



A holistic approach for classifying dental conditions from textual reports and panoramic radiographs

Bernardo Silva^{a,f}, Jefferson Fontinele^b, Carolina Letícia Zilli Vieira^c,
João Manuel R.S. Tavares^d, Patricia Ramos Cury^e, Luciano Oliveira^a^{*,}

^a Intelligent Vision Research Lab, Institute of Computing, Federal University of Bahia, Bahia, Brazil

^b Federal University of Maranhão, Maranhão, Brazil

^c Harvard School of Public Health, Boston, MA, United States

^d Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, Porto, Portugal

^e Dentistry Faculty, Federal University of Bahia, Bahia, Brazil

^f Federal Institute of Bahia, Bahia, Brazil

ARTICLE INFO

MSC:

68T05

68T50

Keywords:

Large language models

Semi-supervised learning

Human-in-the-loop

Masked autoencoders

Noun phrases

Textual reports

ABSTRACT

Dental panoramic radiographs offer vast diagnostic opportunities, but the shortage of labeled data hampers the training of supervised deep-learning networks for the automatic analysis of these images. To address this issue, we introduce a holistic learning approach to classify dental conditions on panoramic radiographs, exploring tooth segmentation and textual reports, without a direct tooth-level annotated dataset. Large language models were used to identify the prevalent dental conditions in these reports, acting as an auto-labeling procedure. After an instance segmentation network segments the teeth, a linkage approach is in charge of matching each tooth with the corresponding condition found in the textual report. The proposed framework was validated using two of the most extensive datasets in the literature, specially gathered for this study, consisting of 8,795 panoramic radiographs and 8,029 paired reports and images. Encouragingly, the results consistently exceeded the baseline for the Matthews correlation coefficient. A comparative analysis against specialist and dental student ratings, supported by statistical evaluation, highlighted its effectiveness. Using specialist consensus as the ground truth, the system achieved precision comparable to final-year undergraduate students and was within 8.1 percentage points of specialist performance.

1. Introduction

Imaging significantly impacts dentistry, enabling specialists to identify problems that might not be visible during a clinical examination. Imaging modalities like X-rays, computerized tomography scans, and magnetic resonance imaging provide detailed views of teeth, bones, and soft tissues (White and Pharoah, 2014). These tools enhance the precision of diagnoses and treatments, ensuring better patient outcomes. Among the current imaging exams, radiographs are the most common in dentistry (White and Pharoah, 2014; Langlais and Miller, 2016), being requested to identify various pathologies like cavities, periodontal disease, impacted teeth, and bone infections (Chang et al., 2020; Yüksel et al., 2021) and track the progress of dental treatments.

One of the most commonly used radiographs in dentistry is the panoramic radiograph (White and Pharoah, 2014; Langlais and Miller, 2016; Silva et al., 2018), which is an extraoral imaging technique in

which the X-ray film or sensor remains outside the patient's mouth during acquisition. In a single image, the panoramic radiograph provides a comprehensive view of both the upper and lower jaws, but with fewer details of the structure of the mouth (Haring and Jansen, 2000; Silva et al., 2018; Jader et al., 2018; Pinheiro et al., 2021). Fig. 1 shows an example of a panoramic radiograph, revealing the structures and their overlaps, which can lead to cluttered readings.

While panoramic radiography offers considerable flexibility in diagnosing periodontal bone loss, bone irregularities in the maxilla and mandible, temporomandibular joint disorders, and other conditions, the teeth remain central to the diagnostic process (Whaites and Drage, 2013; Silva et al., 2020; Pinheiro et al., 2021). Because of the significance of the teeth and to expedite communication, dentists usually do not identify the teeth by their complete anatomical designations but rather by numbers defined by a system called *Fédération Dentaire*

* Corresponding author.

E-mail address: irebouca@ufba.br (L. Oliveira).

<https://doi.org/10.1016/j.media.2025.103709>

Received 26 June 2024; Received in revised form 2 April 2025; Accepted 22 June 2025

Available online 2 July 2025

1361-8415/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

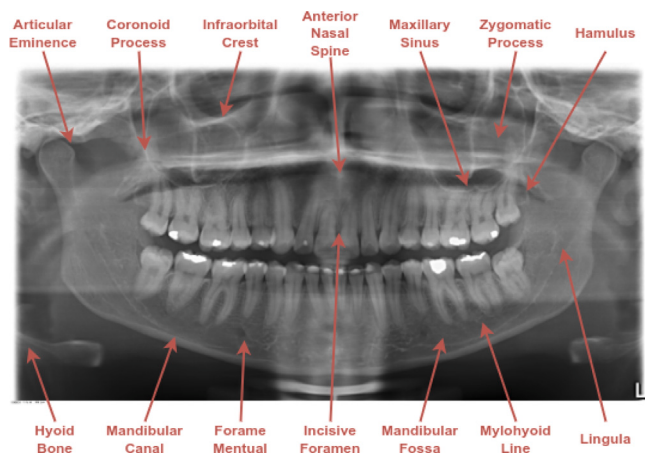


Fig. 1. Sample of a human panoramic radiograph with the oral structures identified.

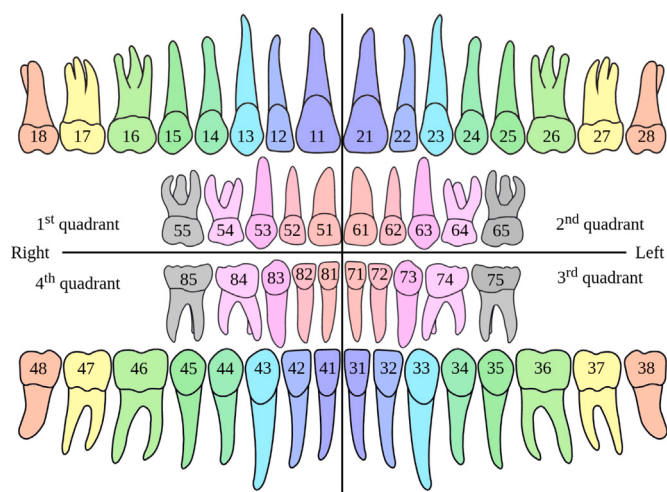


Fig. 2. Illustration of FDI notation: A two-digit system that determines each tooth (a custom color code was added to show qualitative results).
Source: Adapted from Silva et al. (2023).

Internationale (FDI) notation (Pinheiro et al., 2021). Fig. 2 illustrates the two-digit FDI notation. The first digit of the FDI system denotes one of the four mouth quadrant and the dentition type (permanent, 1–4, or deciduous, 5–8); the second digit represents one of the eight tooth types, in each quadrant. The axes are mirrored relative to the Cartesian system because the radiograph itself is acquired in a mirrored orientation. In Fig. 2, we introduced a custom color code for each tooth, which will be used later to present the qualitative results.

Dental reports consist of unstructured data pairs: images and text. These textual records represent the most comprehensive and valuable database of dental conditions associated with panoramic radiographs, providing detailed descriptions of the entire oral cavity (Jing et al., 2017). Images are challenging to interpret due to the vast amount of visual information they contain. Nowadays, computer vision (CV) researchers have been addressing image complexities with relative success through supervised machine-learning techniques. Supervised strategies rely on labeled data, which are manually annotated, restraining their scalability. Text manipulation also presents its challenges, which have previously hindered researchers from utilizing text reports in their work with panoramic radiographs. These challenges arise from the intricacies of natural languages, where nuances, idioms, and contextual meanings can significantly affect interpretation (Névéol et al., 2018). Recently, with the advent of large language models (LLMs), the application of artificial intelligence (AI) to text has proven successful.

These LLMs are trained in a self-supervised manner and considerably ease the natural language processing (NLP) tasks (Brown et al., 2020). Consequently, the extraction of dental conditions proved to be feasible yet unexplored by the literature of computational dentistry.

The advances in NLP due to LLMs have turned CV into the most significant bottleneck, primarily because it relies heavily on labeled data. Consequently, no prior work in this domain has conducted its experiments on a dataset of a considerable magnitude. Motivated by this gap, a framework for the classification of thirteen dental conditions in panoramic radiographs is introduced in the current article. The proposed holistic approach incorporates a semi-supervised learning paradigm to address the scarcity of labeled data, a common issue in the medical field. Additionally, a novel auto-labeler strategy that leverages LLMs to extract dental conditions from textual reports is proposed.

1.1. Related work

Recent studies in the area of dental radiographic analysis have leveraged machine learning to address various diagnostic challenges, predominantly relying on supervised learning. For instance, Ekert et al. (2019) developed a custom seven-layer neural network to detect apical lesions in teeth using manually cropped panoramic radiographs, achieving an Area under the ROC Curve (AUC) of 0.85. Similarly, Fukuda et al. (2020) employed the DetectNet CNN architecture to identify vertical root fractures in a dataset of 330 panoramic radiographs, reporting a promising F1-score of 0.83.

In the odontogenic cystic lesions (OCLs) domain, Lee et al. (2020) used GoogLeNet Inception-v3 to distinguish between odontogenic keratocyst, dentigerous cysts, and periapical cysts. The study gathered 2126 high-quality images and found the method particularly effective when trained with cone-beam computed tomography (CBCT) images. Building on this, Kwon et al. (2020) employed YOLOv3 to identify four types of dental cysts using 1282 labeled panoramic radiographs, achieving high accuracy despite the limited dataset size.

