BMC Oral Health

RESEARCH

Open Access

Combining public datasets for automated tooth assessment in panoramic radiographs

Niels van Nistelrooij[1,3], Khalid El Ghoul[4], Tong Xi[2], Anindo Saha[2], Steven Kempers[1], Max Cenci[6], Bas Loomans[6], Tabea Flügge[3,5*], Bram van Ginneken[2] and Shankeeth Vinayahalingam[1]

Abstract

Objective Panoramic radiographs (PRs) provide a comprehensive view of the oral and maxillofacial region and are used routinely to assess dental and osseous pathologies. Artificial intelligence (AI) can be used to improve the diagnostic accuracy of PRs compared to bitewings and periapical radiographs. This study aimed to evaluate the advantages and challenges of using publicly available datasets in dental AI research, focusing on solving the novel task of predicting tooth segmentations, FDI numbers, and tooth diagnoses, simultaneously.

Materials and methods Datasets from the OdontoAI platform (tooth instance segmentations) and the DENTEX challenge (tooth bounding boxes with associated diagnoses) were combined to develop a two-stage AI model. The first stage implemented tooth instance segmentation with FDI numbering and extracted regions of interest around each tooth segmentation, whereafter the second stage implemented multi-label classification to detect dental caries, impacted teeth, and periapical lesions in PRs. The performance of the automated tooth segmentation algorithm was evaluated using a free-response receiver-operating-characteristics (FROC) curve and mean average precision (mAP) metrics. The diagnostic accuracy of detection and classification of dental pathology was evaluated with ROC curves and F1 and AUC metrics.

Results The two-stage AI model achieved high accuracy in tooth segmentations with a FROC score of 0.988 and a mAP of 0.848. High accuracy was also achieved in the diagnostic classification of impacted teeth (F1 = 0.901, AUC = 0.996), whereas moderate accuracy was achieved in the diagnostic classification of deep caries (F1 = 0.683, AUC = 0.960), early caries (F1 = 0.662, AUC = 0.881), and periapical lesions (F1 = 0.603, AUC = 0.974). The model's performance correlated positively with the quality of annotations in the used public datasets. Selected samples from the DENTEX dataset revealed cases of missing (false-negative) and incorrect (false-positive) diagnoses, which negatively influenced the performance of the AI model.

Conclusions The use and pooling of public datasets in dental AI research can significantly accelerate the development of new AI models and enable fast exploration of novel tasks. However, standardized quality assurance is essential before using the datasets to ensure reliable outcomes and limit potential biases.

Keywords Panoramic radiograph, Artificial intelligence, Public datasets, Tooth segmentation, Diagnostic classification

*Correspondence:
Tabea Flügge
tabea.fluegge@charite.de

Full list of author information is available at the end of the article

Introduction

Panoramic radiographs (PRs) provide a two-dimensional (2D) radiographic view of the upper and lower jaw, including the teeth and adjacent osseous structures. PRs are commonly used in dentistry and oral and maxillofacial surgery for diagnostic purposes due to their easy acquisition, limited radiation exposure, and comprehensive field of view. Although frequently used for assessing tooth impaction and identifying cysts, tumors, and other bony or osteolytic pathologies, the diagnostic accuracy of PRs is limited by interobserver variability and the 2D representation of complex 3D maxillofacial structures [1–3].

Pathologies of odontogenic origin, such as caries, periapical lesions, and impacted teeth, are routinely diagnosed through clinical and radiographic assessment [4]. A bitewing or a periapical radiograph can be acquired to obtain a high-resolution view centered on the crowns and/or roots of teeth. However, these radiographs have a limited field of view and can be challenging to obtain for certain patients (e.g. small mouth, gag reflex), so that acquiring a PR is envisaged in such cases. While conventional PRs demonstrated limited efficacy in the diagnosis of caries and associated pathologies [5], a potential improvement in their diagnostic accuracy through artificial intelligence (AI) has been suggested [6]. Previous studies have reported on applying convolutional neural networks (CNNs) for the automated segmentation and labeling of dentition on PRs [7–9]. Using transfer learning and transformer-based models, other studies achieved further improvement in the accuracy of these tasks, including the detection and segmentation of teeth and associated numbering [10, 11].

Few studies have performed concurrent tooth detection and classification of caries or periapical lesions on PRs [12–14]. Lower resolution in PRs has been identified as an explanation for their underperformance compared to bitewings and periapical radiographs [15]. Nevertheless, a recently proposed AI method has shown diagnostic capabilities comparable to dentists with 3 to 10 years of experience in diagnosing tooth pathologies while significantly reducing assessment time [6]. These algorithms show considerable potential in improving the detection rates of various pathologies while reducing the work-load associated with radiographic examination [16].

To accelerate the development and benchmarking of AI techniques, several studies have made their PRs and corresponding annotations publicly available in data depositories [13, 17–21]. These public datasets vary in size, from 115 to 4,000 unique PRs, with the annotations ranging from mandible segmentations to the segmentation and labeling of teeth and abnormalities. Despite the availability of these datasets, the adoption of these datasets in dental AI research still needs to be improved. Most studies rely on an in-house dataset, which requires a considerable time investment for annotation. The use of in-house datasets also makes comparisons between studies more difficult, as the annotation guideline to construct each dataset may differ and the local population may be overrepresented.

Therefore, the current study combined two public datasets (OdontoAI [19], DENTEX [21]) to develop an automated method for the novel task of concurrent tooth segmentation, FDI labeling, and diagnosis classification, including caries, impacted tooth, and periapical lesions. This study aimed to evaluate the advantages and challenges of using publicly available datasets in dental AI research, focusing on improving the diagnostic accuracy of caries, impacted teeth, and periapical lesions in PRs.

Methods

This study was conducted following the code of ethics of the World Medical Association (Declaration of Helsinki) and the checklist of artificial intelligence (AI) in dental research has been consulted for reporting [22]. No informed consent was required as all image data were publicly available and were anonymized.

Data

A dataset with tooth segmentations (OdontoAI) was pooled with another dataset comprising tooth bounding boxes and associated diagnoses (DENTEX) to develop a two-stage AI method to segment, label, and classify teeth and odontogenic pathologies, such as caries, impacted teeth, and periapical lesions in PRs.

OdontoAI: The OdontoAI platform provides a public dataset with 4,000 PRs, of which 2,000 are annotated [19]. The annotations include tooth segmentations with corresponding FDI numbers. The FDI notation describes the location of a tooth by the quadrant in which the tooth resides (1–4) and the numerical order of individual teeth within the quadrant from midline to the back (1–8). PRs were excluded based on the following exclusion criteria: the presence of an artefact ($n=1$), incorrect or missing tooth segmentation ($n=85$), edentulous jaws ($n=15$) or the presence of a mixed dentition ($n=238$). The remaining 1,661 PRs were split into 1,337 PRs for training and validation and 324 PRs for testing using multi-label stratification based on the FDI numbers [23].

DENTEX: The Dental Enumeration and Diagnosis on Panoramic X-rays (DENTEX) challenge provides a public dataset with 705 unique PRs with tooth diagnosis annotations [21, 24]. More specifically, each diagnosed tooth is annotated within a bounding box and

labeled as impacted, early caries, deep caries, and/or periapical lesion. Additionally, a subset of 260 PRs includes bounding boxes for teeth without a diagnosis (without pathology). PRs were excluded based on the following exclusion criteria: the presence of an artifact ($n=8$), incorrect or missing tooth annotations ($n=12$), or the presence of a mixed dentition ($n=3$). The remaining 682 PRs were split into 548 PRs for training and validation and 134 PRs for testing using multi-label stratification based on the FDI numbers and tooth diagnoses [23].

