Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Development and external validation of an AI pipeline for fully automated DMFT determination from panoramic radiographs

Babak Saravi

babak.saravi@med.uni-duesseldorf.de

University Hospital Düsseldorf, Heinrich-Heine-University Düsseldorf

Veronika Shavlokhova

University Clinic Ruppin-Brandenburg

Ashraf Türky

University Hospital Düsseldorf, Heinrich-Heine-University Düsseldorf

Daman Deep Singh

University Hospital Düsseldorf, Heinrich-Heine-University Düsseldorf

Felix Schrader

University Hospital Düsseldorf, Heinrich-Heine-University Düsseldorf

Andreas Vollmer

University Hospital of Würzburg

Norbert Kübler

University Hospital Düsseldorf, Heinrich-Heine-University Düsseldorf

Christoph Sproll

University Hospital Düsseldorf, Heinrich-Heine-University Düsseldorf

Lara Schorn

University Hospital Düsseldorf, Heinrich-Heine-University Düsseldorf

Research Article

Keywords: artificial intelligence, deep learning, panoramic radiography, dental caries, DMFT index, external validation

Posted Date: June 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10083595/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Clinical Oral Investigations on October 15th, 2026. See the published version at https://doi.org/10.1007/s00784-026-07214-x.

Abstract

Objectives: To develop and externally validate a deep-learning pipeline for fully automated Decayed-Missing-Filled-Teeth (DMFT) determination from a single panoramic radiograph, benchmarked against trained dentists.

Materials and Methods: A single-stage detector trained on 705 annotated DENTEX 2023 radiographs was deployed with four-fold test-time augmentation and Weighted-Box-Fusion. It was externally validated on 100 anonymised panoramic radiographs from an independent university hospital against reference DMFT counts from two dentists, adjudicated by component-wise averaging. Agreement was assessed by intraclass correlation coefficient (ICC), Pearson correlation, mean absolute error (MAE) and Bland–Altman analysis (10,000-resample bootstrap CIs).

Results: On the external cohort, the pipeline reached an ICC (2,1) of 0.90 and a Pearson r of 0.91 against the adjudicated DMFT, comparable to the two dentists' inter-rater agreement (ICC 0.88; ICC difference +0.02, 95% CI -0.04 to 0.09). Agreement was high for missing (r = 0.90) and filled (r = 0.91) teeth but weak for decayed teeth (r = 0.33). Bland–Altman bias was +1.25 DMFT units (95% LoA -4.0 to 6.5) and MAE 2.21, comparable to the inter-rater MAE of 2.13; 84% of radiographs fell within ±3 units.

Conclusions: A deep-learning pipeline determines DMFT from a single panoramic radiograph with overall agreement against dentists comparable to their agreement with each other, and generalises to an independent cohort. Agreement is strongest for missing and filled teeth; the decayed component remains weak and DMFT is modestly over-estimated, reflecting the limits of panoramic radiography for caries.

Clinical Relevance: Automated DMFT determination from routinely acquired panoramic radiographs may support standardised dental charting and population-level oral-health surveillance, reducing examiner workload and inter-rater variability.

Introduction

Untreated dental caries in the permanent dentition remains the single most prevalent chronic condition in the Global Burden of Disease 2017 estimates, affecting an estimated 2.3 billion people globally; the corresponding Global Burden of Disease 2021 update confirms that this burden has neither shifted geographically nor declined materially over the intervening years[1, 2]. Oral diseases continue to rank among the most costly and most under-served categories of non-communicable disease, and their distribution remains markedly socio-economically skewed[3]. Within this clinical landscape, the Decayed-Missing-Filled-Teeth (DMFT) index has, since the late 1930s, served as the standard epidemiological summary of an individual's lifetime caries experience and is the metric required for World Health Organization-aligned oral-health surveillance; recent comparative work shows that even when more granular criteria such as ICDAS are available, DMFT remains the dominant operationalisation in adult-population studies because of its computational simplicity and historical comparability[4, 5].

In the routine clinical workflow, the DMFT score is reconstructed from a combination of intra-oral inspection, bitewing imaging, and panoramic radiography (orthopantomograms, OPGs), with the radiograph providing the only single-acquisition view of the full dentition. Manual reconstruction of the index is time-consuming and known to be subject to inter-rater variability, particularly for the discrimination between filled and otherwise restored teeth and for the radiographic resolution of small carious lesions[6]. The combination of digital archiving, the global standardisation of OPG acquisition protocols and the maturation of deep-learning methods has therefore made fully automated radiographic DMFT determination an obvious target for clinical translation[7].

Recent systematic and umbrella reviews of artificial-intelligence applications to dental radiology document rapid progress on several adjacent tasks — caries detection on periapical and panoramic images[8, 9], periodontal bone-loss assessment[10, 11], restoration detection and dental charting[12, 13], and tooth detection and FDI numbering[14] — and identify the most common deep-learning architecture families used in current dental work[15, 16]. With the public release of the DENTEX 2023 dataset, which provides 1,005 hierarchically annotated panoramic radiographs and an open evaluation protocol, the field has acquired its first widely-shared benchmark for hierarchical OPG annotation[17], and recent work has begun to extend AI detection of restorations to more challenging clinical conditions such as caries hidden under fixed prostheses[18].

Despite this progress, external validation of dental-AI systems on institutionally independent cohorts remains comparatively uncommon[13], and model performance is seldom benchmarked against patient-level clinical endpoints with the uncertainty quantification called for by current reporting frameworks for clinical prediction models and medical-imaging AI[19, 20]. To the best of our knowledge, no deep-learning system for the fully automated determination of the DMFT index from panoramic radiographs has yet been published. Here, we develop and externally validate such a pipeline and benchmark its agreement against an adjudicated two-rater reference on an external clinical cohort.