To enhance the diagnosis of various dental conditions, Chen et al. (2021) developed a CNN-based auxiliary system for detecting lesions in 2900 periapical radiographs, focusing on different disease categories and severity levels. The results were notably promising for severe lesions. Additionally, Yüksel et al. (2021) introduced DENTECT, a three-stage pipeline for classifying five dental conditions using panoramic radiographs, demonstrating performance comparable to dental clinicians.

Segmentation of dental conditions has also been explored. Khan et al. (2021) evaluated U-Net-based architectures for identifying caries, alveolar bone loss, and interradicular radiolucencies in 206 periapical radiographs, suggesting that custom-built solutions could improve performance. Similarly, Vinayahalingam et al. (2021) used a Mask R-CNN to segment and classify teeth and five dental conditions in 2000 radiographs, achieving high F1-scores and underscoring the potential of deep learning in clinical diagnostics.

Several studies have compared different neural network architectures for dental condition classification. Liu et al. (2023) evaluated ResNet-50 (He et al., 2016), VGG-16, InceptionV3, and DenseNet-121 on 188 periapical radiographs, with DenseNet-121 achieving the highest accuracy of 99.5%. Bonfanti-Gris et al. (2022) assessed Denti.AI™ for detecting five dental conditions, finding it effective for implants, crowns, and metal fillings but limited in classifying resin-based restorations.

Innovative approaches include DiagnoCat (Amasya et al., 2024), which used Mask R-CNN and Cascade R-CNN to detect periodontal bone loss in 6000 panoramic radiographs, achieving accurate results in comparison with clinician assessments. Additionally, Ranjbar and Zamanifar (2023) focused on predicting eight future dental treatments using YOLOv7, showing high accuracy and potential for clinical application.

Lastly, Gao et al. (2024) introduced YOLO-DENTAL for detecting and classifying dental conditions in 413 periapical radiographs,

Table 1

Comparison of the current study with others regarding dataset, task, learning paradigm, proposed solution, and investigated classes.

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

achieving a notable mean average precision (mAP) of 86.81%. Similarly, [Tassoker et al. \(2024\)](#) used YOLOv5 to detect idiopathic osteosclerosis in 175 panoramic radiographs, reporting high detection accuracy despite dataset challenges. These studies collectively highlight the evolving landscape of machine learning applications in dental radiographic analysis.

1.2. Contributions

Table 1 compares the current study with other research in dental radiographic automation, highlighting common limitations. The table provides a comprehensive summary of each study, including the type of radiographs analyzed, the number of radiographs utilized, the

tasks addressed (e.g., detection or auto-labeling), the learning approach employed (supervised, semi-supervised, or self-supervised), the neural network architecture adopted, and the number of classes, along with their respective names. Most studies, except for one using a dataset of 6000 images (Amasya et al., 2024), relied on datasets with 2900 or fewer samples (Chen et al., 2021; Lee et al., 2020; Vinayahalingam et al., 2021). Such small datasets can compromise model generalizability. Additionally, many studies excluded challenging cases, limiting the practical applicability of their findings (Fukuda et al., 2020; Lee et al., 2020; Chen et al., 2021; Liu et al., 2023; Amasya et al., 2024; Gao et al., 2024). The focus on a narrow range of target classes (Ekert et al., 2019; Fukuda et al., 2020; Lee et al., 2020; Khan et al., 2021; Liu et al., 2023; Amasya et al., 2024) further restricts the comprehensiveness of these models.

Technically, most studies relied on supervised learning (Ranjbar and Zamanifar, 2023; Tassoker et al., 2024; Amasya et al., 2024; Bonfanti-Gris et al., 2022), which requires extensive labeled data, which is a significant limitation. There was also a trend towards using off-the-shelf solutions (Lee et al., 2020; Kwon et al., 2020; Vinayahalingam et al., 2021; Liu et al., 2023; Ranjbar and Zamanifar, 2023).

Our study introduces a novel framework for diagnosing dental conditions using panoramic radiographs and textual reports, addressing key limitations in existing methodologies. Our framework is built on a holistic approach that combines a human-in-the-loop (HITL) strategy (Silva et al., 2023) with masked autoencoders (MAE) (He et al., 2022), to extract the cropped teeth from unlabeled panoramic radiographs. A central idea of our proposed framework is the automated generation of tooth-level datasets by leveraging pseudolabeled radiographs to create tooth crops. To assign meaningful labels to these crops, we developed an auto-labeler provided by a large language models (LLMs) and a text-image linkage procedure. The auto-labeler extracts dental conditions from textual reports and maps them to the corresponding teeth using the FDI numbering system, covering distinct dental conditions. The framework also includes a statistical agreement analysis to ensure the accuracy and validity of the generated labels. As part of this study, we introduce the largest dataset in the field, comprising 16,824 panoramic radiographs. This dataset includes 8795 independent radiographs and 8029 paired textual reports, representing an unprecedented resource for dental diagnostics research. By integrating panoramic images and textual data, this work provides a scalable and efficient solution for diagnosing dental conditions, setting a new benchmark for the field.

2. Classifying multiple dental conditions through a holistic framework

The proposed framework consists of four major steps: (i) Dataset construction, (ii) tooth pseudolabeling and crop generation, (iii) classification network pretraining and label extraction, and (iv) dental conditions classification. These steps are depicted in Fig. 3 and described in the following.

2.1. Dataset construction

Two distinct groups of datasets were built, as detailed in Tables 2 and 3,¹ all acquired using the same imaging device (Silva et al., 2023). The first group comprises a set of panoramic radiographs in total dimensions. The second group, derived from the first, comprises tooth crops for training binary classification models. It is noteworthy that images produced by different equipments are similar when the radiographic technique is performed correctly. This assumption is supported

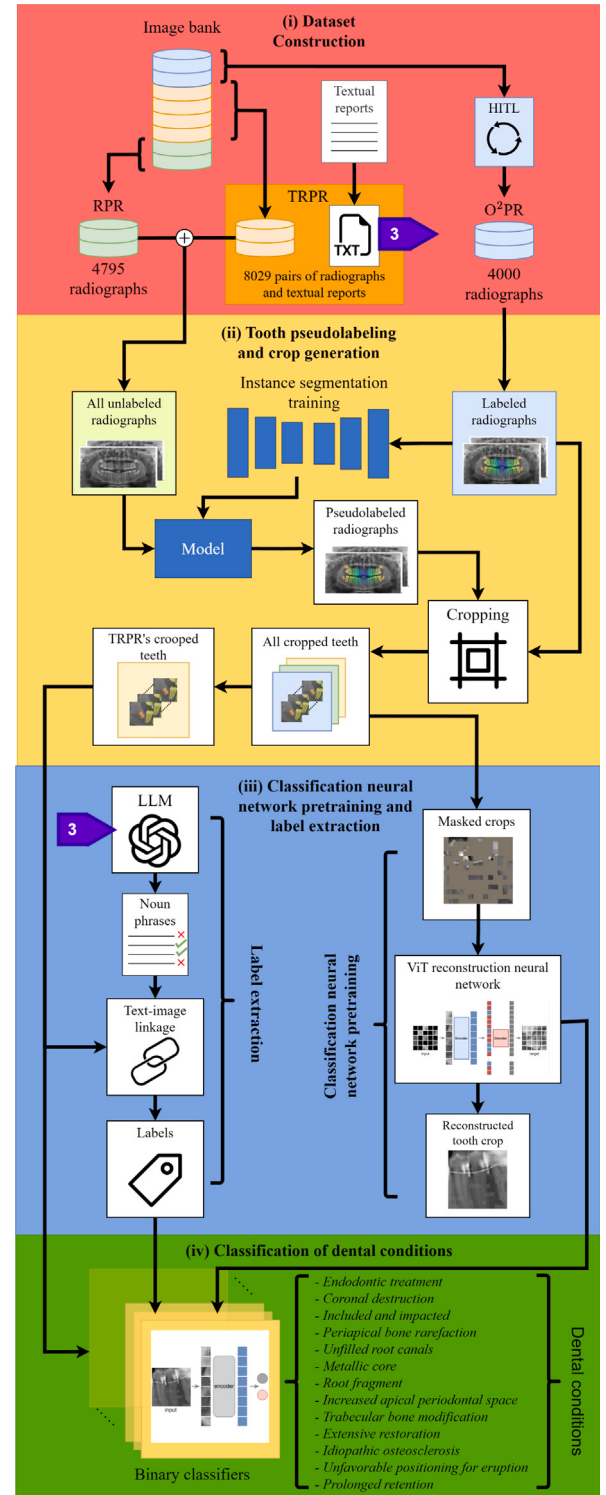


Fig. 3. Holistic framework for classifying dental conditions. (i) **Dataset construction:** Combines Textual Report Panoramic Radiographs (TRPR), Raw Panoramic Radiographs (RPR), and O²PR. (ii) **Tooth pseudolabeling and crop generation:** Uses an instance segmentation neural network to generate tooth pseudolabels on unlabeled radiographs, creating tooth crops. (iii) **Classification network pretraining and label extraction:** Cropped teeth and text information pretrain a model via MAE, extracting noun phrases with a large language model (LLM). Text-image linkage generates labels for binary classifiers. (iv) **Classification of dental conditions:** Trains a binary classifier for each dental condition.

¹ The National Commission for Research Ethics (CONEP) and the Research Ethics Committee (CEP) authorized the use of the radiographs in research under the report number 646.050/2014.