Deep learning method

The current method consists of two stages: tooth segmentation and multi-label diagnosis classification. See Fig. 1 for an overview of the method. The first stage for tooth instance segmentation predicted tooth segmentations with corresponding FDI numbers. A region of interest (ROI) was extracted from a PR based on the tooth segmentation; and the second stage for multi-label diagnosis classification predicted multiple abnormalities of the tooth. Training and inference was performed on a workstation with 128GB of system memory and an RTX A6000 GPU with 48GB of memory.

Tooth segmentation

Mask DINO was used for tooth segmentation (Fig. 1b). This is a recent end-to-end deep learning pipeline for instance segmentation using vision transformers [25]. For this study, Mask DINO was implemented using the MMDetection framework (v3.1.0) based on PyTorch 2.0.1 [26, 27].

The model was initialized with a ResNet-50 backbone [28] and was pre-trained on the COCO dataset [29]. The pre-trained model was fine-tuned for a maximum of 50 epochs using the AdamW optimizer with a weight decay of 0.05 [30]. The initial learning rate was 10^{-4} and subsequently adjusted, divided by 10 after epochs 44 and 48. Data augmentation included flipping, resizing, and copy-and-pasting of a tooth on top of the contralateral tooth with the same tooth number [31]. PRs were processed in mini-batches of two. A multi-task loss function was used to supervise the predicted bounding boxes, segmentations, and labels.

For inference, test-time augmentation was applied in the form of flipping, while non-maximum suppression was used to select the final tooth predictions with a set threshold of 0.1 [32]. The minimum bounding box around a tooth segmentation was used as the bounding box for this tooth.

Tooth ROI extraction

The fine-tuned Mask DINO model was used to predict all teeth in the PRs from the DENTEX challenge. These tooth predictions were used as inputs for the subsequent diagnosis classification stage.

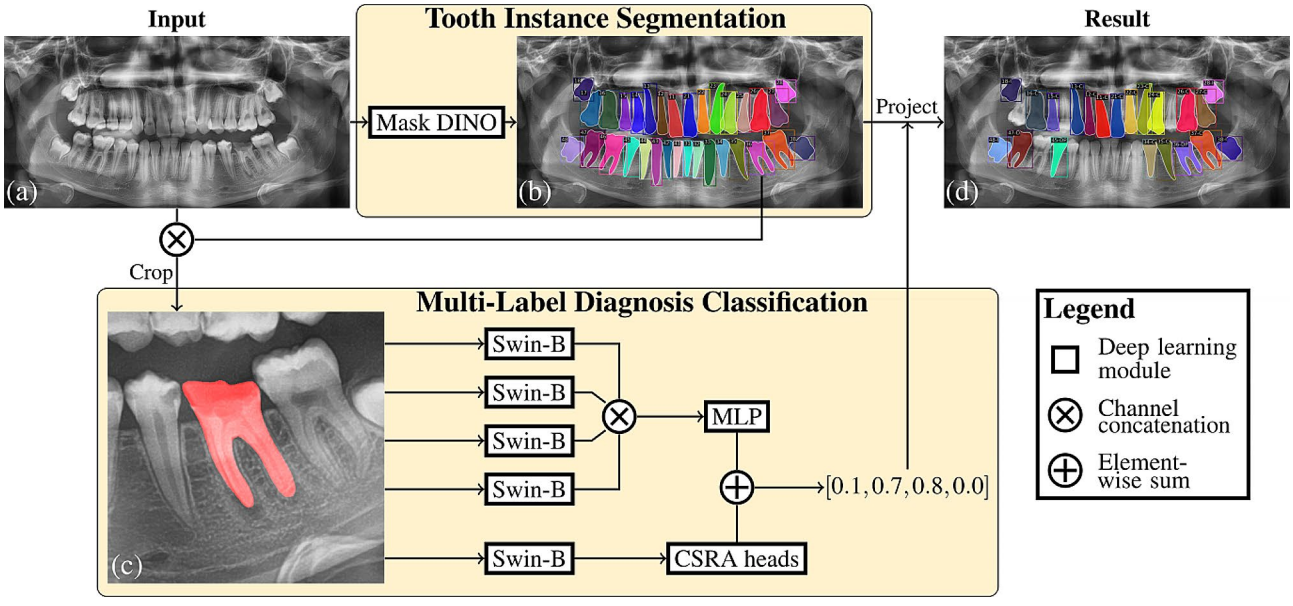

Fig. 1 Overview of methodology. The teeth in the input PR **(a)** are segmented by Mask DINO and their FDI numbers are predicted**(b)** [25]. For each predicted tooth, a cropped image is made with the segmentation as extra channel **(c, ROI extraction)**. This image is processed by four binary classifiers, one for each diagnosis, whose predictions are aggregated using an MLP [34], and by a multi-label classifier who returns multiple predictions via its CSRA heads [40]. All multi-label predictions are summed and these final scores are used to add diagnoses to the tooth segmentations **(d)**. Non-diagnosed teeth are predicted, but not shown in the result for clarity. The label is the tooth's FDI number with C = early caries, D = deep caries, P = periapical lesion, I = impacted

A predicted tooth bounding box was matched with the reference tooth bounding box with a maximum intersection over union (IoU) of at least 0.25. The diagnosis labels of a predicted tooth were assigned based on the diagnosis labels of the matched reference tooth. Predicted teeth that were not matched were excluded, which resulted in 5,887 non-diagnosed teeth, 593 impacted teeth and 2,110 teeth with early caries, 536 teeth with deep caries, and 150 teeth with a periapical lesion. For each matched tooth, a classification image was generated (Fig. 1c). An additional color channel representing the tooth's binary segmentation was added to the grayscale PR. This two-channel image was then cropped around the tooth segmentation with a margin of 10%. This margin provided extra contextual information to the classification model, improving its diagnostic effectiveness.

Multi-label diagnosis classification

The classification stage was implemented using MMPreTrain (v1.0.1) based on PyTorch 2.0.1 [27, 33]. Classification images extracted for each tooth in the PRs from the DENTEX challenge were used for training and evaluation (Fig. 1c).

Pre-training A Swin-B backbone [34] was first pre-trained on the ImageNet dataset [35] using the SimMIM method [36]. SimMIM is a self-supervised pre-training technique that removes parts of the input image and predicts the missing pixels. This allows it to effectively model the relationships between foreground objects and their context objects. After pre-training on ImageNet for 800 epochs, the Swin-B backbone underwent further pre-training on the train/validation classification images for an additional 100 epochs.

Binary classification Four binary classifiers were trained to distinguish diagnosed teeth from non-diagnosed teeth for each diagnosis. Each classifier comprised a Swin-B backbone followed by a fully-connected layer. The backbone was initialized using the pre-trained model parameters and each classifier was fine-tuned for a maximum of 80 epochs using the AdamW optimizer with a weight decay of 0.05 [30]. The learning rate increased linearly to 0.0002 during the first 5 epochs and subsequently followed a cosine annealing schedule. Data augmentation included flipping, resizing, spatial and intensity transformations [37], and copy-and-pasting a tooth to another classification image with the same FDI number [31]. PRs were processed in mini-batches of 256 and the predictions were supervised using the cross-entropy loss function with label smoothing [38]. Class imbalance was addressed by sampling diagnosed and non-diagnosed teeth equally [39].

