rec	F1
D8	1.968	89.403	87.798	0.971	0.954	0.960
D11	0.750	92.147	92.446	0.971	0.994	0.982

Our observations demonstrated that D11 consistently outperformed D8 in terms of ValD mAP and TestD mAP. D8 achieved mAP values of 89.403 in the validation set and 87.798 in the test set, whereas the D11 model scored 92.147 and 92.446 in these metrics. It is noteworthy that the average time per iteration in D8 is significantly higher than in D11, representing the time required for the model to make predictions on a single image.

The superior performance of D11 is also evident in recall, and consequently, the F1-score. D8 achieved a recall of 0.954 and an F1-score of 0.960, while D11 exhibited a

Figure 14 – Samples of dentomaxilo region detection generated on *Stage B*


Model 8 - Faster R-CNN with ResNeXt
Model 11 - RetinaNet with ResNet101

Source: Author's collection

recall of 0.994 and an F1-score of 0.982. Both models achieved the same precision, with a value of 0.971. Both models marked improvement in this stage compared to *Stage A*, thanks to the largest number of iterations.

Table 1 presents that each cross-validation iteration contains 93 images in the test set, resulting in a total of 465 radiographs across all five test sets, with 392 being dentate and 73 edentulous. During the testing phase, models D8 and D11 processed these 465 radiographs in their respective test sets. Combining the predictions from these tests into a single analysis, the five iterations of the D8 model failed to detect 6 dentate and 5 edentulous dentomaxilo regions, resulting in a 2.366% false-negative rate. In contrast, the five iterations of the D11 model missed 4 dentate dentomaxilo regions but successfully identified all edentulous dentomaxilo regions, resulting in only a 0.860% false-negative rate.

It is worth noting that the task of separating the dentomaxilo area can be accomplished using simpler image processing techniques, as demonstrated in previous studies (MURAMATSU et al., 2020; MURAMATSU et al., 2012). These techniques might offer advantages in terms of time efficiency during the model setup and hyperparameter search phases compared to DL. However, once our model is trained, it provides rapid predictions (approximately 0.750 seconds per prediction), high accuracy (precision of 0.971), and consistency (0.860% false negatives) for new image predictions. Furthermore, by applying a similar approach to train the models used in detection and segmentation, replicating strategies and routines, we reduce planning time and achieve the final result more efficiently. The detection of the dentomaxilo area can be considered a preparatory task for segmentation, making these results an integral part of *Task 2*.

In *Task 2*: Teeth Segmentation, during *Stage A*, we deployed twelve unique Mask

R-CNN models alongside the sole Cascade Mask R-CNN version available in the Detectron2 library. Our primary goals in this phase were twofold: first, to identify the most effective Mask R-CNN configuration within our specific domain, and second, to assess whether Cascade Mask R-CNN delivered comparable results. These experiments were exclusively conducted using the first cross-validation iteration. Table 5 offers a comprehensive overview of these 13 network configurations in its second column (Configuration), detailing aspects such as the backbone network, bounding box head, pre-trained weight routine, and additional information (LVIS: backbone pre-trained on the Large Vocabulary Instance Segmentation - LVIS - dataset; DConv: backbone with deformable convolution; GN: group normalization). Each model is distinguished by an “S” character denoting instance segmentation, followed by a numerical identifier, as shown in the first column (Mod) of the table. The third column (s/it) presents the average time per iteration during the training process, while the subsequent two columns provide the metrics for network predictions. These metrics encompass bounding box generation in the validation set (ValD mAP) as well as polygon generation in the validation set (ValS mAP).

Among these models, S8 emerged as the fastest, boasting an average iteration time of 0.481 seconds per iteration, closely followed by S3 at 0.484 seconds per iteration. This outcome aligns with expectations, considering that both models share the same architecture, differing only in the pre-training of the backbone network weights (S8 was trained on the LVIS dataset, whereas S3 utilized the Common Objects in Context - COCO - instance segmentation dataset). In contrast, models S7 and S10 displayed slower performance due to their reliance on the ResNeXt backbone network. These time-related results are consistent with those observed in *Stage A* of Dentomaxilo Region Detection. Moreover, regarding the backbone network, the ranking from fastest to slowest is as follows: ResNet50, ResNet101, and ResNeXt. Concerning the bounding box head configuration, FPN proved to be the fastest, followed by C4 and DC5.