Table 2

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

Dataset	# Images	Inst. Segm. labels	Textual reports	Pseudolabeling	Mode dimensions	Inst. Segm. training data
RPR	4,795			✓		None
O ² PR	4,000	✓			2440 × 1292	All
TRPR	8,029		✓	✓		None

Table 3

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

Dataset	Crop dimensions	Source dataset	# Crops	Textual reports	Pretraining	Training	Validation	Testing
Less context	224 × 224	RPR O ² PR	132,497 112,842		All	None	None	None
More context	380 × 380 to 224 × 224	TRPR	213,395	✓	Train only (70%)	70%	15%	15%

by [Alahmad and Alnafea \(2023\)](#), who discuss the importance of proper technique in achieving optimal image quality across various imaging systems. Under these conditions, images from the same device often show minimal intra-class variability. We used a large dataset collected over ten years from a radiology clinic, encompassing a wide range of dental conditions referred by various specialists. However, [Krois et al. \(2021\)](#) showed that models trained on single-center data performed poorly on images from other centers when used a fully supervised segmenter; incrementally adding data from a second center improved performance in the new domain, though with a slight decline in the original domain, highlighting the importance of data diversity. Unlike fully supervised CNN-based approach as in [Krois et al. \(2021\)](#), we used a ViT with a masked autoencoder. This way, semi- and self-supervised learning, as explored in our work, shows potential for improving generalizability in clinical settings, but its relationship with multi-center data integration requires further investigation, lying beyond the scope of our work.

The first group's preparation step is illustrated in [Fig. 3\(i\)](#). This group of radiographs was sourced from an image bank with 16,824 samples. These images were gathered into three distinct, non-overlapping datasets. The first dataset encompasses 4795 unlabeled images retained in their original, raw format. Therefore, it is designated here as the Raw Panoramic Radiographs (RPR) dataset. The second dataset, the O²PR, was released by [Silva et al. \(2023\)](#) and built under the human-in-the-loop (HITL) concept. It includes 4,000 radiograph images along with tooth instance segmentation labels and numbering. Finally, the third dataset, devoid of tooth segmentation labels, comprises 8,029 panoramic radiographs and the corresponding textual reports of two radiologists working independently. It is designated here as the Textual Report Panoramic Radiographs (TRPR) dataset. The main characteristics of these datasets² are presented in [Table 2](#).

The images had widths ranging from 2,272 to 2,692 and heights ranging from 1,292 to 1,304. The most common size was 2440 × 1292, accounting for 79% of the images. The textual reports were preprocessed from the original .odt format and converted to the .txt format for consistency and easier manipulation. The final .txt files had the format exemplified in [Table 4](#): A two-digit number followed by a colon and the report line. In [Table 4](#), the teeth are highlighted in bold according to the FDI system described in [Fig. 2](#), to emphasize their significance for the radiograph screening. In the used textual reports, as commonly done, the digits of the textual reports denote a tooth through the FDI notation, while other numbers are written out in full.

The second group of datasets is synthesized from the previous datasets (RPR, O²PR, and TRPR). It comprises tooth crop images, which will be detailed in the following section.

Table 4

Sample of a panoramic radiograph preprocessed report of the TRPR dataset. It is highlighted in bold the mentioning of the teeth according to the FDI system described in [Fig. 2](#) to reveal their importance.

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

2.2. Tooth pseudo labeling and crop generation

This step serves two purposes, as illustrated in [Fig. 3\(ii\)](#): (i) to automatically generate pseudo labels for the teeth in the radiographs of the RPR and TRPR datasets using an instance segmentation network, as they do not contain instance segmentation labels; and (ii) to create tooth crops from the tooth labels and pseudo labels, which are later used to train and test the classification neural networks. These procedures allow consistent, uniformly sized image crops over all datasets around each tooth. The standardization is crucial for subsequent steps, as the classification neural networks are trained on fixed-size inputs. Notably, the instance segmentation network's output is achieved by training it with "pseudolabels".

An instance segmentation neural network was trained on all 4,000 labeled images of the O²PR dataset. The hybrid task cascade (HTC) architecture ([Chen et al., 2019](#)) was selected because it was the best model in the benchmark conducted by [Silva et al. \(2023\)](#). HTC ensures more accurate object boundaries and improved detection results by leveraging information from tasks like semantic segmentation. Using this network, the teeth of all the remaining unlabeled radiographs from the RPR (4,795) and the train set of TRPR (8,029) datasets were segmented. Finally, 4,000 labeled radiographs and 12,824 pseudo labeled radiographs were obtained.

After training, our HTC instance segmentation neural network was employed to generate two distinct datasets of tooth crops from all labeled and pseudolabeled radiographs. [Table 3](#) summarizes the characteristics of these datasets, detailing the crop dimensions, the original datasets used, the total number of crops, the availability of corresponding textual reports, whether the crop dataset was employed in the

² Instructions to request the datasets can be found in <https://github.com/VisionLab/MEDIA-datasets>.

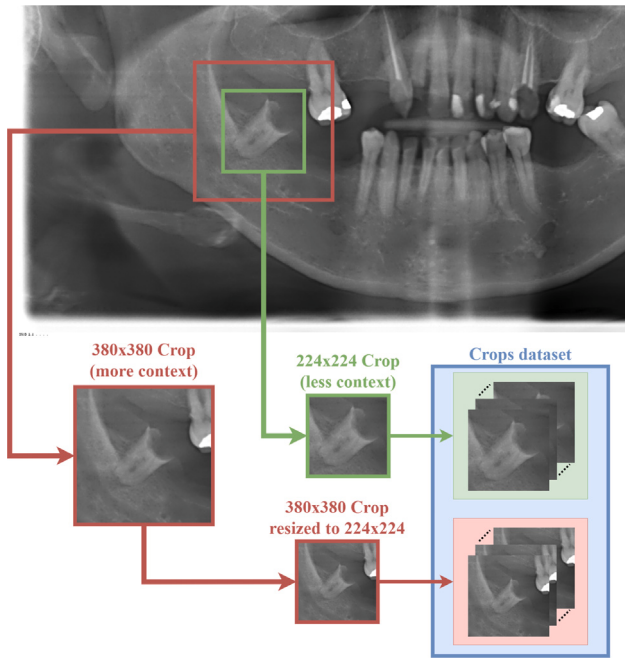


Fig. 4. Two tooth crop variants used in this study: The first, termed the “less context” crop, was taken from a panoramic radiograph of a tooth and measures 224×224 pixels. The second, termed the “more context” crop, was resized to 224×224 pixels from an original crop of size 380×380 pixels. These two sets comprise the Crops dataset.

pretraining process, and how much it was used in training, validation, or testing routines. The primary crop variant, referred to as “less context”, consists of 224×224 crops centered around each tooth. This approach offers the advantage of focusing closely on the teeth but has the limitation of excluding surrounding context, potentially omitting parts of the teeth. To mitigate this limitation, a second crop variant, termed “more context”, was introduced. These crops initially captured a broader 380×380 area around the tooth, which was subsequently resized to 224×224 to meet the input requirements of the classification neural network architecture. Both configurations yielded approximately 460,000 tooth crops. Fig. 4 illustrates this cropping process.

2.3. Classification network pretraining and label extraction

The next step is illustrated in Fig. 3(iii). This stage consists of two distinct processes that can be executed independently: the neural network pretraining and label extraction (via auto-labeler). The neural network pretraining was designed to enhance the performance of classification models for the final tasks, leveraging the unlabeled data to learn significant features. Besides the baselines discussed in Section 3.2, this process was run independently twice, once for each set of the tooth Crops dataset. Tooth crops of 224×224 pixel dimensions were used to meet the requirements of the chosen classification model: the Vision Transformer (ViT) (Dosovitskiy et al., 2020). The ViT was selected for the used pretraining strategy because of its superior performance in benchmarks (Khan et al., 2022), where it achieved state-of-the-art results by exploiting transformers, and due to its convenience for employing the Mask AutoEncoder (MAE) strategy. The current study used MAE, a variant of deep-learning autoencoders trained by deliberately masking out portions of the input data, as a pretraining strategy. This “masking” approach challenges the network to reconstruct the original data, including the intentionally obscured (masked) patches, resulting in a more robust latent representation. Benefiting from transfer learning principles, MAEs can leverage pretrained models to further enhance

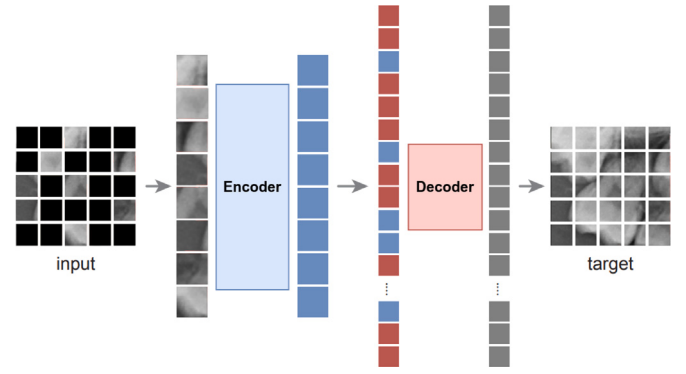


Fig. 5. Illustration of a MAE: Selected patches from an input image are obscured, and the remaining visible patches are processed through an encoder. The obscured patches are subsequently reconstructed using a decoder from the latent space representations.

their performance and generalization capabilities. Fig. 5 illustrates an MAE within the context of the ViT architecture. Given an input image segmented into a grid with some sections obscured, the encoder compresses this partial image into a compact representation. These encoded data are then processed by the decoder, aiming to regenerate the full image. In the selected configuration, the decoder has fewer parameters than the encoder to focus on efficient reconstruction (He et al., 2022). After completing the pretraining phase, the decoder is discarded, emphasizing its role in learning robust image representations without contributing to the final task performance.