Multi-label classification The four binary classifiers were frozen and a multi-layer perceptron (MLP) was used to aggregate their predictions into four diagnostic probabilities. A fifth pre-trained Swin-B backbone was used to learn visual features for multi-label classification. The final feature maps of the Swin-B backbone were further processed by the class-specific residual attention (CSRA) module [40]. This module employed multi-head self-attention to focus on multiple locations of the input image simultaneously. Each head of the CSRA module predicted four diagnostic probabilities. The predictions from the binary classifiers and the predictions from the CSRA module are aggregated by element-wise summing. The same training setup was used as in the binary classification stage, with the exception that class imbalance was addressed by more frequent sampling of classification images with rare diagnoses [41]. Furthermore, the focal loss function was used [42].

Inference During the inference stage, test-time augmentation was applied with flipping, and the predictions were aggregated by averaging the diagnostic probabilities. The probabilities were updated to incorporate prior knowledge. More specifically, each diagnostic probability was multiplied by the score of the tooth segmentation. Mutual exclusions were downscaled between early caries and deep caries, as well as between impacted teeth and other diagnoses, and between impacted teeth and teeth other than third molars. Finally, the predicted diagnoses of the extracted ROI were projected back to the same tooth of the input PR (Fig. 1d).

Evaluation and comparison

Five-fold cross-validation splits were determined for the train/validation PRs from OdontoAI and DENTEX datasets using multi-label stratification [23]. Both stages are trained five times given a different cross-validation split to investigate the variability of the method's results. Both the tooth segmentation and diagnosis classification stages were trained five times with different cross- validation splits to investigate the variability of the method's results.

The performance of the tooth segmentation stage was evaluated using a free-response receiver-operating-characteristics (FROC) analysis and mean average precision (mAP) metrics. These analyses and metrics were calculated as the mean performance on the test split that was held out for each round of cross-validation.

The diagnosis classification stage was evaluated based on a ROC analysis for each type of diagnosis. Additional metrics such as accuracy, F1-score, and AUC were also computed and reported. To gain

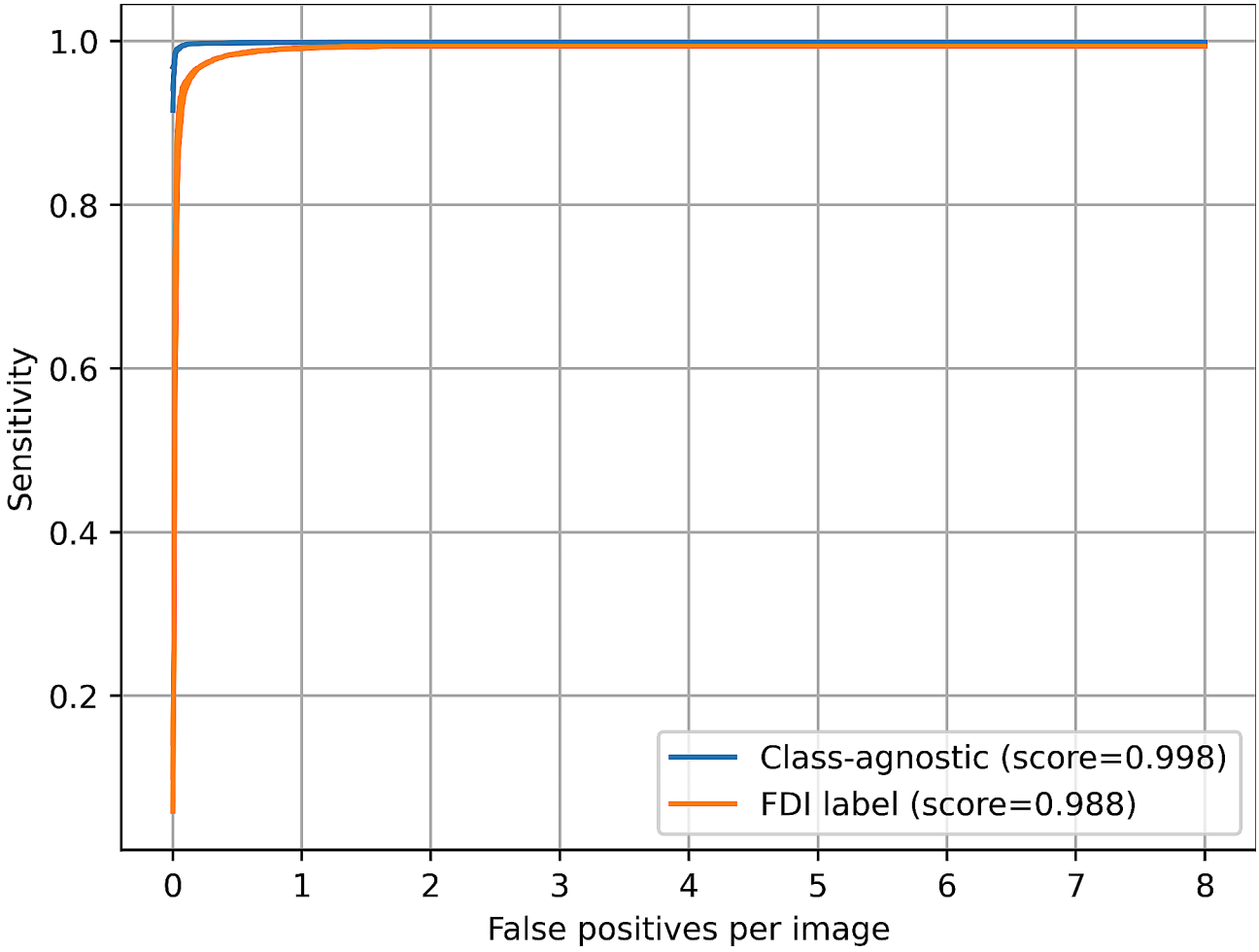

Fig. 2 Free-response receiver operating characteristic (FROC) curves of tooth instance segmentation on PRs. Results are based on 324 held-out test PRs from the OdontoAI platform. The score is computed as the mean sensitivity at false-positive rates of 0.25, 0.5, 1, 2, 4, and 8 following [48]

Table 1 Tooth instance segmentation metrics.The results on the held-out test split of PRs from the OdontoAI platform are averaged for five models trained with 5-fold cross-validation. mAP@IoU=x denotes mean average precision at intersection over union threshold(s) x

mAP@IoU =	0.5	0.75	0.5:0.95
Class-agnostic	0.990	**0.989**	**0.849**
FDI	**0.993**	0.985	0.848

insights into the model's limitations, failure cases were shown and analyzed to assess the method's errors.

Statistical analysis

The model predictions on the test splits were compared to the reference annotations using scikit-learn (v1.3.0). Classification metrics were reported as follows: accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ and F1-score = $\frac{2TP}{2TP+FP+FN}$, where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Furthermore, the area under the receiver-operating-characteristics curve (AUC) and the confusion matrix were presented.

Results

The current method demonstrated high performance in automated tooth segmentation and labeling with a mean FROC score of 0.988 and a mAP of 0.848 (Fig. 2; Table 1).

When the FDI numbers were excluded from consideration, the performance increased further, with a mean FROC score of 0.998 and a mAP of 0.849. Upon visual examination, the tooth segmentations were accurate, even in cases of overlapping teeth (Fig. 1). Errors that were present could be attributed to poor image quality or uncommon dental anomalies, such as horizontally impacted canines or the presence of a syndromic disease (Fig. 3).