In terms of mAP values, S13 excelled in the validation set, followed by S12. Specifically, S13 achieved a detection mAP of 79.464 and a segmentation mAP of 80.611 in the validation set, whereas S12 attained a detection mAP of 78.463 and a segmentation mAP of 79.855 in the validation set. Among the Mask R-CNN models (S1 to S12), S12 consistently outperformed its peers across both mAP metrics. Consequently, both S12 and S13 were designated as the top-performing models for this task and were subsequently chosen for testing in *Stage B*, which involves five-fold cross-validation.

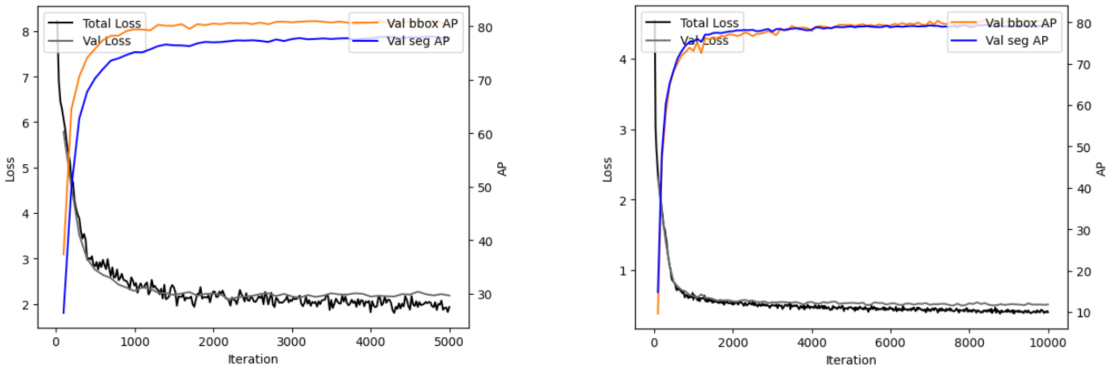
In line with *Task 1*, for *Stage B* of *Task 2*, we trained the selected models until the first signs of overfitting became apparent. S12 (Mask R-CNN with GN) underwent 5000 iterations of training, while S13 (Cascade Mask R-CNN) completed 10000 iterations. Figure 15 presents the training graphs for models S12 and S13. The graph illustrates the loss values of the models measured on the training set at each iteration (“Total Loss”) and

Table 5 – *Stage A* Model selection for *Task 2*

Mod	Configuration	s / it	ValD mAP	ValS mAP
S1	ResNet50 C4 3x	0.568	77.914	78.625
S2	ResNet50 DC5 3x	0.737	76.369	76.212
S3	ResNet50 FPN 3x	0.484	76.592	77.313
S4	ResNet101 C4 3x	0.757	76.298	77.860
S5	ResNet101 DC5 3x	0.966	76.675	77.671
S6	ResNet101 FPN 3x	0.665	76.984	78.042
S7	ResNeXt FPN 3x	2.109	76.812	78.454
S8	ResNet50 FPN 1X LVIS	0.481	74.897	76.132
S9	ResNet101 FPN 1X LVIS	0.656	76.796	77.681
S10	ResNeXt FPN 1X LVIS	2.184	77.741	78.095
S11	ResNet50 FPN 3x DConv	0.618	76.548	77.572
S12	ResNet50 FPN 9x GN	0.630	78.463	79.855
S13	Cascade ResNet50 FPN 3x	0.514	79.464	80.611

on the validation set every 100 iterations ("Val Loss"). The loss scale is depicted on the Y-axis to the left. Simultaneously, on the same graph, using the scale on the right side of the Y-axis, we showcase the evolution of mAP measured on the validation set every 100 training iterations. Specifically, mAP is displayed for the bounding box head ("Val bbox AP") and the mask head ("Val seg AP"). Similar to the patterns observed in the models of *Task 1* (Figure 13) and as anticipated, there is a progressively slower improvement in model precision (mAP) as the iterations advance. Analyzing the loss curves, a earlier and more pronounced divergence emerges between the validation loss, on an ascending trajectory, and the training error, on a descending path. This divergence, as reflected in the growing validation set loss, suggests that the model is starting to overfit the training data, compromising its ability to generalize. In *Task 2*, this overfitting indication is more apparent due to the task's heightened complexity and a constrained number of training images, compared to *Task 1*.