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

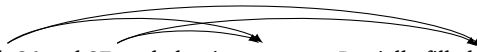
The state-of-the-art GPT-4 (OpenAI, 2023) was used to automate and expedite noun phrase extraction. The adopted methodology is centered on prompt engineering – a technique where specific inputs are crafted to elicit desired outputs from language models. Guided by this approach, the following prompt was formulated to be executed on the textual reports:

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

With the input above, for instance, the noun phrases extracted from sentence “04” in the report shown in Table 4 were:

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

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


 Tooth 36 and 37: endodontic treatment. Partially filled root canals.

The linkage process proves effective in the context of dental reports, which are organized by specific conditions. Radiologists are more favorable to note the presence of conditions rather than the absence. This tendency is illustrated in the case of the patient mentioned in Table 4, sentence “03”. For example, it is less common for radiologists to document each tooth lacking conditions, such as ‘Teeth 17, 16, 15, 14, 13, ... without endodontic treatment or unfilled root canals’, focusing instead on those with notable conditions.

Sentences detailing the presence or absence of teeth were excluded from the analysis, as illustrated in sentence “02” of Table 4. This decision was based on the understanding that this task is better conducted using object detection or instance segmentation techniques specifically designed for identifying and delineating objects within images (Silva et al., 2020; Pinheiro et al., 2021). To filter out these sentences, straightforward regular expressions were implemented.

2.4. Dental conditions classification

In Fig. 3(iv), the focus is on training the ViT networks (binary classifiers). Initially, a baseline model without pre-training was used. Pretrained weights from MAEs were also explored to broaden the proposed approach. Specifically, weights from MAEs pretrained on two datasets: The widely recognized ImageNet dataset and the previously constructed Crops dataset were used.

The Matthews correlation coefficient (MCC) metric (Matthews, 1975) was used for the proposed network evaluation. The MCC takes into account true positives (TP), false positives (FP), and false negatives (FN), including the frequently overlooked true negatives (TN):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \quad (1)$$

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

3. Experimental analysis

3.1. Training for tooth pseudolabeling

The main goal of this phase was training an instance segmentation neural network, specifically the HTC. Following the methodology outlined by Silva et al. (2023), the trained HTC used a ResNeXt neural network as its backbone with 101 layers and a cardinality of 64 (Xie et al., 2017). The initial weights of this network were derived from the training on the ImageNet dataset to leverage the transfer of the learning technique later. The training data comprised 4000 images from the O²PR dataset. No data was allocated for testing, emphasizing that the goal was not efficiency measurement but tooth crop generation for subsequent phases.

Data augmentation was purely horizontal flips, carefully changing the labels of the teeth from the right side to the left and vice versa. To optimize network performance, the radiographs were cropped from their prevailing 2440 × 1292 pixels to 1876 × 1036, removing 159 pixels from the top, which resulted in more focused teeth. The batch size was 1 (one), and the optimizer was stochastic gradient descent, with a learning rate of 0.0015, momentum of 0.9, and no weight decay. The

threshold value for tooth detection was 0.5. The network was trained for 20 epochs in an NVIDIA GeForce GTX TITAN X. After training the neural network; it was applied to the 12,824 unlabeled images from the RPR and TRPR datasets. This allowed one to create tooth crops for all the radiographs in the used image dataset. Fig. 6 shows two samples of the instance segmentation results, in an adult’s and a child’s mouth, using the color code introduced in Fig. 2, demonstrating its promising performance (refer to Silva et al. (2023) for further details on instance segmentation results).

3.2. Neural network pretraining

MAE was utilized to pretrain Vision Transformer (ViT) neural networks, which were subsequently fine-tuned through transfer learning for final classification tasks, addressing each dental condition individually. All tooth crops were used for pretraining the ViTs, reserving some images for validation and testing purposes (see Table 3). The experiment encompassed two scenarios: the first involved ViT networks pretrained on the ImageNet dataset and the second used custom-generated tooth crops. Each scenario was executed twice, accommodating both crop configurations, culminating in four distinct experimental setups.

- Less context crops pretrained on ImageNet dataset
- Less context crops pretrained on Crops dataset
- More context crops pretrained on ImageNet dataset
- More context crops pretrained on Crops dataset

In pretraining scenarios, data augmentation techniques used horizontal flip and random resized crop, with scales ranging from 0.2 to 1. The batch size was 512, and the optimizer was AdamW with a learning rate of 9.5×10^{-4} , betas of 0.9 and 0.95, and no weight decay. The network was trained for 800 epochs with a linear warm-up in the first 40 epochs. The hardware used for training was eight NVIDIA A100 of 80 GB. The depicted sample in Fig. 7 showcases the considerable qualitative success of the reconstruction outcomes from the pretraining configuration using tooth crops.

3.3. Label extraction

In this phase, OpenAI’s LLM GPT-4 was used to streamline and expedite the extraction of noun phrases from textual reports of the TRPR dataset. For the current study, the frequency of all noun phrases was gauged, and only those with occurrences higher than 150 were considered. This threshold was chosen arbitrarily, believing it to represent the minimum necessary for a network to learn effectively. Afterward, similar phrases, such as “unfilled root canal” with its plural form “unfilled root canals”, were manually grouped. As not all noun phrases are dental conditions, the selection was refined through manual filtering. For example, “endodontic treatment” is a dental condition, whereas “clinical assessment” is not and, therefore, was excluded from the analysis. Fig. 8 displays a bar chart depicting the 27 most common noun phrases, their frequency, and trends, evincing their long tail distribution. Noun phrases identified as dental conditions are marked with red bars, whereas those not selected are shown with blue bars.

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

1. Endodontic treatment (4,994) - a procedure that treats infections inside the tooth, typically involving the removal of the pulp and nerves, followed by the filling and sealing of the pulp chamber and root canals.
2. Coronal destruction (1,866) - Damage or decay to the crown portion of the tooth.
3. Included and impacted (1,532) - Teeth trapped within the jawbone or gums and cannot erupt naturally.

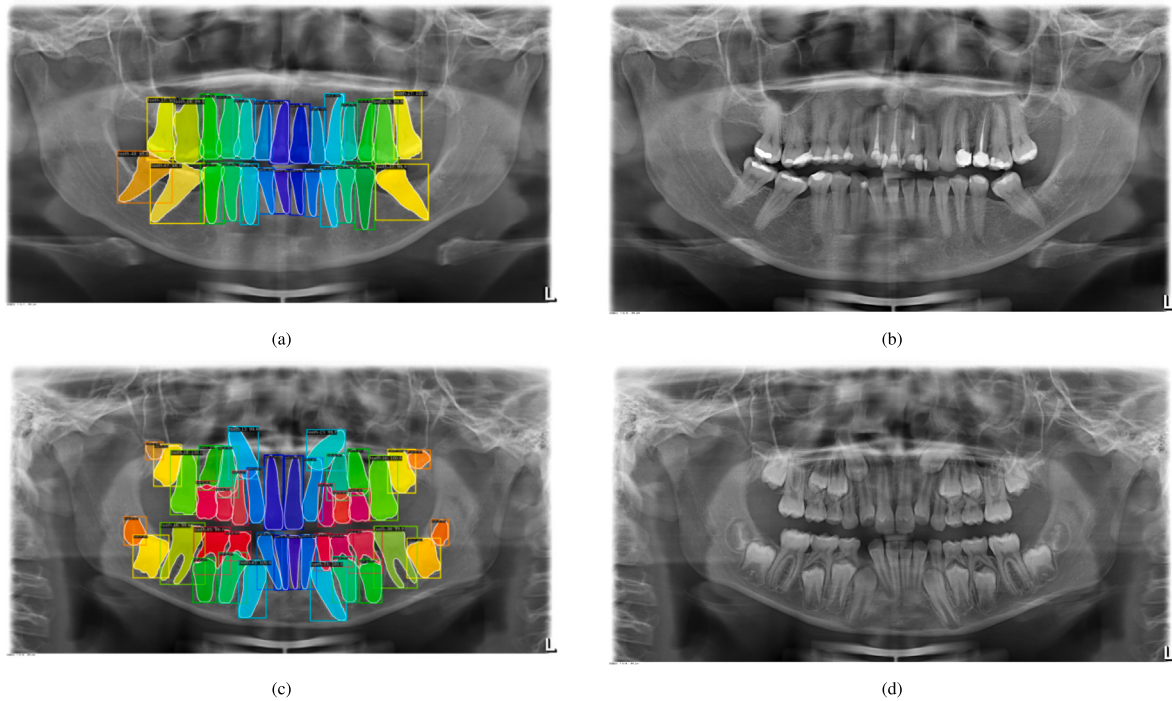


Fig. 6. Qualitative results of the trained instance segmentation neural network, using the color code introduced in Fig. 2. (a) and (b) Instance segmentation results on an adult's mouth, paired with the original image. (c) and (d) Instance segmentation results on a child's mouth, paired with the original image.

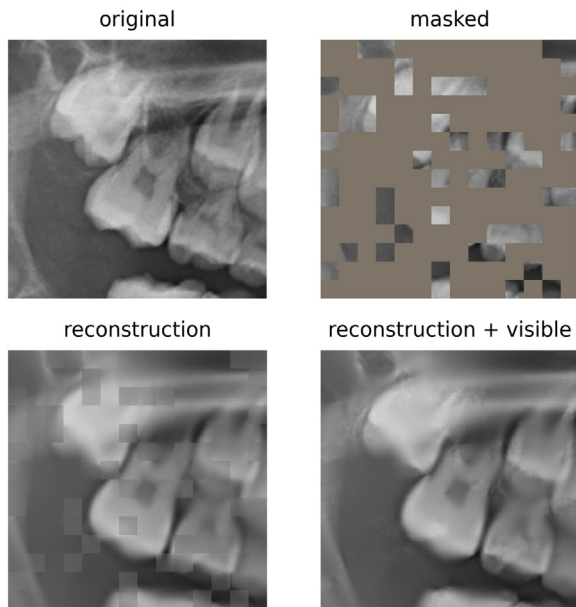


Fig. 7. Reconstruction sample from a pretrained neural network using MAE as a pretraining strategy, showcasing the efficacy of MAE in enhancing the network's reconstructive capabilities.