Materials and Methods

Datasets

Training and internal validation data. The system was trained on the publicly available DENTEX 2023 dataset (MICCAI 2023; CC-BY research license)[17]. The dataset contains 1,005 anonymised panoramic radiographs collected from three European hospitals; 705 of these carry bounding-box annotations for four pathological categories (caries, deep caries, periapical lesion, impacted tooth). Because DMFT scoring requires tooth-status rather than pathology labels, we re-coded the annotations into three target classes (Healthy, Caries, Filled): the two caries grades were merged into a single Caries class, and a tooth-level restoration status was added by integrating an external list of filled teeth — compiled by a study dentist and mapped to each tooth through its FDI position — so that restored teeth were labelled Filled and the remaining sound teeth Healthy. Bounding boxes for all teeth, including sound (Healthy) teeth, derive from the per-tooth enumeration layer of the DENTEX annotations rather than from the

pathology layer alone. To avoid the data leakage that arises when the validation images duplicate the training images, we created a disjoint, image-wise 80/20 split of the 705 annotated radiographs (564 training, 141 held-out internal validation; random seed 42).

External validation cohort. From a series of 269 consecutively acquired OPGs extracted in an earlier project at the Department of Oral, Maxillofacial and Facial Plastic Surgery of University Hospital Düsseldorf, 100 anonymised radiographs were drawn at random for external validation. The selected radiographs were re-identified with blinded codes D001 – D100 in randomised order prior to annotation.

Reference standard for the external cohort

Two dentists annotated the 100 blinded OPGs independently using the WHO DMFT-only convention: per OPG, each rater recorded the integer counts of decayed (D), missing (M) and filled (F) permanent teeth in a structured spreadsheet, together with an integer self-reported confidence rating (1 = low, 2 = medium, 3 = high). Both raters received the same OPGs in the same randomised order and a written annotation manual specifying the WHO-style decision rules and exclusion criteria. Each rater was blind to the other rater's annotations and to all AI predictions. The adjudicated reference DMFT was computed as the component-wise mean of the two raters, rounded to the nearest integer, with the final DMFT = D + M + F.

Detection pipeline

We used a YOLO26x detector (Ultralytics, Frederick, Maryland, USA), a native end-to-end (non-maximum-suppression-free) single-stage object detector, pretrained on COCO and fine-tuned on the three target classes (Healthy, Caries, Filled)[21]. Training used AdamW with an initial learning rate of 1×10^{-3}, a cosine-annealing schedule, an image size of 1024 pixels, mosaic data augmentation, and an early-stopping patience of 20 epochs on validation mean Average Precision (training stopped after 67 epochs). Training and inference ran on a single NVIDIA GeForce RTX 3090 GPU (NVIDIA Corporation, Santa Clara, California, USA).

At inference time, each OPG was passed through four geometric augmentations — identity, horizontal flip and two scale variants (0.83× and 1.20×)[22, 23]. Per-augmentation predictions were back-projected into the original image coordinate system, yielding four prediction sets per OPG which were then fused with a Weighted-Box-Fusion (WBF) implementation at an IoU threshold of 0.55[24]. Per-class confidence thresholds for the count read-out were fixed on the internal validation split by minimising the per-image count error (0.30 for Healthy and Caries, 0.45 for Filled) and then applied unchanged to the external cohort. For each OPG, the surviving boxes yielded the per-OPG counts of decayed (D), filled (F) and present teeth (D + F + Healthy). Missing teeth (M) were inferred as M = max(0, 28 − number of present teeth), using a 28-tooth convention that excludes the third molars, which on panoramic radiographs are frequently unerupted, impacted, or truncated at the image margin and cannot be reliably localised. The same 28-tooth convention was applied to the adjudicated reference standard and to all reported DMFT analyses. The DMFT index per OPG was the sum D + M + F.

Evaluation metrics and statistical analysis

Internal performance on the held-out validation split was characterised by per-class Average Precision (AP, IoU 0.50:0.95), AP at IoU 0.50 (AP50), AP at IoU 0.75 (AP75), precision and recall, following the COCO conventions used by DENTEX. The contribution of test-time augmentation was assessed both by the per-image count error on the internal split and by the external DMFT agreement obtained with and without TTA.

The primary external endpoint was the agreement between AI-predicted DMFT and the adjudicated rater DMFT per OPG, characterised by Bland–Altman bias and 95% LoA[25], Pearson r, ICC (2,1) and component-wise MAE; the AI was additionally compared with each rater individually, and the difference between the AI-to-reference and the human inter-rater ICC was quantified by a paired 10,000-resample bootstrap. Inter-rater reliability between A and B was characterised by ICC (2,1) and linear-weighted Cohen's κ[26, 27]. The cumulative-agreement curve reported the fraction of OPGs for which $|AI - reference| \leq N$ for each integer DMFT tolerance $N \in \{0, \ldots, 8\}$.

Pre-specified subgroup analyses were performed by DMFT severity (DMFT = 0, 1 – 5, 6 – 12, ≥ 13), by rater A self-reported confidence (low / medium / high), and by image sharpness tercile (Laplacian variance). Per-class detection thresholds were tuned only on the internal validation split. All statistical analyses were performed in Python 3.10 (Python Software Foundation, Wilmington, Delaware, USA) with a custom bootstrap-confidence-interval module (10,000 resamples for the primary endpoints, 2,000 resamples for the subgroup MAE). Following established benchmarks, ICC values between 0.75 and 0.90 were interpreted as good and values above 0.90 as excellent[28].