Figure 15 – Training Progress Charts for Models S12 and S13 in *Stage B* of *Task 2*



Model 12 - Mask R-CNN with ResNet50

Model 13 - Cascade Mask R-CNN with ResNet50

Source: Author's collection

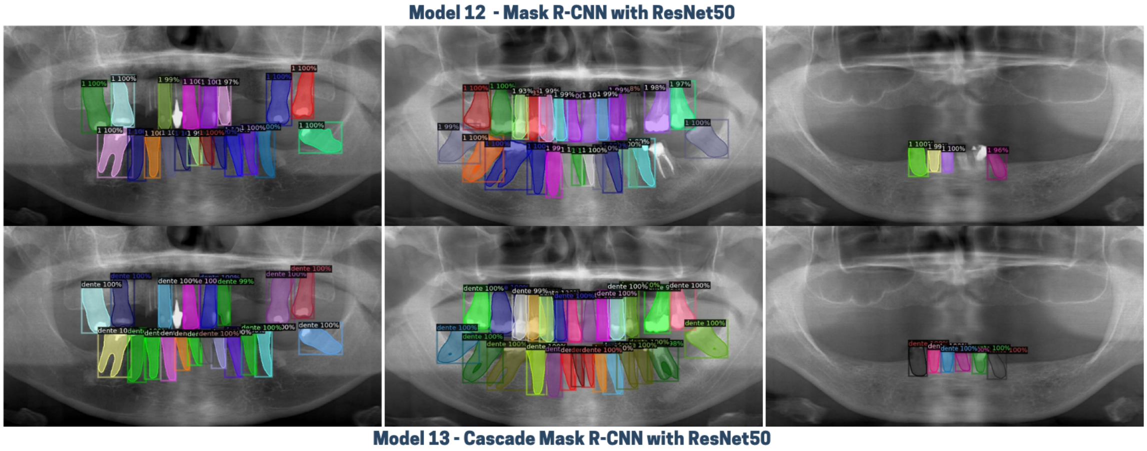
Table 6 presents the outcomes achieved by models S12 and S13 utilizing the five-fold cross-validation technique of *Task 2*'s *Stage B*. Each value in the table represents the arithmetic mean of the model's performance across the five cross-validation folds. This table closely follows the structure of Table 4, providing insights into time per iteration (s/it), bounding box mAP for both the validation (ValD mAP) and test sets (TestD mAP), precision (Pre), recall (Rec), and F1-score (F1). This task deals with a single class, eliminating the need for additional precision and recall calculations beyond the arithmetic mean of the model's fold-wise performance. Another noteworthy distinction concerns to the segmentation aspect of the task, where the table presents the results of bounding box predictions in terms of ValD mAP and TestD mAP, along with the model's performance in mask creation, assessed in both the validation and test sets. These results are meticulously detailed in the columns labeled (ValS mAP) and (TestS mAP), respectively.

Figure 16 displays 3 samples of teeth segmentation generated by models S12 and S13 on the test set.

Table 6 – *Stage B* Results for *Task 2*

Mod	s / it	ValD mAP	TestD mAP	ValS mAP	TestS mAP	Pre	Rec	F1
S12	0.744	81.124	78.291	81.324	77.999	0.986	0.988	0.987
S13	0.857	79.971	79.222	80.188	78.954	0.991	0.991	0.989

Figure 16 – Samples of teeth segmentation generated on *Stage B*


Model 12 - Mask R-CNN with ResNet50
Model 13 - Cascade Mask R-CNN with ResNet50

Source: Author's collection

Analyzing the data in Table 6, we note that S12 exhibits a slightly better average processing time than S13, with a marginal difference of 0.113 seconds per iteration.

In terms of performance metrics, S12 stands out in the validation set, boasting higher mAP scores (81.124 in detection mAP and 81.324 in segmentation mAP). Conversely, S13 outperforms S12 in the test set, achieving superior results (79.222 in detection mAP and 78.954 in segmentation mAP). Additionally, S13 demonstrates superior precision (0.991), recall (0.991), and consequently, a higher F1-score (0.989) compared to S12.