The current method also achieved moderate to high accuracies in classifying dental pathology (Fig. 4). The method was most effective in classifying impacted teeth (AUC=0.996), followed by teeth with a periapical lesion (AUC=0.974) and/or a deep caries lesion (AUC=0.960), and the model was least effective in detection of teeth with an early caries lesion (AUC=0.881).

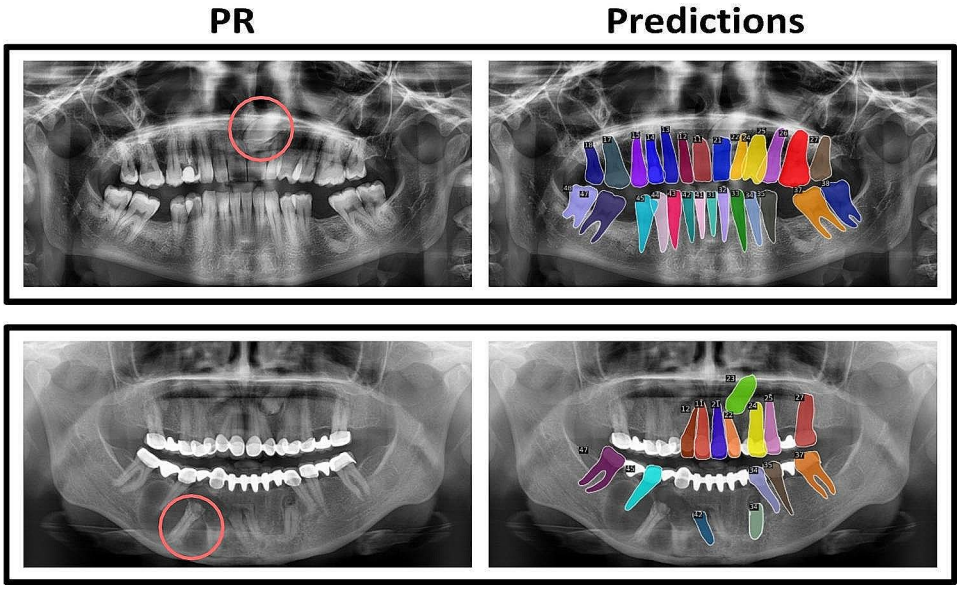

Fig. 3 Tooth segmentation failure cases. Two PRs from the DENTEX challenge dataset are shown with predictions. A horizontally impacted canine (in red circle) is not segmented in the first PR and multiple teeth (e.g. in red circle) are not segmented in the second PR due to a syndromic disease. The label is the tooth's FDI number

The present two-stage AI model had difficulties in distinguishing between teeth with early caries and non-diagnosed teeth, as shown in Fig. 5. In cases where the reference or the predictions indicated a diagnosis (excluding the true negatives, see Table 2), classifying impacted teeth yielded the highest effectiveness (F1=0.901), followed by deep caries (F1=0.683), early caries (F1=0.662), and periapical lesions (F1=0.603).

Discussion

This study aimed to explore the advantages and challenges of employing publicly available research data for AI-based dental research. Based on two public datasets, OdontoAI and DENTEX, a two-stage AI model was developed for automated tooth segmentation and diagnosis classification in PRs, to detect impacted teeth as well as teeth with (deep) caries and/or periapical lesions. Using the Mask DINO model (vision transformer), segmentation and labeling of teeth was accurate, obtaining a mAP of 0.848. The AI model demonstrated high effectiveness in the multi-label diagnosis classification of impacted teeth (F1-score=0.901; AUC=0.996). However, it showed limitations in detecting teeth with early caries (F1-score=0.662; AUC=0.881).

The present study highlighted the potential benefits of utilizing public research datasets to develop AI-based approaches to perform specific tasks, to aid clinicians in daily practice. By combining two distinct datasets, data collection and data annotation time could be considerably reduced, yet time was required for adequate data selection. Furthermore, the two-stage AI model was trained to use reference annotations from each dataset effectively. This demonstrated that using and combining public research datasets is a viable way to develop innovative dental AI solutions.

Many studies have been performed investigating tooth segmentation in PRs [43]. A recent study annotated 6,046 PRs and developed a two-stage model that first segmented and cropped around a region of interest, whereafter tooth segmentations were predicted and labeled with FDI numbers [44]. The authors reported an mAP of 0.966 at an intersection over union (IoU) threshold of 0.75 on the validation set, which was less effective compared to our method. Another study used a dataset comprising 1,500 PRs and incorporated individual models for tooth segmentation and tooth labeling using collaborative learning [7]. The results showed an mAP of 0.973 at an IoU threshold of 0.5, which was less effective compared to our model. The high effectiveness of the current method could be partly explained by the curation of the dataset from the OdontoAI platform; dental implants and bridges were not annotated and were not included for model evaluation [19].

One study investigated the efficacy of automatic software for classifying dental conditions such as restorations, caries, and periapical lesions [45]. This software achieved F1-scores of 0.593 and 0.479 for classifying teeth with caries and periapical lesions, respectively. In contrast, the present study achieved higher F1-scores of 0.662 and 0.603 for these conditions. Another study used a two-stage approach similar to the present

study, using manually segmented and cropped third molars in PRs to determine the presence of caries lesions [46]. The study reported an F1-score of 0.86 and an AUC of 0.90, showing better results compared to the current study. However, this AI method has limited utility compared to the current method, as it requires clinician interactions and only assesses caries in third molars. The highlighted studies made use of datasets without public access and the source code of their methods was unavailable, making a thorough reproduction unfeasible. Direct comparisons between the results of the current method and the highlighted studies should thus be made with caution.

Several studies have focused on predicting the segmentation of caries and periapical lesions on PRs. CariesNet was trained on 1159 PRs to predict a segmentation of shallow, moderate, and deep caries lesions and achieved a Dice similarity coefficient (DSC) of 0.935 [12]. A second study used a U-net model to segment periapical lesions in 470 PRs and detected the lesions with an F1-score of 0.88 at an IoU threshold of 0.70 [47]. These studies suggest that a direct segmentation of caries and periapical lesions provides a more precise reference annotation that can be used to develop more effective AI methods. However, acquiring reference segmentations is considerably more laborious, time-consuming, and resource-intensive than identifying or labeling tooth diagnoses. Achieving a consensus among dental experts on the exact boundaries and nature of the segmented lesions poses an additional challenge. This makes it more difficult to establish a unanimous reference for training AI models.

A limitation of this study is the inconsistency of the dataset with tooth diagnosis annotations, as shown in Figs. 6 and 7. As the tooth bounding box can be notably larger than the assessed tooth, some reference teeth with diagnoses could not be matched to a predicted tooth during the construction of classification images (subsubsection 3.2.2).