Ethics and reporting standards

Use of the DENTEX 2023 dataset was governed by its CC-BY research license. Use of the 100 Düsseldorf OPGs was approved by the Ethics Committee of the Medical Faculty of University Hospital Düsseldorf (reference number 2024-3049). All radiographs were anonymised prior to transfer and analysis. Reporting follows the CLAIM checklist for medical AI publications (Online Resource 6) and the relevant items of the TRIPOD-AI Statement.

Results

Study cohort

The flow of OPGs through training, internal validation and external validation is shown in Figure 1. The external cohort consisted of n = 100 anonymised OPGs from University Hospital Düsseldorf, with a rater-adjudicated DMFT distribution that spanned the full clinical range (median 4, range 0 – 28).

Internal performance and contribution of TTA

On the held-out internal validation split (n = 141 OPGs), the detector achieved an overall AP of 0.59 (AP50 = 0.81, AP75 = 0.73, precision = 0.76, recall = 0.79; Figure 2). Performance was strongest for the Healthy class (AP 0.71) and weakest for caries (AP 0.43), the most difficult of the three classes. On the internal split, four-fold TTA reduced the summed per-image count error relative to single-pass inference (4.19 vs 4.40) and was therefore specified a priori as the deployed configuration; the single-pass (no-TTA) result is reported below as a sensitivity analysis.

Inter-rater reliability of the external reference standard

The two dentists agreed strongly on the per-OPG DMFT of the 100 external OPGs (Table 1). Their counts correlated at r = 0.911 with an ICC (2,1) of 0.876 and a linear-weighted Cohen's κ of 0.704, placing the reference standard above the conventionally accepted threshold for clinically usable inter-rater reliability (κ ≥ 0.6)[29] and within the "good" ICC band (0.75 – 0.90)[28].

External validation against the adjudicated rater DMFT

Against the adjudicated rater DMFT, the AI pipeline achieved a Pearson correlation of r = 0.91 and an ICC (2,1) of 0.90 (Figure 3; Table 2). This agreement was comparable to the inter-rater agreement of the two dentists (ICC 0.88): the difference between the AI-to-reference and the human inter-rater ICC was +0.02 (95% CI -0.04 to 0.09), i.e. not statistically distinguishable. Agreement was similar for the two individual raters (Rater A: ICC 0.87, MAE 2.69; Rater B: ICC 0.86, MAE 2.60). Bland–Altman analysis showed a systematic positive bias of +1.25 (95% CI 0.73 to 1.77) DMFT units, with 95% LoA of [-3.95, 6.45]; the AI tended to over-estimate DMFT. The mean absolute error across the cohort was 2.21 (95% CI 1.85, 2.60) DMFT units, comparable to the human inter-rater MAE of 2.13 (difference +0.08, 95% CI -0.61 to 0.76).

Cumulative agreement

The cumulative-agreement curve in Figure 4 shows the fraction of OPGs for which the AI prediction falls within a given absolute DMFT tolerance of the adjudicated reference. 42% of OPGs were within ±1 DMFT unit, 68% within ±2, 84% within ±3 and 90% within ±5. The AI and the between-dentist curves were closely comparable across the clinically relevant tolerances (e.g. 84% vs 74% within ±3 DMFT units), with the two dentists agreeing more often only at the strictest zero-tolerance (42% vs 15% exact agreement).

Component-wise performance

Per-component analysis (Online Resource 1 [Fig. S1]; Table 2) showed strong agreement for filled teeth (Pearson r = 0.91, MAE = 1.20) and missing teeth (r = 0.90, MAE = 0.65), but only weak agreement for the decayed component (r = 0.33, MAE = 1.90). The weak caries agreement, together with the systematic over-estimation of DMFT, is the dominant source of error and is consistent with the well-documented difficulty of identifying carious lesions on standard-resolution OPGs and with the more liberal caries labelling of the training data relative to the WHO convention applied by the raters[9]. The AI pipeline performs strongest on the components most relevant for treatment-planning and oral-health surveillance

— filled and missing teeth — and supports human readers on the more difficult caries component rather than replacing them.

Subgroup robustness

Pre-specified subgroup analyses (Online Resource 2 [Fig. S2]) showed that the MAE of the AI was relatively uniform across image-sharpness terciles (1.88, 2.47 and 2.26 DMFT units) and across DMFT severity strata (2.29 for caries-free mouths and 2.04, 2.12 and 2.79 for the mild, moderate and severe strata). The error did not fall in caries-free mouths, reflecting the systematic over-detection of caries, indicating that the AI error is driven by image content rather than by case difficulty as perceived by the human rater.

Qualitative examples

Representative qualitative examples of the pipeline output on the external Düsseldorf cohort are shown in Online Resource 3 [Fig. S3]. Examples are sampled to span the full diagnostic spectrum: a caries-free mouth on which AI and reference agree (D003), a mid-range case (D004), a high-burden case (D066; DMFT 16), and the two largest disagreements — one where the AI under-estimates DMFT (D015) and one where it over-estimates (D069).

Inference time

Mean per-image inference time on a single consumer-grade NVIDIA RTX 3090 GPU was 81 ms for single-pass inference (95th percentile 132 ms); four-fold TTA increases this approximately four-fold to roughly 0.3 s per OPG. Even with TTA, throughput is several OPGs per second on a single GPU, far faster than the typical clinical OPG reading time of 1 – 5 minutes per image.