Each test set within the five iterations of the cross-validation comprises 60 images, encompassing an average of 1425 teeth, as outlined in Table 2. Cumulatively, across all five iterations, we evaluated a total of 350 images and 7125 teeth. Out of these 7125 teeth, S12 failed to detect 78, while S13 missed 75 (equivalent to 1.095% and 1.053% false-negative predictions, respectively). The most prevalent error scenarios involve overlapping teeth or teeth positioned closely to other bone structures. Additionally, teeth with pins and extensive restorations pose challenges for accurate prediction.

Considering the mAP values calculated in the test set, the F1-score, and the number of false negatives, Model 13, Cascade Mask R-CNN with a ResNet50 backbone, delivered the most favorable results (79.222 mAP for detection, 78.954 mAP for segmentation, and a 0.989 F1-score) for *Task 2*. Nevertheless, these results closely align with the performance achieved by the Mask R-CNN network using the ResNet50 backbone and the Group Normalization technique.

Some related studies have tackled the task of instance segmentation of teeth in PANs, and a concise summary of these works can be located in Section 2.2. In Table 7, we present a comparative analysis of our tooth segmentation results alongside those achieved in these related studies, highlighting the quantity of images used and the chosen network architecture. These papers report their findings using diverse sets of metrics. The papers (LEE et al., 2020; JADER et al., 2018; LEITE et al., 2021; SILVA et al., 2020) employ metrics such as precision, recall, and F1-score, while (PINHEIRO et al., 2021) and, once again, (SILVA et al., 2020) use the mAP metric. Across all these metrics, our study consistently demonstrates superior performance.

Table 7 – *Task 2* results comparison

Paper	#imgs	Network	mAP	Pre	Rec	F1
(LEE et al., 2020)	50	Mask R-CNN	-	0.858	0.893	0.875
(JADER et al., 2018)	1500	Mask R-CNN	-	0.94	0.84	0.88
(LEITE et al., 2021)	153	DeepLabV3 + FCN	-	0.969	0.983	0.975
(SILVA et al., 2020)	543	PaNet	71.3	0.944	0.891	0.916
(PINHEIRO et al., 2021)	450	Mask R-CNN PointRend	77.3	-	-	-
Our	605	Cascade Mask R-CNN	78.954	0.991	0.991	0.989

We chose not to compare our study with (SCHNEIDER et al., 2023), because the paper, despite also delving into tooth instance segmentation, does not explicitly disclose the F1-score achieved by the final model, after the Federated Learning step of parameters aggregation. Instead, it reports F1-scores for local models generated in the middle of the Federated Learning process. Similarly, we refrained from drawing comparisons with (CHEN et al., 2019) since that study concentrates solely on tooth detection, rather than segmentation, making it unsuitable for a direct comparison between different tasks. Nevertheless, even if such a comparison were feasible, our precision and recall values would still surpass those reported in the paper, specifically registering values of 0.988 and 0.985, respectively.

The study that achieved performance closest to ours was conducted by (PINHEIRO et al., 2021), achieving an mAP of 77.3. The authors utilized the Mask R-CNN network in combination with the PointRend module, a neural network designed to enhance the precision of segmented object boundaries. It's important to note that since each article used its own dataset, the comparison presented in Table 7 serves as a benchmark. A fair assessment of performance across different systems necessitates similar conditions, such as a standardized reference test collection.

We abstained from comparing the results of *Task 1* with related studies because similar papers often treat dentomaxilo region detection as a preliminary stage of segmentation and typically do not disclose the results of this initial phase. Consequently, they only present the final outcomes (ESTAI et al., 2022; MURAMATSU et al., 2020; MURAMATSU et al., 2012). Nevertheless, we have chosen to share the results obtained at all stages of our study.

5

Conclusion

Our approach revolves around a modular solution designed for tooth segmentation, preceded by the detection and classification of dentate and edentulous dentomaxilo regions (mandible, maxilla and teeth) in panoramic radiographs (PANs). Through the development of specialized and complementary components, we have successfully implemented two distinct models, each dedicated to its respective task, delivering satisfactory performance. In the task of dentomaxilo region detection, among the various models we evaluated, the RetinaNet with a ResNet101 backbone emerged as the top performer, achieving 92.446 mAP and 0.982 F1-score in the test set. Transitioning to tooth segmentation, we accomplished an mAP of 79.222 for detection, 78.954 for segmentation, and F1-score of 0.989 in the test set, leveraging the Cascade Mask R-CNN. These results not only met but surpassed the performance of similar studies (LEE et al., 2020; JADER et al., 2018; LEITE et al., 2021; SILVA et al., 2020; PINHEIRO et al., 2021).