4. Periapical bone rarefaction (1,486) - A reduction or loss of bone density around the apex of a tooth root, often due to inflammation or infection.
5. Unfilled root canals (1,194) - Root canals that have not been filled or sealed after an endodontic procedure.
6. Metallic core (1,091) - A metal post used to support a restoration or crown, especially in a tooth undergoing endodontic treatment.

7. Root fragment (964) - A piece or portion of a tooth root left behind, typically after tooth extraction or breakage.
8. Increased apical periodontal space (922) - Enlargement of the space around the tooth root's apex, which may indicate an inflammatory response.
9. Trabecular bone modification (773) - Changes in the spongy part of the bone, which can be indicative of disease or other conditions.
10. Extensive restoration (573) - Large dental fillings or excessive material used in a dental restoration.
11. Idiopathic osteosclerosis (470) - A localized increase in bone density without a known cause.
12. Unfavorable positioning for eruption (200) - The positioning of a tooth that hinders its natural eruption process.
13. Prolonged retention (181) - The extended presence of a tooth or dental element beyond its normal duration, often referring to baby teeth that do not fall out on time.

Fig. 9 displays examples of each condition. Upon determining the conditions to be evaluated, the adopted linkage process, which associates every tooth mentioned in a sentence with all the dental conditions stated in that sentence, was applied as described in Section 2.3.

3.4. Training of the classification network

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

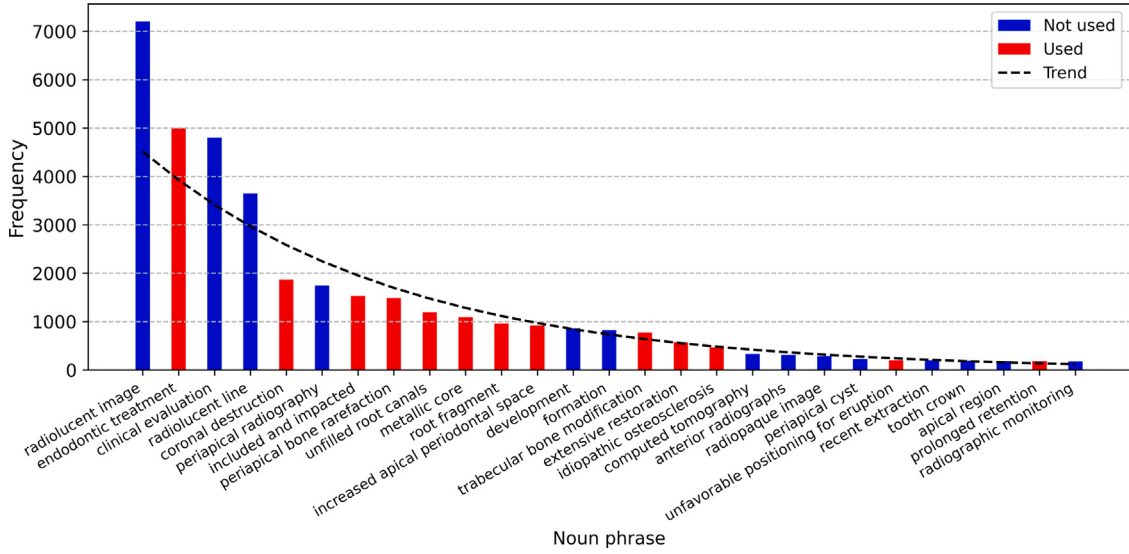


Fig. 8. Bar chart of the 27 most common noun phrases, showing their frequency and trends, and illustrating their long tail distribution. (Noun phrases identified as dental conditions are highlighted with red bars, while those not selected are represented with blue bars.).

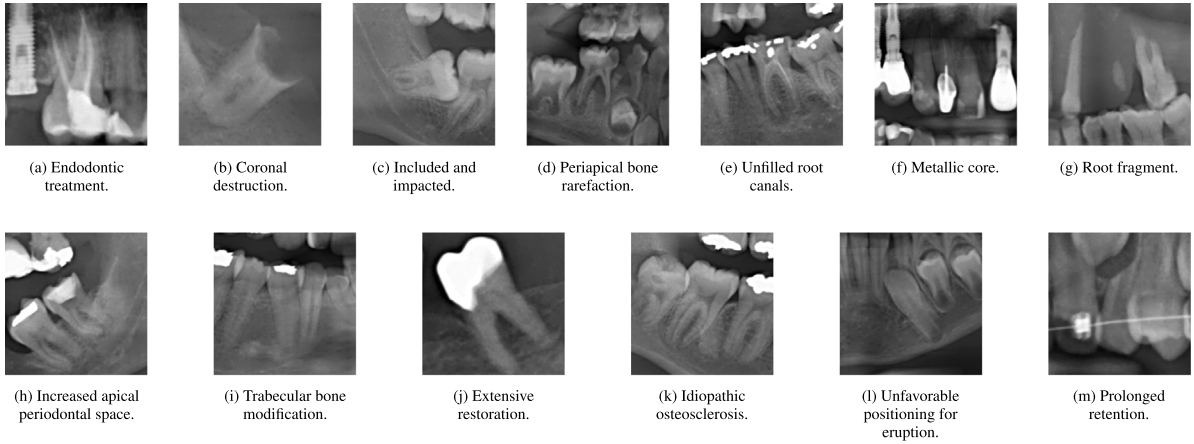


Fig. 9. Dental conditions considered in this study. (They were selected according to their frequency in the textual reports.)

eight NVIDIA A100 of 80 GB. Code implementation was carried out using PyTorch, version 1.11. Finally, the loss was binary cross entropy (BCE) plus MCC loss were calculated:

$$\text{BCE}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^N (y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)), \quad (2)$$

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

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

The final loss is

$$\text{Loss} = \alpha \cdot \text{BCE}(y, \hat{y}) + (1 - \alpha) \cdot (1 - \text{MCC}(y, \hat{y})). \quad (3)$$

α was determined through empirical experimentation. Its value was systematically varied in increments of 0.1 within the range [0.1, 0.9], using the training and validation sets to evaluate performance. The optimal value was obtained as $\alpha = 0.5$.

3.5. Results and discussions

We compared our proposed framework with a ResNet with 50 layers (He et al., 2016). The inclusion of ResNet in the experiments serves as a strong baseline for comparison, as it is widely recognized

as a robust deep-learning classifier, ensuring a fair evaluation of the proposed method without the need to compare against other similarly performant classifiers. The ResNet-50 was trained using the same data split strategy as our proposed method, maintaining consistency across training, validation, and testing sets. The configurations for training epochs, crop size, data augmentation, optimizer, and warm-up strategy remained the same, as detailed in Section 3.4.

Table 5 summarizes results on validation and test datasets. For each tooth condition (label), it is shown the number of (samples) presented in the experiment, results on validation dataset and test dataset. ResNet presented the worst results both on the less-context and more-context tooth cropped scenarios, considering all labels. In contrast, our proposed framework pretrained on the tooth crops (pseudolabels), indicated in Table 5 by crops, had the best average results in both cases (less and more context). ViT using MAE pretrained with tooth crops outperformed ViT pretrained with ImageNet, on average, by 1.6% and 9.5% percentage points in the less-context and more-context tooth crops, respectively. Despite containing approximately 460,000 images – far fewer than the ImageNet dataset's more than 17 million – pretraining with tooth crops proved more efficient due to the field-oriented data context. Table 6 shows that tooth crop pretraining configurations perform more optimally in fewer epochs than those pretraining with ImageNet.

Table 5

Results based on MCC across the validation and test datasets. Values in bold indicate the best results in each line, considering validation and test columns, separately.

Label	#Samples	Validation						Test	
		Less context (224×224)			More context (380×380)			ResNet	Ours
		ResNet (ImageNet)	VIT (ImageNet)	MAE (Crops)	ResNet (ImageNet)	VIT (ImageNet)	MAE (Crops)		
1	4,994	0.714	0.903	0.904	0.640	0.847	0.846	0.716	0.865
2	1,866	0.132	0.668	0.714	0.117	0.663	0.675	0.139	0.658
3	1,532	0.153	0.767	0.740	0.146	0.776	0.790	0.165	0.683
4	1,486	0.500	0.589	0.598	0.442	0.498	0.523	0.273	0.397
5	1,194	0.161	0.454	0.445	0.112	0.611	0.595	0.158	0.436
6	1,091	0.124	0.677	0.695	0.095	0.711	0.750	0.090	0.632
7	964	0.090	0.532	0.510	0.086	0.728	0.668	0.087	0.583
8	922	0.255	0.275	0.270	0.229	0.399	0.405	0.245	0.327
9	773	0.315	0.301	0.309	0.339	0.506	0.458	0.116	0.218
10	573	0.032	0.385	0.286	0.048	0.284	0.314	0.016	0.252
11	470	0.057	0.182	0.182	0.060	0.424	0.414	0.066	0.347
12	200	0.142	0.336	0.420	0.143	0.430	0.456	0.167	0.353
13	181	0.107	0.577	0.666	0.091	0.386	0.545	0.030	0.426
Average		0.214	0.511	0.519	0.196	0.559	0.572	0.174	0.475

Table 6

Epoch convergences based on the highest MCC value on the validation dataset.