Discussion

This study externally validates a single-detector deep-learning pipeline that determines the DMFT index from a single panoramic radiograph and reaches an overall agreement with adjudicated dentist DMFT (ICC 2,1 = 0.90; Pearson r = 0.91) that is statistically comparable to the inter-rater agreement of the two dentists themselves on the same external cohort (ICC 2,1 = 0.88; ICC difference + 0.02, 95% CI -0.04 to 0.09). Both the rank agreement and the absolute error (MAE 2.21 vs the inter-rater 2.13; difference + 0.08, 95% CI -0.61 to 0.76) were statistically indistinguishable from the between-dentist agreement. The comparability is nonetheless qualified by a systematic over-estimation of DMFT (Bland–Altman bias + 1.25, 95% CI 0.73 to 1.77 DMFT units). On the cumulative-agreement scale, 84% of OPGs were within ± 3 DMFT units of the adjudicated reference and 90% within ± 5. The pipeline operates at sub-second speed per OPG on consumer-grade GPU hardware.

A methodological observation is that detection-level gains do not translate automatically into clinical-endpoint gains: four-fold test-time augmentation modestly reduced the per-image count error on the internal validation split but did not improve the external DMFT endpoint (MAE 2.21 with vs 2.10 without

TTA). Because the augmentation policy was fixed a priori on the internal split, we report single-pass inference as a sensitivity analysis rather than re-selecting on the external endpoint; given the small difference, a single-pass deployment is equally defensible. This dissociation between detection-level and count-level gains echoes the broader evidence on view-invariant evidence aggregation[8, 9]. It also underscores the importance of validating augmentation strategies on the clinical target rather than on intermediate detection metrics, as emphasised in adjacent medical-imaging applications[22, 23]. Combined with Weighted-Box-Fusion across the augmented predictions[24], this approach provides a robust and computationally cheap alternative to multi-architecture ensembles for dental-AI work.

Two subgroup observations are relevant for deployment. First, the per-OPG mean absolute error of the AI was relatively uniform across image-sharpness terciles (1.88, 2.47 and 2.26 DMFT units), indicating that performance was not strongly associated with image sharpness within this selected cohort. Second, the error was relatively flat across DMFT severity strata (2.0–2.8 DMFT units) and did not fall in caries-free mouths, consistent with a systematic, content-driven over-detection of caries rather than an error concentrated in the most complex cases.

The inter-rater reliability of the two-dentist adjudicated reference standard (ICC 2,1 = 0.876; linear-weighted $\kappa = 0.704$; mean absolute difference 2.13 DMFT units) provides two methodological lessons. First, the WHO DMFT-only annotation convention, in which raters record per-OPG counts of D, M and F rather than per-tooth bounding boxes, produces a clinically usable reference standard at substantially lower annotation effort than per-tooth annotation. Second, the workflow scales to multi-centre studies and could readily be adapted to other patient-level radiographic endpoints. The substantial inter-rater κ (0.704; Landis–Koch band 0.61–0.80) is broadly consistent with the literature on weighted-κ agreement for dental staging tasks on panoramic radiographs[6]. The mean inter-rater difference of 2.13 DMFT units underlines that whole-mouth DMFT scoring from a single panoramic radiograph is intrinsically difficult even for trained dentists, and defines the realistic ceiling against which any automated system should be judged.

Two methodological choices warrant comment in relation to recent comparable work. Hamdan et al. recently reported strong dental-restoration detection on panoramic radiographs using a no-code computer-vision platform and achieved very high per-tooth F^1 (0.98) for restoration presence detection[12]; Ayhan et al. extended automatic detection to the difficult case of caries underneath fixed prostheses[18]; and Mureșanu et al. provided a recent comprehensive overview of the strengths and pitfalls of fully automated dental-condition detection on panoramic radiographs[13]. None of these systems, however, reports a multi-rater external Bland–Altman validation against the clinical DMFT endpoint, which is the primary methodological innovation of the present work. In contrast to these task-specific detection studies, our system targets the patient-level DMFT endpoint directly and reaches an ICC of 0.90 against an externally-adjudicated two-rater consensus rather than against the training-distribution annotation, while transparently reporting a systematic over-estimation of DMFT driven by the decayed component. We position our pipeline as a decision-support and documentation tool rather than

as a substitute for clinical examination, an alignment with the prevailing consensus on AI deployment in dental clinical practice[16, 20].

Several limitations should be acknowledged. First, the external cohort, while institutionally independent, originates from a single tertiary referral centre and consists of OPGs from a maxillofacial-surgery population. A further limitation concerns the decayed component: the pipeline over-estimated DMFT (bias + 1.25 units), driven by caries over-detection that reflects both the intrinsic difficulty of caries assessment on panoramic radiographs and the more liberal caries labelling of the DENTEX training data relative to the WHO convention used by the raters; the system is consequently best suited to missing- and filled-tooth assessment, with caries read-outs requiring clinical confirmation. Broader generalisation will require additional cohorts from general dental practice and from independent national health-care systems. Second, the DMFT-only reference standard does not enable per-FDI position evaluation of the AI on the external cohort; per-tooth annotation in a follow-up external cohort would extend the present analysis to localisation generalisation. Third, missing teeth are inferred from the absence of detection rather than explicitly classified, a strategy that is robust on the present datasets but could fail in severe edentulism. Fourth, the pipeline does not yet emit calibrated probability scores; explicit probabilistic calibration via temperature scaling on the validation split is a natural extension. Fifth, although the analysis includes pre-specified subgroup analyses by DMFT severity, rater confidence and image sharpness, it does not include a formal demographic-fairness audit; the DMFT-only annotation convention does not require demographic metadata, but a follow-up reader study with consented demographic capture would enable such an audit, as recommended by current reporting standards for clinical-AI systems[19, 20].