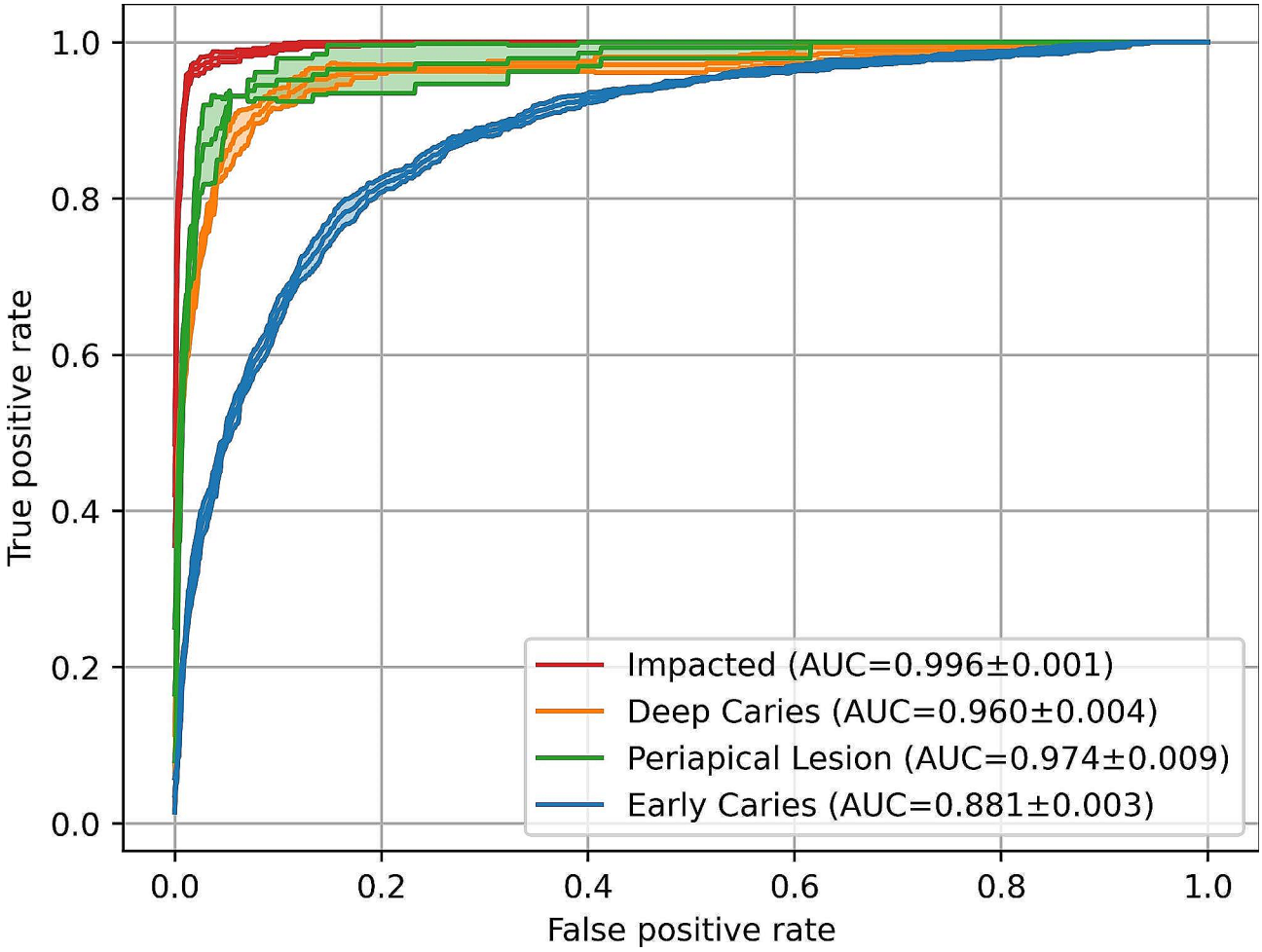

Fig. 4 Receiver operating characteristic (ROC) curves illustrating the multi-label classification results of tooth diagnoses on PRs. A varying effectiveness can be observed for different tooth diagnoses. Results are based on 134 held-out test PRs from the DENTEX challenge. AUC = area under ROC curve

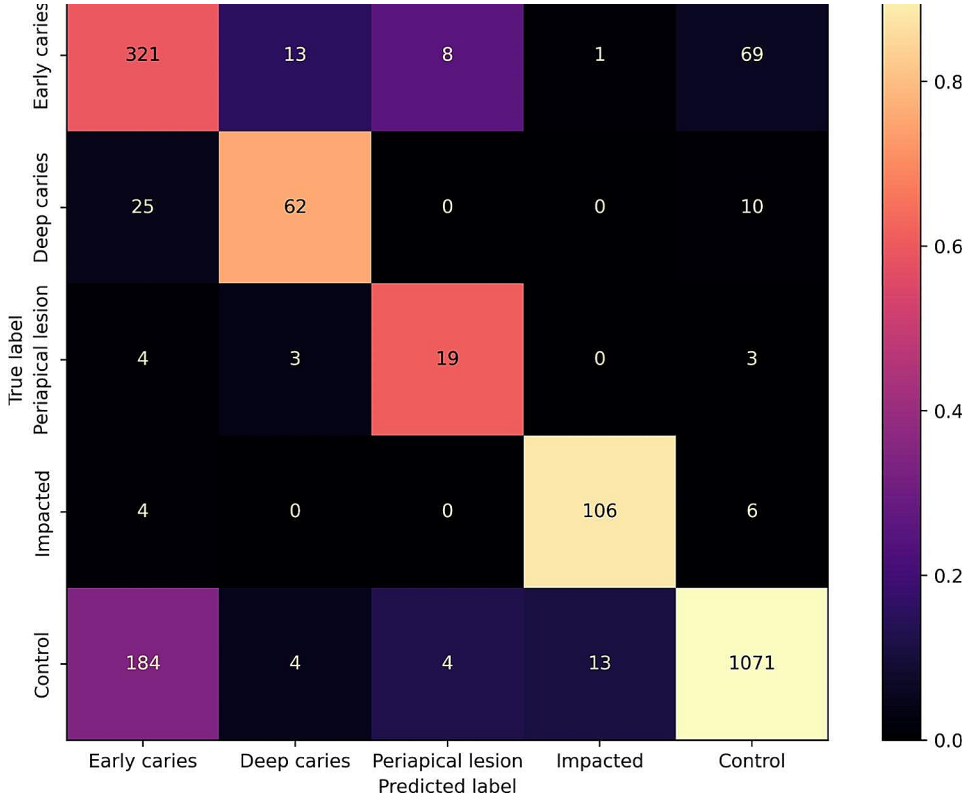

Fig. 5 Confusion matrix illustrating the multi-label classification results of tooth diagnoses on PRs. Results on 134 held-out test PRs from the DENTEX challenge are shown for the most effective model. The colorbar is normalized according to the number of PRs per predicted label

Table 2 Multi-label tooth diagnosis classification metrics. The results on the held-out test split are averaged for five models trained with 5-fold cross-validation. See subsection 3.4 for elaboration on the metrics

	Accuracy	F1-score	AUC
Early Caries	0.843	0.662	0.881
Deep Caries	0.969	0.683	0.960
Periapical Lesion	0.987	0.603	0.974
Impacted	0.988	0.901	0.996

Moreover, the dataset contained missing diagnoses (false negatives) and misdiagnoses (false positives). As a result, a re-assessment of the dataset could be performed based on the largest discrepancies between the reference annotations and the AI method's predictions to improve the consistency of the dataset and the model's effectiveness. Biases present in the dataset and model could only be identified and corrected with a diagnostic re-assessment of (a sample of) the dataset involving a dental expert. Another limitation is the long processing time of the current two-stage model (40 s), compared to single-stage object detectors or segmentation models. This could potentially hinder the adoption of the AI method, as a clinician expects immediate results upon acquiring a PR.