Our findings substantiate the effectiveness of our modular solution, which harnesses specialized networks, proving to be as effective, if not more so, than a single neural network attempting to tackle all the tasks proposed within our system. It is important to highlight that this work also makes a contribution by introducing the intermediate stage of detecting the dentomaxilo region in the tooth segmentation process. This step involves cropping the dentomaxilo region, which, on average, eliminates 35% of pixels from the original PAN. This not only optimizes computational resources but also mitigates the risk of false positives during tooth segmentation, particularly in regions removed where the presence of a tooth is anatomically impossible. Another contribution of this research is the segmentation of teeth without the simultaneous enumeration attempt within the same neural network (a common approach in many existing studies). Instead, we designed a specialized module with the exclusive responsibility of tooth segmentation. Given the structural similarities among teeth, this approach enhances the network's generalization capability, allowing it to learn the fundamental characteristics of a tooth through exposure to a more diverse set of examples.

While our tests have yielded promising results, it is essential to acknowledge certain limitations associated with the nature and volume of the images used. The availability of a limited set of segmented images posed a constraint in our testing process. Despite implementing data augmentation strategies for both tasks, the extent of teeth segmentation performed by professionals significantly influenced the results. In contrast, some studies trained their models using images segmented by researchers from diverse fields outside of Dentistry or Radiology. Our approach stands out as all our images were meticulously annotated and checked by experienced professionals in these specialized domains. Additionally, the use of an unbalanced dataset for dentomaxilo area detection tasks presents a potential challenge to the generalization of results. Furthermore, we relied on images from a specific location, all generated using the same radiographic unit (Veraviewepocs from J Morita), inherently representing a confined community.

Despite these acknowledged limitations, our methodologies exhibit potential for future investigations in the context of diagnostic aids. To enhance the outcomes of the proposed methodology, implementing strategies such as expanding the dataset size, diversifying the dataset's sources (including data from various locations and different radiographic units), including images removed from our dataset based on exclusion criteria, and exploring explainable AI techniques like decision heat maps could be beneficial. Furthermore, evaluating the performance of our networks across various types of radiographs (periapical, bitewing, occlusal, CBCT, etc.) would provide valuable insights. Another interesting test would involve training the proposed segmentation networks on our dataset without the dentomaxilo region removal stage and compare their performance with the original approach. The modular structure of our solution allows for the integration of additional modules, each offering distinct functionalities within the system. As previously highlighted, this research is part of the InReDD group. Within this research group, active work is already underway on the subsequent module, which involves assigning numerical identifiers to segmented teeth using the Federation Dentaire Internationale (FDI) nomenclature and classifying them among deciduous or healthy. Furthermore, the framework is flexible enough to accommodate various applications. For example, one can think about the development and integration of a module dedicated to the identification of dental braces and pins within segmented teeth. Similarly, another module could be incorporated to identify restored teeth or teeth with prior endodontic treatment. This adaptable approach empowers the utilization of diverse network architectures, enabling the selection of the most suitable model for each task's specific domain, whether originating from our research or integrated into the system.

Bibliography

ABDALLA-ASLAN, R. et al. An artificial intelligence system using machine-learning for automatic detection and classification of dental restorations in panoramic radiography. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, v. 130, n. 5, p. 593–602, 2020. ISSN 2212-4403.

ABDINIAN, M. et al. Accuracy of digital bitewing radiography versus different views of digital panoramic radiography for detection of proximal caries. *Journal of Dentistry (Tehran, Iran)*, v. 12, p. 290–297, 04 2015.

AKKAYA, N. et al. Comparing the accuracy of panoramic and intraoral radiography in the diagnosis of proximal caries. *Dentomaxillofacial Radiology*, v. 35, n. 3, p. 170–174, 2006. PMID: 16618850.

ARI, T. et al. Automatic feature segmentation in dental periapical radiographs. *Diagnostics*, v. 12, n. 12, 2022. ISSN 2075-4418.