Three future directions follow from these results. First, a multimodal extension that combines OPG-based DMFT prediction with bitewing imaging will likely narrow the caries-detection gap, since bitewings resolve approximal lesions that the OPG does not. Second, a prospective reader-comparison study, in which the same OPGs are read by independent dentists with and without AI assistance, would quantify the clinical impact on time-to-diagnosis and on inter-reader agreement. Third, the per-OPG inference profile makes the pipeline a candidate engine for population-level retrospective DMFT surveillance on archived OPG collections, complementing existing decennial national oral-health surveys[1].

In summary, a single-detector deep-learning pipeline determines the DMFT index from a single panoramic radiograph with overall agreement against trained dentists that is statistically comparable to the inter-rater agreement of the dentists themselves on the same external cohort. Agreement is strongest for missing and filled teeth, whereas the decayed component remains weak and the pipeline systematically over-estimates DMFT. With these caveats, and pending prospective evaluation, the pipeline may serve as a decision-support and documentation aid — particularly for restorative and missing-tooth status — and as a scalable engine for population-level oral-health surveillance, rather than as a replacement for clinical examination.

Declarations

Author contributions

BS and AT conceived and designed the study, developed the deep learning pipeline, conducted all model training and evaluation, performed statistical analyses, interpreted the results, and wrote the original manuscript draft. DDS contributed equally to study design, assisted with data preprocessing and quality control, performed literature review, and contributed to manuscript writing. AV assisted with model implementation, figure generation, and manuscript editing. NK and VS provided resources and administrative support, contributed to clinical interpretation of imaging findings, and supervised aspects of the project. CS and FS contributed to study conception, provided overall supervision, contributed to data interpretation, and critically revised the manuscript for important intellectual content. LS provided clinical expertise, conceived and designed the study, contributed to interpretation of results, performed critical revision of the manuscript, and supervised aspects of the project. All authors read and approved the final manuscript.

Funding

This work received no specific external funding.

Conflict of interest

The authors declare no conflict of interest.

Ethics approval

All procedures performed in this study involving human data were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki Declaration and its later amendments. The study was approved by the Ethics Committee of the Medical Faculty of University Hospital Düsseldorf (reference number 2024-3049).

Consent to participate

The requirement for individual informed consent was waived by the Ethics Committee of the Medical Faculty of University Hospital Düsseldorf, given the retrospective analysis of fully anonymised imaging data.

Consent to publish

Not applicable. The manuscript does not contain any individual person’s identifiable data.

Data and code availability

The DENTEX 2023 dataset is publicly available from the DENTEX challenge organisers (https://dentex.grand-challenge.org/). The complete training, inference, evaluation and figure-generation source code is openly available at https://github.com/Freiburg-AI-Research/automated-dmft-from-opg

under the MIT License. The 100 anonymised external panoramic radiographs and the two-rater DMFT annotation files contain potentially identifying clinical imaging data and are therefore not publicly deposited; they can be made available from the corresponding author upon reasonable request and after completion of a data-transfer agreement, subject to the approval of the Ethics Committee of the Medical Faculty of University Hospital Düsseldorf.

References

1. GBD 2017 Oral Disorders Collaborators, Marcenes BE (2020) Global, Regional, and National Levels and Trends in Burden of Oral Conditions from 1990 to 2017: A Systematic Analysis for the Global Burden of Disease 2017 Study. J Dent Res 99:362–373. 10.1177/0022034520908533
2. GBD 2021 Oral Disorders Collaborators (2025) Trends in the global, regional, and national burden of oral conditions from 1990 to 2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet. 10.1016/S0140-6736(25)00084-3
3. Peres MA, Macpherson LMD, Weyant RJ (2019) Oral diseases: a global public health challenge. Lancet 394:249–260. 10.1016/S0140-6736(19)31146-8
4. Rai A, Sundas S, Dhakal N (2023) Assessment of dental caries based on ICDAS and WHO criteria: A comparative study. Int J Pediatr Dent. 10.1111/ipd.13083
5. Pitts NB, Zero DT, Marsh PD (2017) Dental caries. Nat Reviews Disease Primers 3:17030. 10.1038/nrdp.2017.30
6. Abdul Rahim AH, Davies JA, Liversidge HM (2023) Reliability and limitations of permanent tooth staging techniques. Forensic Sci Int 346:111654. 10.1016/j.forsciint.2023.111654
7. LeCun Y, Bengio Y, Hinton G (2015) Deep Learn Nat 521:436–444. 10.1038/nature14539
8. Mohammad-Rahimi H, Motamedian SR, Rohban MH (2022) Deep learning for caries detection: A systematic review. J Dent 122:104115. 10.1016/j.jdent.2022.104115
9. Negi S, Mathur A, Tripathy S (2024) Artificial Intelligence in Dental Caries Diagnosis and Detection: An Umbrella Review. Clin Experimental Dent Res 10:e70004. 10.1002/cre2.70004
10. Khubrani YH, Thomas D, Slator PJ (2025) Detection of periodontal bone loss and periodontitis from 2D dental radiographs via machine learning and deep learning: systematic review employing APPRAISE-AI and meta-analysis. Dentomaxillofacial Radiol 54:89–108. 10.1093/dmfr/twae053
11. Revilla-León M, Gómez-Polo M, Barmak AB (2022) Artificial intelligence models for diagnosing gingivitis and periodontal disease: A systematic review. J Prosthet Dent 130:816–824. 10.1016/j.prosdent.2022.01.026
12. Hamdan M, Badr Z, Bjork J (2023) Detection of dental restorations using no-code artificial intelligence. J Dent 139:104768. 10.1016/j.jdent.2023.104768
13. Mureșanu S, Hedeșiu M, Iacob L (2024) Automating Dental Condition Detection on Panoramic Radiographs: Challenges, Pitfalls, and Opportunities. Diagnostics 14:2336. 10.3390/diagnostics14202336