Label	#Samples	Validation			
		224 × 224 crops (less context)		380 × 380 crops (more context)	
		ImageNet	Crops	ImageNet	Crops
1	4,994	38	11	7	7
2	1,866	14	23	8	13
3	1,532	23	26	15	30
4	1,486	14	13	9	10
5	1,194	11	11	24	11
6	1,091	25	21	8	11
7	964	25	11	24	7
8	922	3	9	6	6
9	773	23	6	26	17
10	573	5	12	19	1
11	470	16	6	6	2
13	200	4	19	6	3
14	181	8	10	18	11
Average		16	14	14	10

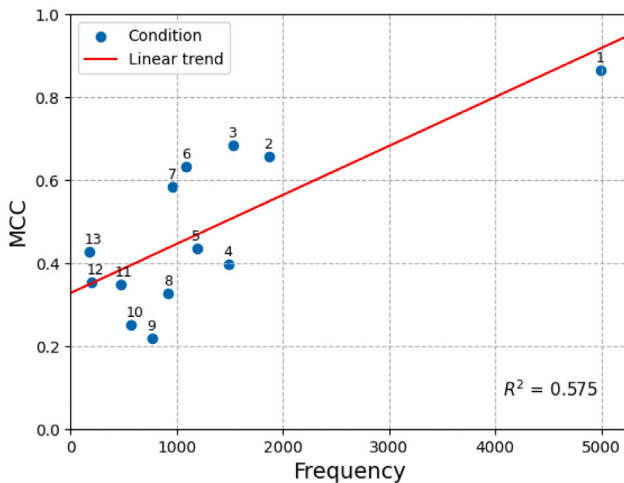


Fig. 10. Scatter plot of the MCC results on the test set. The red line shows the MCC increasing trend according to the frequency of the dental condition in the dataset.

The test set's results in Table 5 were derived from the top-performing network based on the validation sets. Notably, while the test MCC values exhibit considerable variation, they all exceed 0

(zero), indicating performance better than random guessing. The results demonstrate a clear positive trend in MCC values as the sample size of positive examples increases. This relationship is illustrated in Fig. 10, which presents a scatter plot of the number of positive examples (frequency) against the corresponding MCC values. To better visualize this trend, we applied a linear regression model to the data, yielding an R^2 value of 0.575. With an R^2 values between 0.5 and 0.7, the model successfully captures a significant portion of the variance in the data, affirming its reliability.

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

1. Endodontic treatment (less context)

An endodontic treatment appears as white (radiopaque) lines in the tooth canals (refer to Fig. 9(a)). Therefore, an image crop close and centered on the teeth eases the task of identifying this condition. This configuration is the case for the 224×224 crops. Together with the large amount of positive data, this resulted in a MCC higher than 0.900 on the validation data, while the resize crops from 380×380 dimensions reached 0.845 MCC.

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

Coronal destruction appears as darker areas (radiolucencies) because the structure is less dense than a healthy tooth. This decay, which can be caused by caries or other factors, can be seen as disruptions in the continuous outline of the tooth crown, especially around or underneath existing dental restorations. Therefore, a close, near-the-tooth crop is sufficient for detecting coronal destruction, as depicted in Fig. 9(b). In this case, a maximum of 0.714 was reached from the less-context crops against 0.675 of the more-context one.

3. Included and impacted (**more context**)

It refers to a tooth that has not erupted into its expected position in the dental arch due to obstruction by another tooth, bone, or soft tissue (Fig. 9(c)). This phenomenon occurs frequently with wisdom teeth (third molars). Depending on its location, the impacted wisdom tooth can be seen pressing against or tilted towards its neighboring second molar, potentially causing root resorption or displacement. Therefore, a minimal increase in the context of the tooth crop may be beneficial to identify inclusions. The maximum attained results from more-context tooth crops (0.790) were 2.3 (p.p.) higher than the less-context counterpart (0.767).

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

It appears as a darker area (radiolucent) around the tooth's root or at its apex (see Fig. 9(d)). This dark spot indicates bone loss or decreased bone density. The borders of this area can be well-defined or more diffuse, depending on the nature and stage of the condition. Under these circumstances, having a well-focused image around the teeth is better, or even necessary, to better diagnose this condition. Here, the less context reached 0.599 against 0.523.

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

On a panoramic radiograph, unfilled root canals within a tooth appear as relatively dark lines or canals within the lighter, radiopaque outline of the tooth structure (refer Fig. 9(e)). These dark lines represent where the dental pulp once was and should generally be filled with endodontic materials if a root canal treatment has been completed. In those conditions, a well-focused image is better for classification. In the current setup, the networks trained with more-context crops attained an MCC of 0.611 and the less-context crops of 0.454.

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

On a panoramic radiograph, a “metallic core” within a tooth appears as a highly radiopaque area within the tooth structure, often in the shape of a post or a dense filling (see Fig. 9(f)). It stands out distinctly against the less dense surrounding tooth material and any dental restorations that are not metal-based. However, since it is not a part of the tooth, the crop center excludes the metallic core “crown”. More context is expected to help the classification in light of this situation. A value of 0.75 for MCC in the more-context configuration and 0.695 in the less-context was reached.

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

On a panoramic radiograph, a “root fragment” appears as a radiopaque structure, resembling a part of a tooth's root, as shown in Fig. 9(g). It is usually isolated, without a crown portion, and may be surrounded by a darker area if inflammation or bone resorption is present. The need to screen the tooth's surroundings makes having more context in the image important. Results of 0.728 and 0.532 were achieved in less-context and more-context scenarios, respectively.

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

On a panoramic radiograph, an increased apical periodontal space, displayed in Fig. 9(h), appears as an enhanced or widened radiolucent area around the tip of the root of a tooth. This dark gap, known as the periodontal ligament space, is usually uniform

and thin around the roots of healthy teeth. Bearing this in mind, a closed, well-focused image around the tooth's center may exclude the dental condition, making its diagnosis impossible. Therefore, an image with more context is more beneficial for detecting an increased apical periodontal space. The current study reached a 0.405 MCC in the more context scenario against a 0.275 in the less context (i.e., a performance boost of 47.27%).

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

On a panoramic radiograph, a “trabecular bone modification” may appear as changes in the pattern and density of the bone, as shown in Fig. 9(i). Areas with increased density will look whiter, indicating a more solid bone structure, while regions with decreased density will appear darker, suggesting less bone mass. The regular mesh-like pattern of the trabeculae might appear disrupted or altered, which can indicate various dental or bone conditions. These areas appear on the bones surrounding the teeth, not near their center. Therefore, a crop with more context is beneficial for diagnosing trabecular bone modification. In the more-context scenario, we attained 0.506 of MCC; in the less context, we attained 0.309 on the validation datasets.

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

On a panoramic radiograph, “extensive restoration”, or “excess restorative material”, appears as a filling, crown, or other dental work, that extends beyond the natural contours of the tooth, as displayed in Fig. 9(j). Typically used for fillings or crowns, these materials will stand out as they are denser than the tooth and absorb more X-rays. If overfilled, excess material may also be seen beyond the confines of the tooth's normal borders, such as in the interdental spaces or the pulp chamber. Restorative material frequently occurs on the tooth crown, a distance from the tooth center. Therefore, cropping too closely may hinder accurate diagnosis. In the current analysis, the images were not excessively cropped, which allowed for a correct diagnosis with less contextual information. Specifically, an MCC of 0.385 was observed in scenarios with less context compared to 0.314 in scenarios with more context.

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

On a dental radiograph, “idiopathic osteosclerosis” appears as a brighter area due to increased bone density (Fig. 9(k)). It is often seen near the roots of teeth but lacks the characteristic dark border of other lesions. Therefore, the view of the tooth's surroundings could be beneficial for detecting it. The results were 0.424 in the more-context scenario and 0.182 in the less-context one.

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

An “unfavorable positioning for eruption” for a tooth appears as a tooth that is misaligned with the normal arch form, often at an abnormal angle or location that suggests it will not erupt into a functional position without intervention (see Fig. 9(m)). This could be a tooth that is tilted, rotated, or horizontally displaced. The context around the tooth is important to verify and confirm if its position is unfavorable for eruption. Indeed, the results in the more-context scenario were 0.456 against 0.420 in the less-context scenario (an increase of 8.57%).

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

“Prolonged retention” of a tooth is indicated by a tooth that remains in the jaw beyond the typical age of exfoliation without evidence of natural shedding or eruption (Fig. 9(n)). It often appears as a tooth with roots that may be resorbed, situated in the jaw without movement, potentially affecting the positioning of adjacent teeth or the eruption of successor permanent teeth. These deciduous teeth are small and do not require a larger context for diagnosis of prolonged retention. A value of 0.666 was reached in the context scenario against 0.545 in the more context one.

4. Comparative study with dentistry student and specialists

Performance evaluation on a large and diverse dataset indicates that the proposed framework was effective to learn how to classify the dental conditions, as all MCCs exceeded 0 (zero). However, these values do not provide a desirable comparison with the performance of human specialists when evaluating panoramic radiographs. To make this comparison, annotations made by dental specialists were assessed and compared against the results of the classification models' predictions. This was accomplished by inviting five final-year undergraduate students and five radiologist specialists (all of them with more than 10 years of experience) to label some samples of the same test images used to evaluate the classification models.