BAYDAR, O. et al. The u-net approaches to evaluation of dental bite-wing radiographs: An artificial intelligence study. *Diagnostics*, v. 13, n. 3, 2023. ISSN 2075-4418.

BAYRAKTAR, Y.; AYAN, E. Diagnosis of interproximal caries lesions with deep convolutional neural network in digital bitewing radiographs. *Clinical Oral Investigations*, v. 26, 01 2022.

BENKE, K.; BENKE, G. Artificial intelligence and big data in public health. *International Journal of Environmental Research and Public Health*, v. 15, n. 12, 2018. ISSN 1660-4601.

CAI, Z.; VASCONCELOS, N. Cascade r-cnn: Delving into high quality object detection. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2018. p. 6154–6162.

CANTU, A. G. et al. Detecting caries lesions of different radiographic extension on bitewings using deep learning. *Journal of Dentistry*, v. 100, p. 103425, 2020. ISSN 0300-5712.

CARNEIRO, J. A. et al. Deep learning to detect and classify teeth, dental caries and restorations: A systematic mapping. *Submitted to Dentomaxillofacial Radiology (DMFR-D-23-00388)*, 2023.

CHANG, H.-J. et al. Deep learning hybrid method to automatically diagnose periodontal bone loss and stage periodontitis. *Scientific Reports*, v. 10, 05 2020.

CHEN, H. et al. A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films. *Scientific Reports*, v. 9, 03 2019.

CHOLLET, F. *Deep learning with Python*. [S.l.]: Simon and Schuster, 2017.

CLIFTON, T.; TYNDALL, D.; LUDLOW, J. Extraoral radiographic imaging of primary caries. *Dento maxillo facial radiology*, v. 27, p. 193–8, 08 1998.

COSTA, E. D. et al. Development of a dental digital dataset for research in artificial intelligence: importance of labeling by radiologists. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology (Accepted with minor revision)*, 2023.

DAI, J. et al. *Deformable Convolutional Networks*. 2017.

ESTAI, M. et al. Deep learning for automated detection and numbering of permanent teeth on panoramic images. *Dentomaxillofacial Radiology*, v. 51, n. 2, p. 20210296, 2022. PMID: 34644152.

FERNANDES, A. F. A.; DóREA, J. R. R.; ROSA, G. J. d. M. Image analysis and computer vision applications in animal sciences: An overview. *Frontiers in Veterinary Science*, v. 7, 2020. ISSN 2297-1769.

FUKUDA, M. et al. Evaluation of an artificial intelligence system for detecting vertical root fracture on panoramic radiography. *Oral Radiology*, v. 36, 09 2019.

GEETHA, V.; APRAMEYA, K.; HINDUJA, D. Dental caries diagnosis in digital radiographs using back-propagation neural network. *Health Information Science and Systems*, v. 8, 01 2020.

GIRSHICK, R. Fast r-cnn. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2015. p. 1440–1448.

GIRSHICK, R. et al. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 11 2013.

GURSES, A.; OKTAY, A. B. Tooth restoration and dental work detection on panoramic dental images via cnn. In: *2020 Medical Technologies Congress (TIPTEKNO)*. [S.l.: s.n.], 2020. p. 1–4.

HE, K. et al. *Mask R-CNN*. [S.l.]: arXiv, 2017.

HOSNY, A. et al. Artificial intelligence in radiology. *Nature Reviews Cancer*, v. 18, 05 2018.

JADER, G. et al. Deep instance segmentation of teeth in panoramic x-ray images. In: *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. [S.l.: s.n.], 2018. p. 400–407.

KAMBUROGLU, K. et al. Proximal caries detection accuracy using intraoral bitewing radiography, extraoral bitewing radiography and panoramic radiography. *Dento maxillo facial radiology*, v. 41, p. 450–9, 09 2012.

KIM, E. Y.; LIM, K. O.; RHEE, H. S. Predictive modeling of dental pain using neural network. In: *Connecting Health and Humans*. [S.l.]: IOS Press, 2009. p. 745–746.

KITCHENHAM, B. A.; CHARTERS, S. *Guidelines for performing Systematic Literature Reviews in Software Engineering.* [S.l.], 2007.

KOCH, T. L. et al. Accurate segmentation of dental panoramic radiographs with u-nets. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019).* [S.l.: s.n.], 2019. p. 15–19.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, v. 25, 01 2012.