14. Özçelik STA, Üzen H, Şengür A (2024) Enhanced Panoramic Radiograph-Based Tooth Segmentation and Identification Using an Attention Gate-Based Encoder-Decoder Network. Diagnostics 14:2719. 10.3390/diagnostics14232719
15. Katsumata A (2023) Deep learning and artificial intelligence in dental diagnostic imaging. Japanese Dent Sci Rev 59:329–333. 10.1016/j.jdsr.2023.09.004
16. AbuSalim S, Zakaria N, Islam MR (2022) Analysis of Deep Learning Techniques for Dental Informatics: A Systematic Literature Review. Healthcare 10:1892. 10.3390/healthcare10101892
17. Hamamci IE, Er S, Simsar E (2023) DENTEX: An Abnormal Tooth Detection with Dental Enumeration and Diagnosis Benchmark for Panoramic X-rays. arXiv 10.48550/arXiv.2305.19112
18. Ayhan B, Ayan E, Atsü S (2025) Detection of dental caries under fixed dental prostheses by analyzing digital panoramic radiographs with artificial intelligence. BMC Oral Health 25:216. 10.1186/s12903-025-05604-3
19. Collins GS, Moons KGM, Dhiman P (2024) TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 385:e078378. 10.1136/bmj-2023-078378
20. Mongan J, Moy L, Kahn CE (2020) Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. Radiology: Artif Intell 2:e200029. 10.1148/ryai.2020200029
21. Jocher G, Qiu J (2026) Ultralytics YOLO26. Software (Ultralytics)
22. Qasim AB, Motta A, Studier-Fischer A (2024) Test-time augmentation with synthetic data addresses distribution shifts in spectral imaging. Int J Comput Assist Radiol Surg 19:1021–1031. 10.1007/s11548-024-03100-7
23. Ashraf H, Waris A, Ghafoor MF (2022) Melanoma segmentation using deep learning with test-time augmentations and conditional random fields. Sci Rep 12:3948. 10.1038/s41598-022-07885-y
24. Solovyev R, Wang W, Gabruseva T (2021) Weighted boxes fusion: Ensembling boxes from different object detection models. Image Vis Comput 107:104117. 10.1016/j.imavis.2021.104117
25. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1:307–310. 10.1016/S0140-6736(86)90837-8
26. Shrout PE, Fleiss JL (1979) Intraclass correlations: Uses in assessing rater reliability. Psychol Bull 86:420–428. 10.1037/0033-2909.86.2.420
27. Cohen J (1968) Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 70:213–220. 10.1037/h0026256
28. Koo TK, Li MY (2016) A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 15:155–163. 10.1016/j.jcm.2016.02.012
29. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174. 10.2307/2529310

Tables

Table 1. Inter-rater reliability of the two dentists on the external Düsseldorf cohort (n = 100 OPGs). The two raters were blind to each other and to the AI predictions throughout the annotation process.

Statistic	Value
Pearson r (Rater A vs Rater B)	0.911
ICC (2,1)	0.876
Linear-weighted Cohen's κ	0.704
Mean absolute error (DMFT units)	2.13

Table 2. Agreement between AI predictions and adjudicated rater DMFT on the external Düsseldorf cohort (n = 100 OPGs). MAE values are reported with 95% percentile-bootstrap confidence intervals (10,000 resamples). Bland–Altman bias and 95% LoA are reported in DMFT units.

Metric	Value	95% CI
Pearson r (DMFT)	0.91	—
ICC (2,1) (DMFT)	0.90	—
MAE (DMFT units)	2.21	1.85, 2.60
Bland–Altman bias (DMFT units)	1.25	0.73, 1.77
Bland–Altman lower LoA	-3.95	—
Bland–Altman upper LoA	6.45	—
Pearson r — Filled (F)	0.91	—
Pearson r — Missing (M)	0.90	—
Pearson r — Decayed (D)	0.33	—
MAE — Filled (F)	1.20	—
MAE — Missing (M)	0.65	—
MAE — Decayed (D)	1.90	—