The current work can be extended by incorporating additional public research data with segmentations of caries and periapical lesions [13, 20]. Using the current method for tooth segmentation, it is possible to integrate lesion segmentations to form multi-level tooth segmentations, with the tooth segmentation as the first level and the associated lesion segmentations as the second level. However, performing a diagnostic re-assessment to verify and validate these datasets before using them for further research is recommended. Another direction for future work is collecting PRs and bitewings or periapical radiographs from the same patient visit. Clinicians make fewer diagnostic errors when detecting caries and periapical lesions on these higher-resolution radiographs. Subsequently, a dataset of PRs with more reliable annotations of tooth diagnoses can be created by making the diagnoses on the associated higher-resolution radiographs.

Conclusions

This study aimed to investigate the opportunities and challenges of using publicly available datasets in dental AI research. For that purpose, two public datasets with panoramic radiographs were combined to develop an effective method for predicting tooth segmentations, FDI numbers, and tooth diagnoses, concurrently.

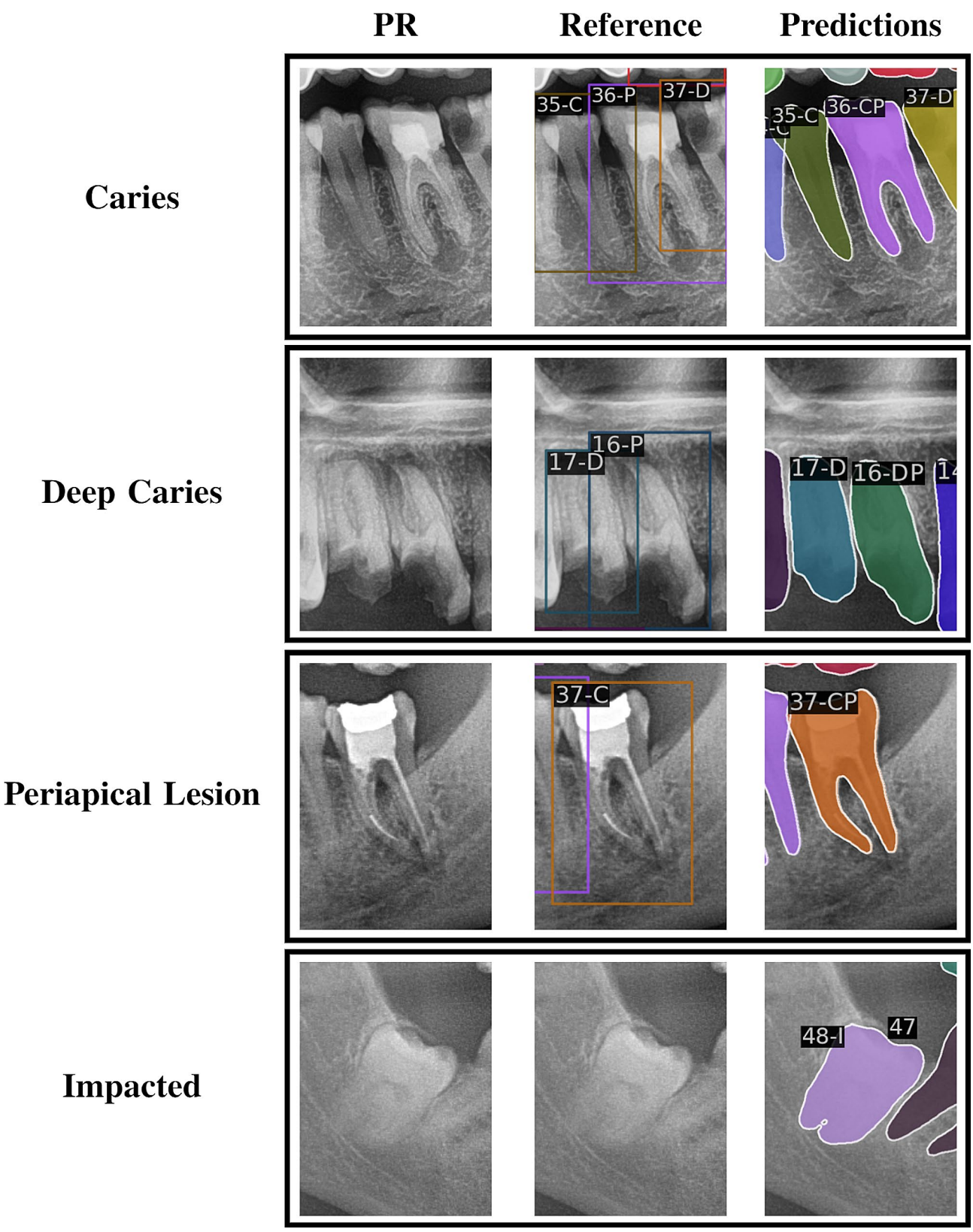

Fig. 6 DENTEX annotations with missing diagnoses. A missed diagnosis (false negative) is shown for each diagnosis, which are selected based on the maximum difference between a diagnosis probability and whether that diagnosis is annotated. Note that the DENTEX dataset only annotates diagnosed teeth, whereas the current method predicts all teeth. The label is the tooth's FDI number with C = early caries, D = deep caries, P = periapical lesion, I = impacted