KUMAR, A.; BHADAURIA, H. S.; SINGH, A. Descriptive analysis of dental x-ray images using various practical methods: A review. *PeerJ Computer Science*, v. 7, 2021.

KUMAR, R.; KHAMBETE, N.; PRIYA, E. Extraoral periapical radiography: an alternative approach to intraoral periapical radiography. *Imaging Science in Dentistry*, v. 41, p. 161 – 165, 2011.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, v. 521, p. 436–44, 05 2015.

LEE, C.-T. et al. Use of the deep learning approach to measure alveolar bone level. *Journal of Clinical Periodontology*, v. 49, n. 3, p. 260–269, 2022.

LEE, J.-H. et al. Application of a fully deep convolutional neural network to the automation of tooth segmentation on panoramic radiographs. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, v. 129, n. 6, p. 635–642, 2020. ISSN 2212-4403.

LEE, J.-H. et al. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. *Journal of Dentistry*, v. 77, p. 106–111, 2018. ISSN 0300-5712.

LEITE, A. et al. Artificial intelligence-driven novel tool for tooth detection and segmentation on panoramic radiographs. *Clinical Oral Investigations*, v. 25, p. 1–11, 04 2021.

LI, H. et al. An interpretable computer-aided diagnosis method for periodontitis from panoramic radiographs. *Frontiers in Physiology*, v. 12, 2021. ISSN 1664-042X.

LIAN, L. et al. Deep learning for caries detection and classification. *Diagnostics*, v. 11, n. 9, 2021. ISSN 2075-4418.

LIN, P. L.; HUANG, P.; HUANG, P. An effective teeth segmentation method for dental periapical radiographs based on local singularity. In: *2013 International Conference on System Science and Engineering (ICSSE).* [S.l.: s.n.], 2013. p. 407–411.

LIN, T.-Y. et al. *Focal Loss for Dense Object Detection.* [S.l.]: arXiv, 2017.

LIU, S. et al. Path aggregation network for instance segmentation. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.* [S.l.: s.n.], 2018. p. 8759–8768.

LIU, S. et al. Path aggregation network for instance segmentation. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.* [S.l.: s.n.], 2018. p. 8759–8768.

MAHDI, F. P.; YAGI, N.; KOBASHI, S. Automatic teeth recognition in dental x-ray images using transfer learning based faster r-cnn. In: *2020 IEEE 50th International Symposium on Multiple-Valued Logic (ISMVL)*. [S.l.: s.n.], 2020. p. 16–21.

MALLYA, S.; ERNEST, L. *White and Pharoah's oral radiology: principles and interpretation*. [S.l.]: Elsevier Health Sciences, 2018. ISBN 9780323543835.

MITCHELL, T. M. *Machine Learning*. New York: McGraw-Hill, 1997. ISBN 978-0-07-042807-2.

MOHAMMAD-RAHIMI, H. et al. Deep learning for caries detection: A systematic review. *Journal of Dentistry*, v. 122, p. 104115, 2022. ISSN 0300-5712.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. *Foundations of machine learning*. [S.l.]: MIT press, 2018.

MURAMATSU, C. et al. Automated measurement of mandibular cortical width on dental panoramic radiographs. *International journal of computer assisted radiology and surgery*, v. 8, 11 2012.

MURAMATSU, C. et al. Tooth detection and classification on panoramic radiographs for automatic dental chart filing: improved classification by multi-sized input data. *Oral Radiology*, v. 37, p. 1–7, 01 2020.

OKTAY, A. B. Tooth detection with convolutional neural networks. In: *2017 Medical Technologies National Congress (TIPTEKNO)*. [S.l.: s.n.], 2017. p. 1–4.

OKTAY, A. B. Human identification with dental panoramic radiographic images. *IET Biometrics*, v. 7, n. 4, p. 349–355, 2018.

PARK, W.; PARK, J.-B. History and application of artificial neural networks in dentistry. *European Journal of Dentistry*, v. 12, p. 594, 10 2018.

PINHEIRO, L. et al. Numbering permanent and deciduous teeth via deep instance segmentation in panoramic x-rays. In: RITTNER, L. et al. (Ed.). *17th International Symposium on Medical Information Processing and Analysis*. [S.l.]: SPIE, 2021. v. 12088, p. 95 – 104.