In the labeling setup, each participant had to label a cropped panoramic radiograph centered on a specific tooth, similar to the images used for training and evaluating the models. The images in the setup were the same as the "more-context" (380×380) crops, but without resizing to 224×224 , which was previously necessary to meet the model input requirements. The dentistry students and specialists had to identify and mark all visible dental conditions in the area of the tooth indicated in the center of the image crop, based on the provided options (all the conditions considered in this study), or mark none, according to their analysis.

A hurdle that needed to be overcome in this procedure was the number of test images to be annotated. According to estimates, there were about 32,000 test images (please refer to Table 3), which would require more than 300 h of continuous work for each participant to label, making it impractical. A natural alternative was to sample a subset of the test set while maintaining the positive/negative class proportions. Unfortunately, this option also proved unworkable due to the highly imbalanced datasets. For instance, the positive/negative ratio for condition 13 (prolonged retention) is 0.106. In this case, the students and specialists would need to annotate approximately 1180 samples to maintain the proportional ratio, with only one being positive. Annotating such a large number of image crops was beyond reasonable feasibility. The issue was overcome by selecting images through a strategy that ensured a minimum number of positive examples and variability. The adopted strategy consisted of selecting 78 samples (six images per condition) using the following pattern: for each condition, two true positives (TP), two false positives (FP), and two false negatives (FN) were selected based on the original models predictions. According to the labels extracted from the reports, this approach ensured at least four positive samples (the two TPs and the two FNs) and two negative samples (FP). It also provided potential variability due to the FP and FN typically being borderline cases. This final set of images is designated as **specialist Image Dataset**.

4.1. Initial assessment

Table 7 shows the results for each annotator and the average results for the students and specialists, considering the labels from the text reports as the ground truth. The outcomes indicate that the specialist group performed moderately better than the students (0.429 vs. 0.455 MCC, with a tiny standard deviation among each group (0.05 and 0.07). The attained MCC by the used models was 0.475 (Table 5), which is higher than the scores of both groups, demonstrating strong performance. However, one cannot assert that the models have reached superhuman performance because the scenario is biased in favor of the models. Rather than being trained to detect dental conditions in general, the models were trained to detect conditions as the primary labeler (reports), giving them an advantage over annotators who did not have access to the labeled samples. This issue was mitigated by combining the specialist labels, as discussed next.

Table 7

Average MCC for each student and specialist (the average value for each group is also included.)

Student	Avg. MCC	Specialists	Avg. MCC
Student 1	0.479	Spec. 1	0.394
Student 2	0.500	Spec. 2	0.589
Student 3	0.360	Spec. 3	0.387
Student 4	0.435	Spec. 4	0.455
Student 5	0.372	Spec. 5	0.452
All students	0.429 $\pm$ 0.05	All specialists	0.455 $\pm$ 0.07

Table 8

Final average MCC of all conditions for each student and specialist, including the average value for each group.

Student	Avg. MCC	Specialists	Avg. MCC
Student 1	0.527	Spec. 1	0.607
Student 2	0.591	Spec. 2	0.689
Student 3	0.490	Spec. 3	0.499
Student 4	0.516	Spec. 4	0.575
Student 5	0.426	Speci. 5	0.574
All students	0.510 $\pm$ 0.05	All specialists	0.589 $\pm$ 0.06

4.2. Definitive assessment with specialist consensus

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

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

Under this new layout, the results of the models and the average results of the students and specialists were 50.8%, 51.0%, and 58.9%, respectively. The detailed results of each participant are included in Table 8. It is possible to draw two conclusions from that summary. Firstly, the MCC values are considerably higher than those attained on the reports' ground truth. This increase can be attributed to the more robust ground truths that were less noisy, as the combination of the labels of several specialists generated them. Furthermore, the similarity between the MCCs attained by the proposed solution and the students (50.82% vs. 51.00%) led to the conclusion that the solution reached the level of a final-year undergraduate student while reaching a precision level within 8.1 percentage points of specialists. This result showcases the potential of the approach as a reliable tool for supporting or improving specialist-level diagnoses.

Fig. 11 depicts a bar chart that further investigates the results, breaking them down by classes and predictors (models, annotators' average, students, and specialists). One can observe that the proposed solution, when compared to the annotators, demonstrates significantly higher performance in Condition 5 (unfilled root canals) and considerably worse performance in Condition 12 (unfavorable positioning for eruption) and 13 (prolonged retention), the ones of less positive samples. However, what stands out the most is the almost null MCC values for condition 9 (trabecular bone modification) class. One can hypothesize that this result stemmed from a lack of agreement among this class's specialists. It can be expected that the higher the agreement between the annotators, the higher the MCC of the model is attained. Consequently, it was decided to conduct a statistical agreement analysis on the ground truth labels.

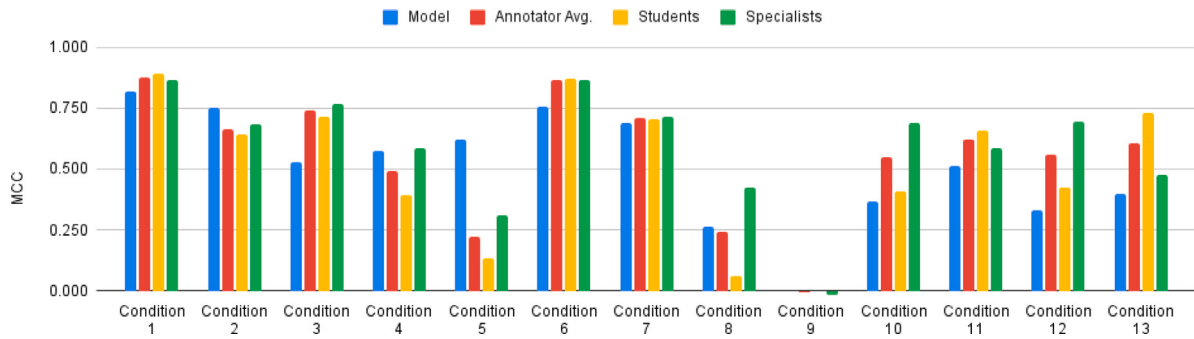
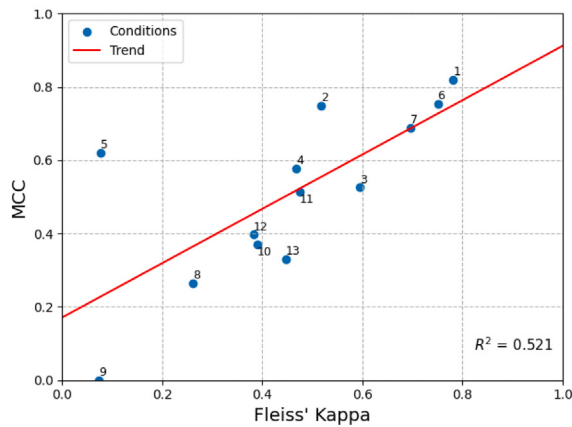
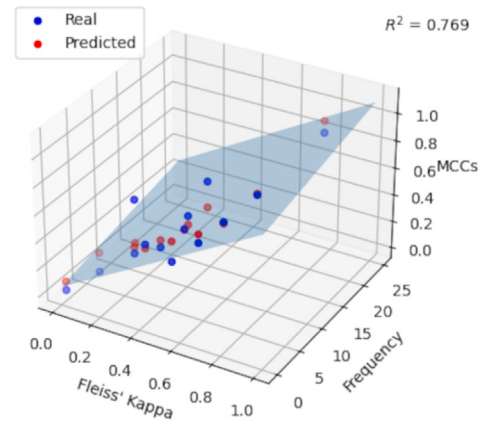


Fig. 11. Bar chart that breaks down the definitive assessment results according to condition and annotator' group. The results on Condition 9 (trabecular bone modification) were closer to 0 (zero).



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



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

Fig. 12. Plots showing the linear trends of MCC results based on Fleiss' Kappa and the frequency of positive samples for each condition.

Table 9

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

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

4.3. Statistical agreement analysis

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

Table 9 contains the frequency of positive samples for each condition, the attained MCC values of the models, and the computed Fleiss' kappa values in the specialist Image Dataset. An inspection of the table's data shows that the hypothesis made holds true: the lack of agreement among the specialists on condition 9 (Kappa of 0.045) was connected to the poor performance of all groups (model, students, and specialists) on the same condition (MCC of 0 (zero)), indicating that the models struggled to learn from inconsistent labels. The disagreement among specialists regarding modifications in trabecular bone structure stems from the complexity and inherent subjectivity of evaluating such changes. Trabecular bone modifications can manifest subtly and may overlap with normal anatomical variations or early stages of several conditions, making their recognition and classification challenging. Additionally, this condition is not exclusive to a single

pathology. Changes in the trabecular bone can indicate a wide range of issues, from osteoporosis to localized effects of dental infections. The multifactorial nature of these modifications may result in differing interpretations among specialists depending on their clinical background and experience.

Table 9 also indicates substantial agreement among the labelers for conditions 1 and 6 (Kappas of 0.776 and 0.750), which are the conditions where the models attained their best results. These results suggest a correlation between the models' performance and the level of agreement among the labelers.

Fig. 12(a) is a scatter plot of the attained MCC results against the computed kappas. The value of R^2 reached 0.521. Similar to the correlation between the Frequency of Positive Samples and MCC of Fig. 10, this correlation was not expected to be flawless. Instead, it was aimed to demonstrate a trend – specifically, an increasing one – where a higher kappa corresponds to a higher MCC of the proposed solution.

The correlation between the independent variables (kappa values and positive sample frequency) and the dependent variable (attained MCC) was further investigated. Fig. 12 illustrates the final result. With these two independent variables, R^2 reaches 0.769, a value considered a good fit.