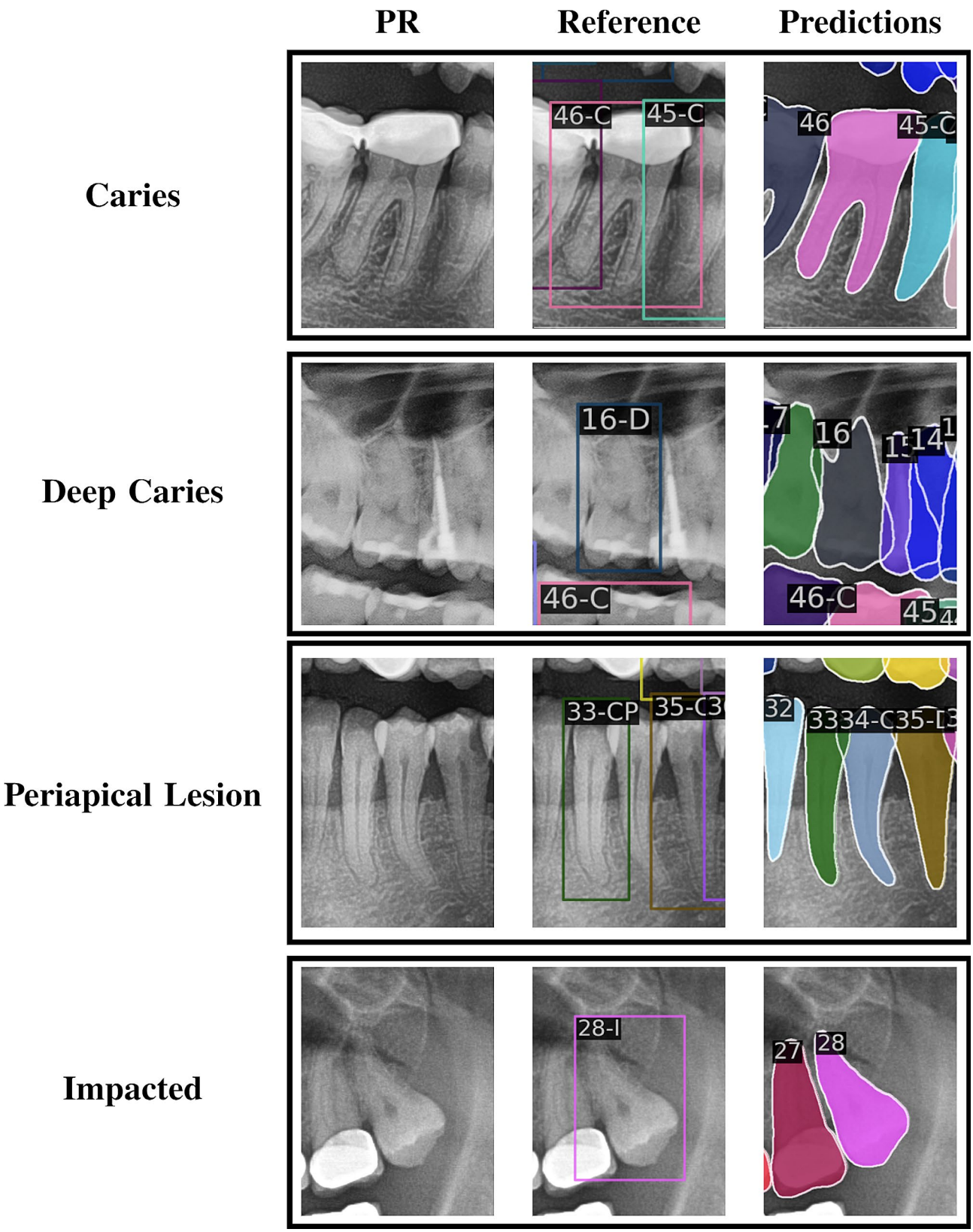

Fig. 7 Misdiagnosed DENTEX annotations. A misdiagnosis (false positive) is shown for each diagnosis, which are selected based on the maximum difference between a diagnosis probability and whether that diagnosis is annotated. Note that the DENTEX dataset only annotates diagnosed teeth, whereas the current method predicts all teeth. The label is the tooth's FDI number with C = early caries, D = deep caries, P = periapical lesion, I = impacted

Using and combining public datasets for AI research in dentistry enables fast exploration of novel tasks and considerably reduces the development time of AI methods. However, the quality of the reference annotations can vary greatly depending on the dataset. Implementing a form of data quality assurance is therefore recommended to optimize the performance of AI models while limiting the risk of biases.

Abbreviations

AI	Artificial Intelligence
AUC	area under receiver operating characteristic curve CNNs = convolutional neural networks
COCO	Common Objects in Context CSRA = class-specific residual attention
DENTEX	Dental Enumeration and Diagnosis on Panoramic X-rays
DINO	DIstillation of knowledge with NO labels
DSC	Dice similarity coefficient
FDI	Fédération Dentaire Internationale
FROC	free-response receiver operating characteristic IoU = intersection over union
mAP	mean average precision
PRs	Panoramic radiographs
ROC	receiver operating characteristic
SimMIM	simple masked image modeling

Acknowledgements
None.

Author contributions
NvN: Method, Investigation, Formal Analysis, Software, Writing - original draft, Visualization. KEG: Software, Validation, Data Curation, Writing - original draft. Tong Xi: Validation, Writing - review & editing. MC: Writing - review & editing. AS: Method, Supervision. SK: Method, Software. TF: Resources, Funding acquisition. BvG: Resources, Project administration. BL: Writing - review & editing. SV: Conceptualization, Investigation, Method, Formal Analysis, Software, Writing – review & editing.

Funding
This research is partially funded by Berlin Institute of Health and Radboud AI for Health.
Open Access funding enabled and organized by Projekt DEAL.

Data availability
The DENTEX challenge dataset used during the current study is available in the Zenodo repository, https://zenodo.org/record/7812323#.ZDQE1uxBwUG, and the dataset from the OdontoAI platform can be downloaded from the platform's website, https://odontoai.com/dataset/.

Declarations

Ethics approval and consent to participate
Ethical approval for this study is waived and no informed consent was required as all image data were publicly available and were anonymized.

Consent for publication
No informed consent was required as all image data were publicly available and were anonymized.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Oral and Maxillofacial Surgery, Radboud University Nijmegen Medical Centre, Postal Number 590, P.O. Box 9101, Nijmegen 6500 HB, The Netherlands
[2]Department of Radiology, Radboud University Medical Center, Geert Grooteplein Zuid 10, Nijmegen 6525 GA, The Netherlands
[3]Department of Oral and Maxillofacial Surgery, Charité - Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Hindenburgdamm 30, 12203 Berlin, Germany
[4]Department of Oral and Maxillofacial Surgery, Erasmus MC, Dr. Molewaterplein 40, Rotterdam, The Netherlands
[5]Einstein Center for Digital Future, Wilhelmstraße 67, Berlin, Germany
[6]Department of Dentistry, Radboud University Medical Center, Ph. Van Leydenlaan 25, Nijmegen 6525 EX, The Netherlands

Received: 30 September 2023 / Accepted: 11 March 2024
Published online: 26 March 2024

References

1. FelsyPremila G, et al. Visual interpretation of panoramic radiographs in dental students using eye-tracking technology. J Dent Educ. Mar. 2022;86. https://doi.org/10.1002/jdd.12899.
2. Peretz B, Gotler M, Kaffe I. Common Errors in Digital Panoramic Radiographs of Patients with Mixed Dentition and Patients with Permanent Dentition. In: *International journal of dentistry* 2012Feb. (2012), p. 584138. https://doi.org/10.1155/2012/584138.
3. Susanne Perschbacher. Interpretation of panoramic radiographs. In: *Australian dental journal* 57 Suppl 1 (Mar. 2012), pp. 40–5. https://doi.org/10.1111/j.1834-7819.2011.01655.x.
4. Lei C, et al. Expert consensus on dental caries management. Int J Oral Sci. Dec. 2022;14. https://doi.org/10.1038/s41368-022-00167-3.
5. Akarslan ZZ et al. A comparison of the diagnostic accuracy of bitewing, periapical, unfiltered and filtered digital panoramic images for approximal caries detection in posterior teeth. In: *Dentomaxillofacial Radiology* 37.8 (2008). PMID: 19033431, pp. 458–463. 10.1259 / dmfr/84698143.
6. Junhua, Zhu, et al. Artificial intelligence in the diagnosis of dental diseases on panoramic radiographs: a preliminary study. BMC Oral Health. June 2023;23. 10.1186/ s12903-023-03027-6.
7. Geetha Chandrashekar E, Bumann, Lee Y. Collaborative deep learning model for tooth segmentation and identification using panoramic radiographs. Comput Biol Med. July 2022;148:105829. https://doi.org/10.1016/j.compbiomed.2022.105829.
8. Gil J et al. Oct. Deep Instance Segmentation of Teeth in Panoramic X-Ray Images. In: 2018, pp. 400–407. https://doi.org/10.1109/SIBGRAPI.2018.00058.
9. Rafic N et al. Aug. Automatic teeth segmentation on panoramic X-rays using deep neural networks. In: 2022, pp. 4299–4305. https://doi.org/10.1109/ICPR56361.2022.9956708.
10. Burak, Dayı, et al. A Novel Deep Learning-Based Approach for Segmentation of different type caries lesions on panoramic radiographs. Diagnostics. Jan. 2023;13:202. https://doi.org/10.3390/diagnostics13020202.
11. Arman H, et al. PaXNet: tooth segmentation and dental caries detection in panoramic X-ray using ensemble transfer learning and capsule classifier. Multimedia Tools Appl. Feb. 2023;82:1–21. https://doi.org/10.1007/s11042-023-14435-9.
12. Haihua, Zhu, et al. CariesNet: a deep learning approach for segmentation of multi-stage caries lesion from oral panoramic X-ray image. Neural Comput Appl. Jan. 2022;35:1–9. https://doi.org/10.1007/s00521-021-06684-2.
13. Yifan, Zhang, et al. Children's dental panoramic radiographs dataset for caries segmentation and dental disease detection. Sci Data. June 2023;10. https://doi.org/10.1038/s41597-023-02237-5.
14. Emel G, et al. Automatic segmentation of Teeth, Crown-Bridge restorations, Dental implants, restorative fillings, Dental Caries, residual roots, and Root Canal fillings on Orthopantomographs: Convenience and pitfalls. Diagnostics. Apr. 2023;13:1–10. https://doi.org/10.3390/diagnostics13081487.
15. Soroush S et al. Deep Learning for Detection of Periapical Radiolucent Lesions: A Systematic Review and Meta-analysis of Diagnostic Test Accuracy. In: *Journal of Endodontics* 49.3 (2023), 248–261.e3. issn: 0099-2399. https://doi.org/10.1016/j.joen.2022.12.007.
16. Manal H, Hamdan et al. The effect of a deep-learning tool on dentists' performances in detecting apical radiolucencies on periapical radiographs. In: *Dentomaxillofacial Radiology* 51.7 (2022). PMID: 35980437, p. 20220122. https://doi.org/10.1259/dmfr.20220122.
17. Julio César Mello Román. Panoramic Dental radiography image Enhancement using Multiscale Mathematical morphology. Sensors. 2021;21. https://doi.org/10.3390/s21093110.