Figures

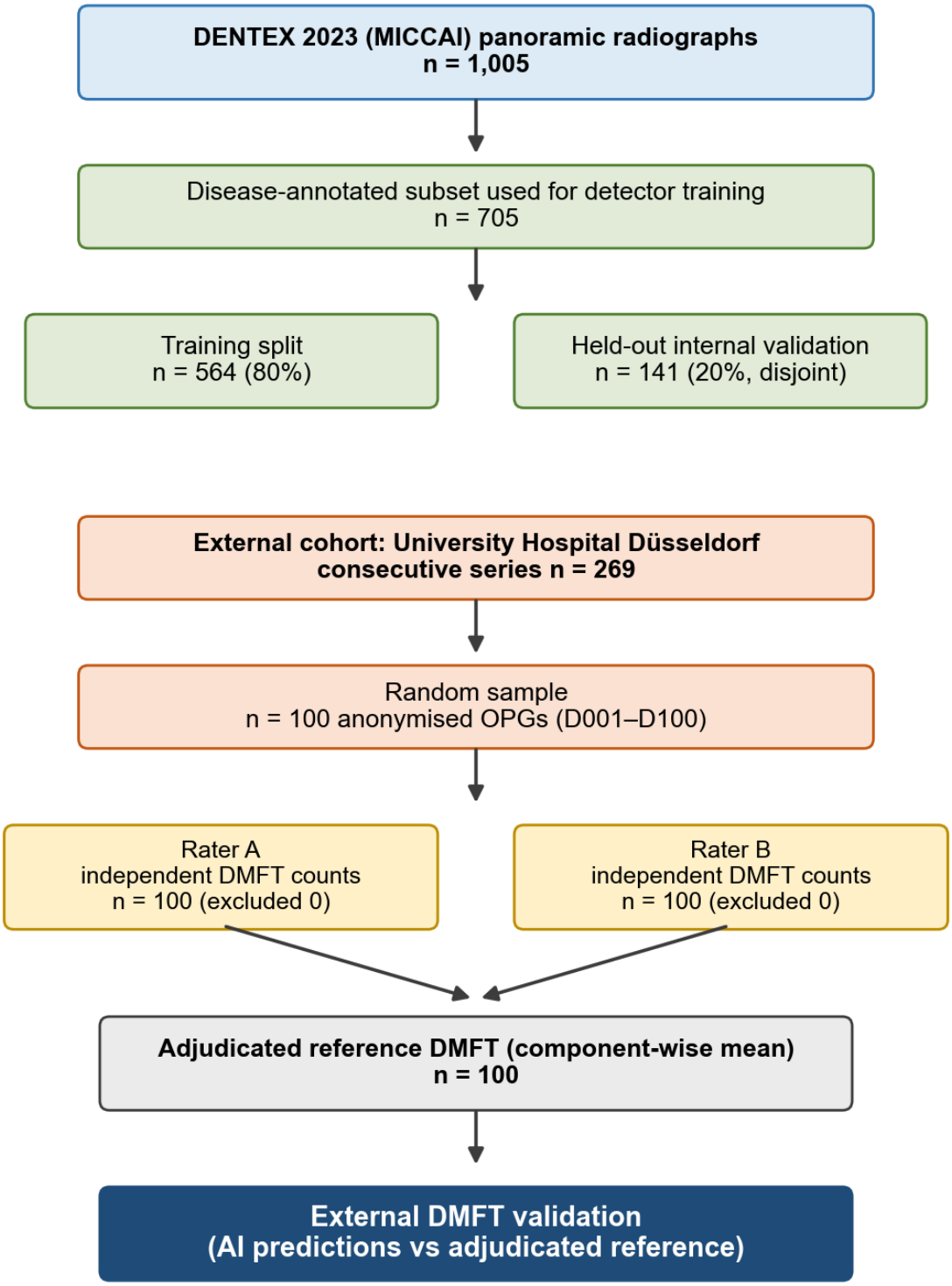


Figure 1

Study flow diagram. The detector was trained on the 705 disease-annotated radiographs of the public DENTEX 2023 dataset (n = 1,005 OPGs), split image-wise into 564 training and 141 held-out internal-validation images. External validation was performed on 100 anonymised OPGs drawn at random from a series of 269 consecutively acquired OPGs at University Hospital Düsseldorf, with reference DMFT counts produced independently by two dentists and adjudicated by component-wise averaging.