PRINCE, S. J. *Computer Vision: Models, Learning, and Inference*. [S.l.]: Cambridge University Press, 2012. 598 p. ISBN 1107011795.

RAD, A. E. et al. Automatic computer-aided caries detection from dental x-ray images using intelligent level set. *Multimedia Tools and Applications*, v. 77, p. 1–20, 11 2018.

RASHID, U. et al. A hybrid mask rcnn-based tool to localize dental cavities from real-time mixed photographic images. *PeerJ Computer Science*, v. 8, p. e888, fev. 2022. ISSN 2376-5992.

REDMON, J. et al. You only look once: Unified, real-time object detection. IEEE Computer Society, Los Alamitos, CA, USA, p. 779–788, jun 2016. ISSN 1063-6919.

REN, S. et al. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 39, n. 6, p. 1137–1149, 2017.

SCHMIDT-ERFURTH, U. et al. Artificial intelligence in retina. *Progress in Retinal and Eye Research*, v. 67, p. 1–29, 2018. ISSN 1350-9462.

SCHNEIDER, L. et al. Federated vs local vs central deep learning of tooth segmentation on panoramic radiographs. *Journal of Dentistry*, v. 135, p. 104556, 2023. ISSN 0300-5712.

SCHWENDICKE, F. et al. Convolutional neural networks for dental image diagnostics: A scoping review. *Journal of Dentistry*, v. 91, p. 103226, 2019. ISSN 0300-5712.

SCHWENDICKE, F.; SAMEK, W.; KROIS, J. Artificial intelligence in dentistry: Chances and challenges. *Journal of Dental Research*, v. 99, n. 7, p. 769–774, 2020. PMID: 32315260.

SHIMIZU, H.; NAKAYAMA, K. I. Artificial intelligence in oncology. *Cancer Science*, v. 111, n. 5, p. 1452–1460, 2020.

SILVA, B. et al. 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)*. [S.l.: s.n.], 2020. p. 164–171.

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. ISSN 0957-4174.

SINGH, N. K.; RAZA, K. Progress in deep learning-based dental and maxillofacial image analysis: A systematic review. *Expert Systems with Applications*, v. 199, p. 116968, 2022. ISSN 0957-4174.

SINGH, P.; SEHGAL, P. Automated caries detection based on radon transformation and dct. In: . [S.l.: s.n.], 2017. p. 1–6.

SINGH, P.; SEHGAL, P. G.v black dental caries classification and preparation technique using optimal cnn-lstm classifier. *Multimedia Tools and Applications*, v. 80, p. 1–18, 02 2021.

TUZOFF, D. V. et al. Tooth detection and numbering in panoramic radiographs using convolutional neural networks. *Dentomaxillofacial Radiology*, v. 48, n. 4, p. 20180051, 2019. PMID: 30835551.

VINAYAHALINGAM, S. et al. Automated chart filing on panoramic radiographs using deep learning. *Journal of Dentistry*, v. 115, p. 103864, 2021. ISSN 0300-5712.

WANG, C.-W. et al. A benchmark for comparison of dental radiography analysis algorithms. *Medical Image Analysis*, v. 31, p. 63–76, 2016. ISSN 1361-8415.

WHAITES, E.; DRAGE, N. *Essentials of dental radiography and radiology*. [S.l.]: Elsevier Health Sciences, 2013.

WHO. Oral health. In: . [S.l.]: WHO Homepage, 2022. Last accessed 17 Jul 2022.

WU, Y.; HE, K. *Group Normalization*. 2018.

ZANCAN, B. Teeth numbering in panoramic radiographs through convolutional neural networks to establish a modular dental assistance system. 2023.

ZHANG, H. et al. *ResNeSt: Split-Attention Networks*. 2020.

ZHANG, K. et al. An effective teeth recognition method using label tree with cascade network structure. *Computerized Medical Imaging and Graphics*, v. 68, p. 61–70, 2018. ISSN 0895-6111.

ZHOU, L. et al. Artificial intelligence in medical imaging of the liver. *World Journal of Gastroenterology*, v. 25, p. 672–682, 02 2019.

ZOU, Z. et al. Object detection in 20 years: A survey. *Proceedings of the IEEE*, v. 111, n. 3, p. 257–276, 2023.