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

5. Concluding remarks

Research on panoramic radiographs has traditionally focused on supervised learning; however, such methods are becoming increasingly impractical due to their dependence on extensive annotated data. This limitation underscores the need for alternative learning paradigms, which form the foundation of our work. Our approach demonstrated that integrating parts-based instance segmentation with an automatic extraction of class labels from dental reports effectively mitigates the challenges of large-scale annotation. By leveraging the natural pairing of radiographs and textual reports, our method bypasses the need for manual annotation of massive image datasets. The strategy of segmenting teeth and amplifying regions of interest in cropped images highlights the critical role of contextual information in improving classifier performance. Indeed, crops with more contextual information yielded better results. However, the dimensions of the cropped regions were determined empirically, pointing to the potential for systematic optimization to further enhance model accuracy. Interestingly, in certain cases, increased contextual information did not translate to improved performance, which warrants further exploration.

Rather than merely benchmarking methods, we validated our approach through real-world experiments involving final-year dental students and specialists. By sampling a representative number of images from our dataset, we demonstrated that our model's performance is comparable to that of final-year dental students, offering promising evidence of its practical applicability. Future directions include advancing the integration of features extracted simultaneously from both textual and image data to achieve a more synergistic representation of these two modalities.

CRediT authorship contribution statement

Bernardo Silva: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jefferson Fontinele:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Carolina Letícia Zilli Vieira:** Writing – review & editing, Validation, Methodology, Conceptualization. **João Manuel R.S. Tavares:** Writing – review & editing, Methodology. **Patricia Ramos Cury:** Writing – review & editing, Validation, Data curation. **Luciano Oliveira:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Formal analysis, Conceptualization.

Funding

Bernardo Silva was supported by the Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB), Brazil under grant BOLA0569/2020, and Luciano Oliveira and Patrícia Cury were supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil under grants 303974/2021-4 and 402796/2021-7, respectively.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

References

- Alahmad, H., Alnafaie, M., 2023. Survey of quality control of panoramic X-ray machines in private dental clinics in Saudi Arabia. *J. Radiat. Res. Appl. Sci.* 16 (2), 100571.
- Amasya, H., Jaju, P.P., Ezhov, M., Gusarev, M., Atakan, C., Sanders, A., Manulius, D., Golitskya, M., Shrivastava, K., Singh, A., et al., 2024. Development and validation of an artificial intelligence software for periodontal bone loss in panoramic imaging. *Int. J. Imaging Syst. Technol.* 34 (1), e22973.
- Bonfanti-Gris, M., Garcia-Canas, A., Alonso-Calvo, R., Rodriguez-Manzanique, M.P.S., Ramiro, G.P., 2022. Evaluation of an artificial intelligence web-based software to detect and classify dental structures and treatments in panoramic radiographs. *J. Dent.* 126, 104301.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* 33, 1877–1901.
- Chang, H.-J., Lee, S.-J., Yong, T.-H., Shin, N.-Y., Jang, B.-G., Kim, J.-E., Huh, K.-H., Lee, S.-S., Heo, M.-S., Choi, S.-C., et al., 2020. Deep learning hybrid method to automatically diagnose periodontal bone loss and stage periodontitis. *Sci. Rep.* 10 (1), 7531.
- Chen, H., Li, H., Zhao, Y., Zhao, J., Wang, Y., 2021. Dental disease detection on periapical radiographs based on deep convolutional neural networks. *Int. J. Comput. Assist. Radiol. Surg.* 16, 649–661.
- Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., et al., 2019. Hybrid task cascade for instance segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4974–4983.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Ekert, T., Krois, J., Meinhold, L., Elhennawy, K., Emara, R., Golla, T., Schwendicke, F., 2019. Deep learning for the radiographic detection of apical lesions. *J. Endod.* 45 (7), 917–922.
- Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. *Psychol. Bull.* 76 (5), 378.
- Fukuda, M., Inamoto, K., Shibata, N., Arijii, Y., Yanashita, Y., Kutsuna, S., Nakata, K., Katsumata, A., Fujita, H., Arijii, E., 2020. Evaluation of an artificial intelligence system for detecting vertical root fracture on panoramic radiography. *Oral Radiol.* 36, 337–343.
- Gao, L., Xu, T., Liu, M., Jin, J., Peng, L., Zhao, X., Li, J., Yang, M., Li, S., Liang, S., 2024. AI-aided diagnosis of oral X-ray images of periapical films based on deep learning. *Displays* 82, 102649.
- Haring, J.I., Jansen, L., 2000. *Dental Radiography: Principles and Techniques*. WB Saunders Philadelphia.
- He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 16000–16009.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 770–778.
- Jader, G., Fontineli, J., Ruiz, M., Abdalla, K., Pithon, M., Oliveira, L., 2018. Deep instance segmentation of teeth in panoramic X-ray images. In: *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images. SIBGRAPI, IEEE*, pp. 400–407.
- Jing, B., Xie, P., Xing, E., 2017. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*.
- Khan, H.A., Haider, M.A., Ansari, H.A., Ishaq, H., Kiyani, A., Sohail, K., Muhammad, M., Khurram, S.A., 2021. Automated feature detection in dental periapical radiographs by using deep learning. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* 131 (6), 711–720.
- Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M., 2022. Transformers in vision: A survey. *ACM Comput. Surv.* 54 (10s), 1–41.
- Krois, J., Cantu, A.G., Chaurasia, A., Patil, R., Chaudhari, P.K., Gaudin, R., Gehrung, S., Schwendicke, F., 2021. Generalizability of deep learning models for dental image analysis. *Sci. Rep.* 11 (6102).
- Kwon, O., Yong, T.-H., Kang, S.-R., Kim, J.-E., Huh, K.-H., Heo, M.-S., Lee, S.-S., Choi, S.-C., Yi, W.-J., 2020. Automatic diagnosis for cysts and tumors of both jaws on panoramic radiographs using a deep convolution neural network. *Dentomaxillofacial Radiol.* 49 (8), 20200185.
- Langlais, R.P., Miller, C., 2016. *Exercises in oral radiology and interpretation-e-book: Exercises in oral radiology and interpretation-e-book*. Elsevier Health Sciences.
- Lee, J.-H., Kim, D.-H., Jeong, S.-N., 2020. Diagnosis of cystic lesions using panoramic and cone beam computed tomographic images based on deep learning neural network. *Oral Dis.* 26 (1), 152–158.

- Liu, F., Gao, L., Wan, J., Lyu, Z.-L., Huang, Y.-Y., Liu, C., Han, M., 2023. Recognition of digital dental X-ray images using a convolutional neural network. *J. Digit. Imaging* 36 (1), 73–79.
- Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. et Biophys. Acta (BBA)- Protein Struct.* 405 (2), 442–451.
- Névél, A., Dalianis, H., Velupillai, S., Savova, G., Zweigenbaum, P., 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *J. Biomed. Semant.* 9, 1–13.
- OpenAI, 2023. *GPT-4 technical report*. arXiv:2303.08774.
- Pinheiro, L., Silva, B., Sobrinho, B., Lima, F., Cury, P., Oliveira, L., 2021. Numbering permanent and deciduous teeth via deep instance segmentation in panoramic x-rays. In: *17th International Symposium on Medical Information Processing and Analysis*. Vol. 12088, SPIE, pp. 95–104.
- Ranjbar, F.R., Zamanifar, A., 2023. Autonomous dental treatment planning on panoramic x-ray using deep learning based object detection algorithm. *Multimedia Tools Appl.* 1–35.
- Silva, G., Oliveira, L., Pithon, M., 2018. Automatic segmenting teeth in X-ray images: Trends, a novel data set, benchmarking and future perspectives. *Expert Syst. Appl.* 107, 15–31.
- Silva, B., Pinheiro, L., Oliveira, L., Pithon, M., 2020. A study on tooth segmentation and numbering using end-to-end deep neural networks. In: *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images*. SIBGRAPI, IEEE, pp. 164–171.
- Silva, B.P.M., Pinheiro, L.B., Sobrinho, B.P.P., Lima, F.P., Sobrinho, B.P.P., Abdalla Buzar Lima, K., Pithon, M.M., Cury, P.R., Oliveira, L.R.d., 2023. Boosting research on dental panoramic radiographs: a challenging data set, baselines, and a task central online platform for benchmark. *Comput. Methods Biomech. Biomed. Eng.: Imaging & Vis.* 1–21.
- Tassoker, M., Öziç, M.Ü., Yuce, F., 2024. Performance evaluation of a deep learning model for automatic detection and localization of idiopathic osteosclerosis on dental panoramic radiographs. *Sci. Rep.* 14 (1), 4437.
- Vinayahalingam, S., Goey, R.-s., Kempers, S., Schoep, J., Cherici, T., Moin, D.A., Hanisch, M., 2021. Automated chart filing on panoramic radiographs using deep learning. *J. Dent.* 115, 103864.
- Whaites, E., Drage, N., 2013. *Essentials of Dental Radiography and Radiology*. Elsevier Health Sciences.
- White, S.C., Pharoah, M.J., 2014. *Oral radiology-E-Book: Principles and interpretation*. Elsevier Health Sciences.
- Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1492–1500.
- Yüksel, A.E., Gültekin, S., Simsar, E., Özdemir, Ş.D., Gündoğar, M., Tokgöz, S.B., Hamamcı, İ.E., 2021. Dental enumeration and multiple treatment detection on panoramic X-rays using deep learning. *Sci. Rep.* 11 (1), 12342.