18. Abdi AH, Kasaei DDSS, Mehdizadeh M. Automatic segmentation of mandible in panoramic x-ray. J Med Imaging. 2015;2:044003. https://doi.org/10.1117/1.JMI.2.4.044003.
19. Bernardo Peters Menezes Silva. Boosting research on dental panoramic radiographs: a challenging data set, baselines, and a task central online platform for benchmark. Computer Methods Biomech Biomedical Engineering: Imaging Visualization. 2023;11:1327–47. https://doi.org/10.1080/21681163.2022.2157747.
20. Karen P, et al. Tufts Dental Database: a Multimodal Panoramic X-Ray dataset for Benchmarking Diagnostic systems. IEEE J Biomedical Health Informatics PP (Oct. 2021;1–1. https://doi.org/10.1109/JBHI.2021.3117575.
21. Ibrahim Ethem, Hamamci et al. *DENTEX: An Abnormal Tooth Detection with Dental Enumeration and Diagnosis Benchmark for Panoramic X-rays*. 2023. arXiv: 2305.19112 [cs.CV].
22. Falk S et al. Artificial intelligence in dental research: Checklist for authors, reviewers, readers. In: *Journal of Dentistry* 107 (2021), p. 103610. issn: 0300–5712. doi: 10. 1016/j.jdent.2021.103610.
23. Piotr Szymański and Tomasz Kajdanowicz. A scikit-based Python environment for performing multi-label classification. Feb. 2017. arXiv: 1702.01460 [cs.LG].
24. Sezgin Er A. 2023. https://doi.org/10.5281/zenodo.7812323.
25. Feng, Li et al. *Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation*. 2022. arXiv: 2206.02777 [cs.CV].
26. Kai, Chen et al. *MMDetection: Open MMLab Detection Toolbox and Benchmark*. 2019. arXiv: 1906.07155 [cs.CV].
27. Adam, Paszke et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. url: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.
28. Kaiming, He. Deep Residual Learning for Image Recognition. In: 2016 *IEEE Conference on Computer Vision and Pattern* Recognition *(CVPR)*. 2016, pp. 770–778. 10.1109/ CVPR.2016.90.
29. Tsung-Yi, Lin et al. *Microsoft COCO: Common Objects in Context*. 2015. arXiv: 1405.0312 [cs.CV].
30. Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. 2019. arXiv: 1711. 05101 [cs.LG].
31. Golnaz G et al. *Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation*. 2021. arXiv: 2012.07177 [cs.CV].
32. Xinlong, Wang et al. SOLOv2: Dynamic and Fast Instance Segmentation. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle Vol. 33. Curran Associates, Inc., 2020, pp. 17721–17732. url: https://proceedings.neurips.cc/paper_files/paper/2020/file/cd3afef9b8b89558cd56638c3631868a-Paper.pdf.
33. MMPreTrain C. *OpenMMLab's Pre-training Toolbox and Benchmark*. https://github.com/open-mmlab/mmpretrain. 2023.
34. Ze, Liu et al. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. 2021. arXiv: 2103.14030 [cs.CV].
35. Jia D et al. Imagenet: A large-scale hierarchical image database. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255. 10.1109/ CVPR.2009.5206848.
36. Zhenda, Xie et al. *SimMIM: A Simple Framework for Masked Image Modeling*. 2022. arXiv: 2111.09886 [cs.CV].
37. Fabian I, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. Feb. 2021;18:1–9. https://doi.org/10.1038/s41592-020-01008-z.
38. Christian S et al. Rethinking the Inception Architecture for Computer Vision. In: 2016 *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Soci- ety, 2016, pp. 2818–2826. https://doi.org/10.1109/CVPR.2016.308.
39. Li Shen Z, Lin, Huang Q. Relay backpropagation for effective learning of deep convolutional neural networks. In: *Computer Vision–ECCV* 2016: *14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*. Springer. 2016, pp. 467–482. https://doi.org/10.1007/978-3-319-46478-7_29.
40. Ke Zhu and Jianxin Wu. Residual Attention: A Simple but Effective Method for Multi-Label Recognition. 2021. arXiv: 2108.02456 [cs.CV].
41. Agrim Gupta P, Dollár, Girshick R. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 5351–5359. https://doi.org/10.1109/CVPR.2019.00550.
42. Tsung-Yi, Lin et al. *Focal Loss for Dense Object Detection*. 2018. arXiv: 1708.02002 [cs.CV].
43. Fahad Umer S, Habib, Adnan N. Application of deep learning in teeth identification tasks on panoramic radiographs. Dentomaxillofacial Radiol. 2022;51:20210504. https://doi.org/10.1259/dmfr.20210504. PMID: 35143260.
44. Mingming, Xu et al. Robust automated teeth identification from dental radiographs using deep learning. In: *Journal of Dentistry* 136 (2023), p. 104607. issn: 0300–5712. doi: 10. 1016/j.jdent.2023.104607.
45. Lukasz Zadrozńy et al. Artificial Intelligence Application in Assessment of Panoramic Radiographs. In: *Diagnostics* 12.1 (2022). issn: 2075–4418. https://doi.org/10.3390/diagnostics12010224.
46. Shankeeth V, et al. Classification of caries in third molars on panoramic radiographs using deep learning. Sci Rep. June 2021;11. https://doi.org/10.1038/s41598-021-92121-2.
47. Ibrahim B et al. A U-Net Approach to Apical Lesion Segmentation on Panoramic Radiographs. In: *BioMed Research International* 2022 (Jan. 2022), pp. 1–7. 10.1155/ 2022/7035367.
48. Babak EB, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in women with breast Cancer. (Dec. 2017. https://doi.org/10.1001/jama.2017.14585.

Publisher's Note