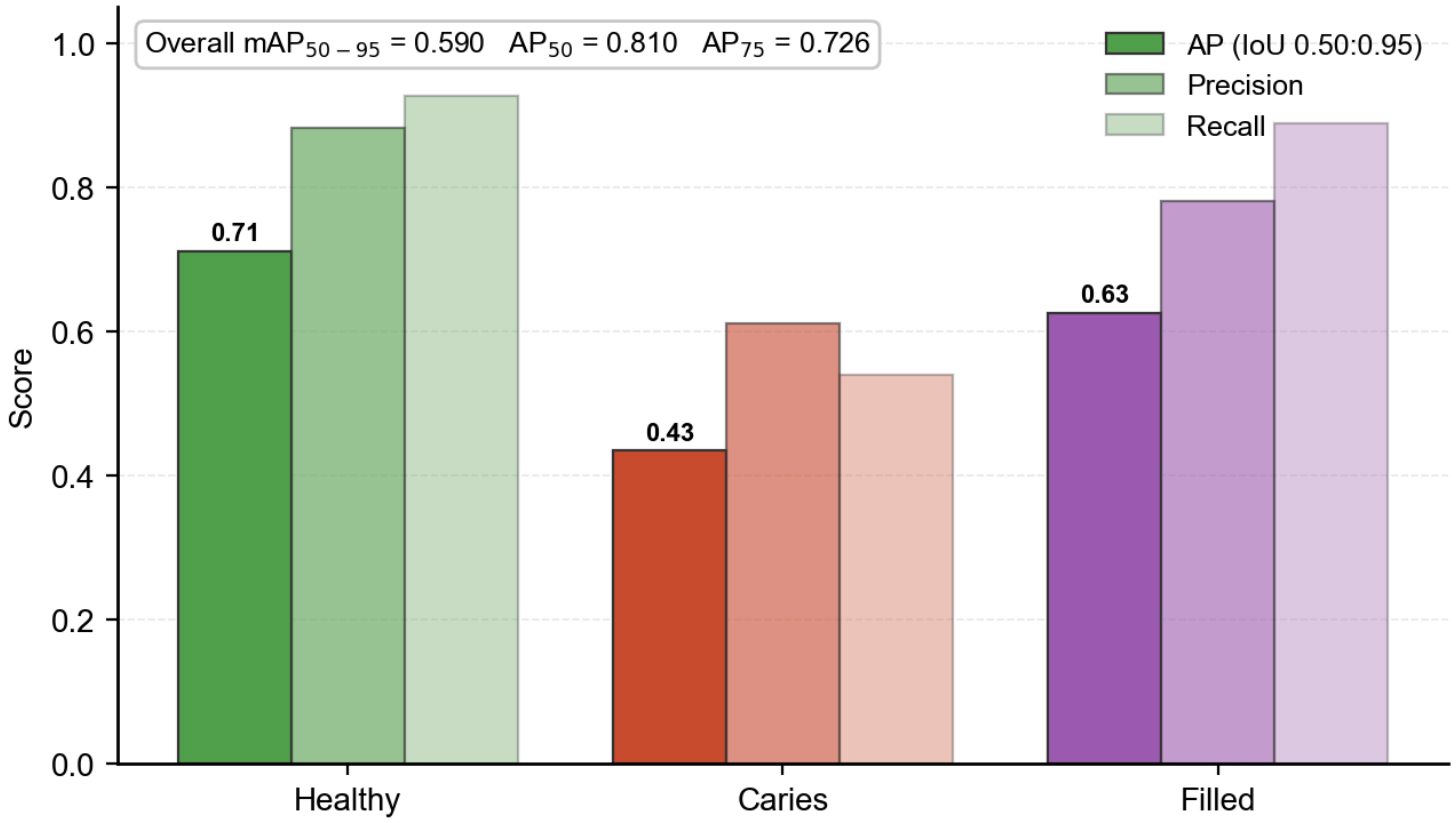


Figure 2

Per-class detection performance on the held-out internal validation set (n = 141): Average Precision (IoU 0.50:0.95), precision and recall for the Healthy, Caries and Filled classes. Caries is the most difficult class (AP 0.43).

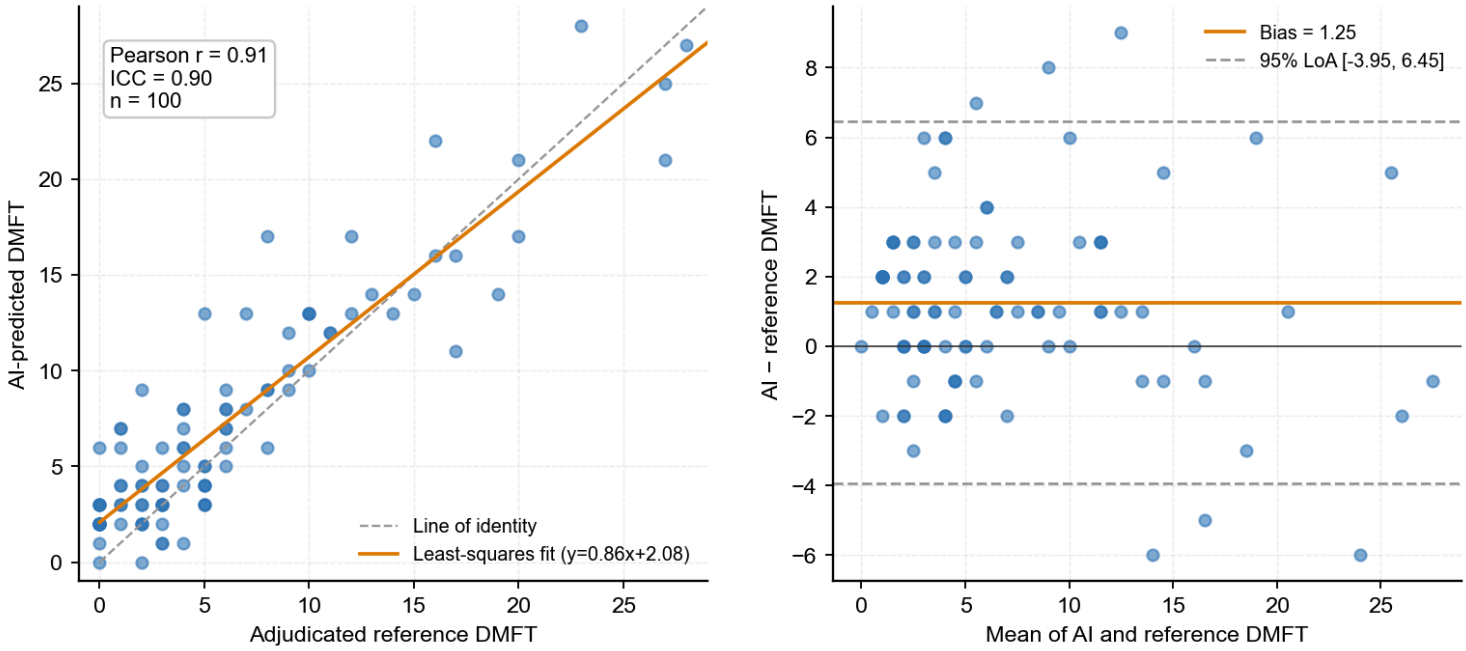


Figure 3

AI vs adjudicated rater DMFT on the external Düsseldorf cohort (n = 100). Left panel: scatter plot of AI-predicted against adjudicated rater DMFT, with line of identity (grey dashed) and least-squares fit (orange). Right panel: Bland–Altman plot with mean bias (solid line) and 95% limits of agreement (dashed lines).

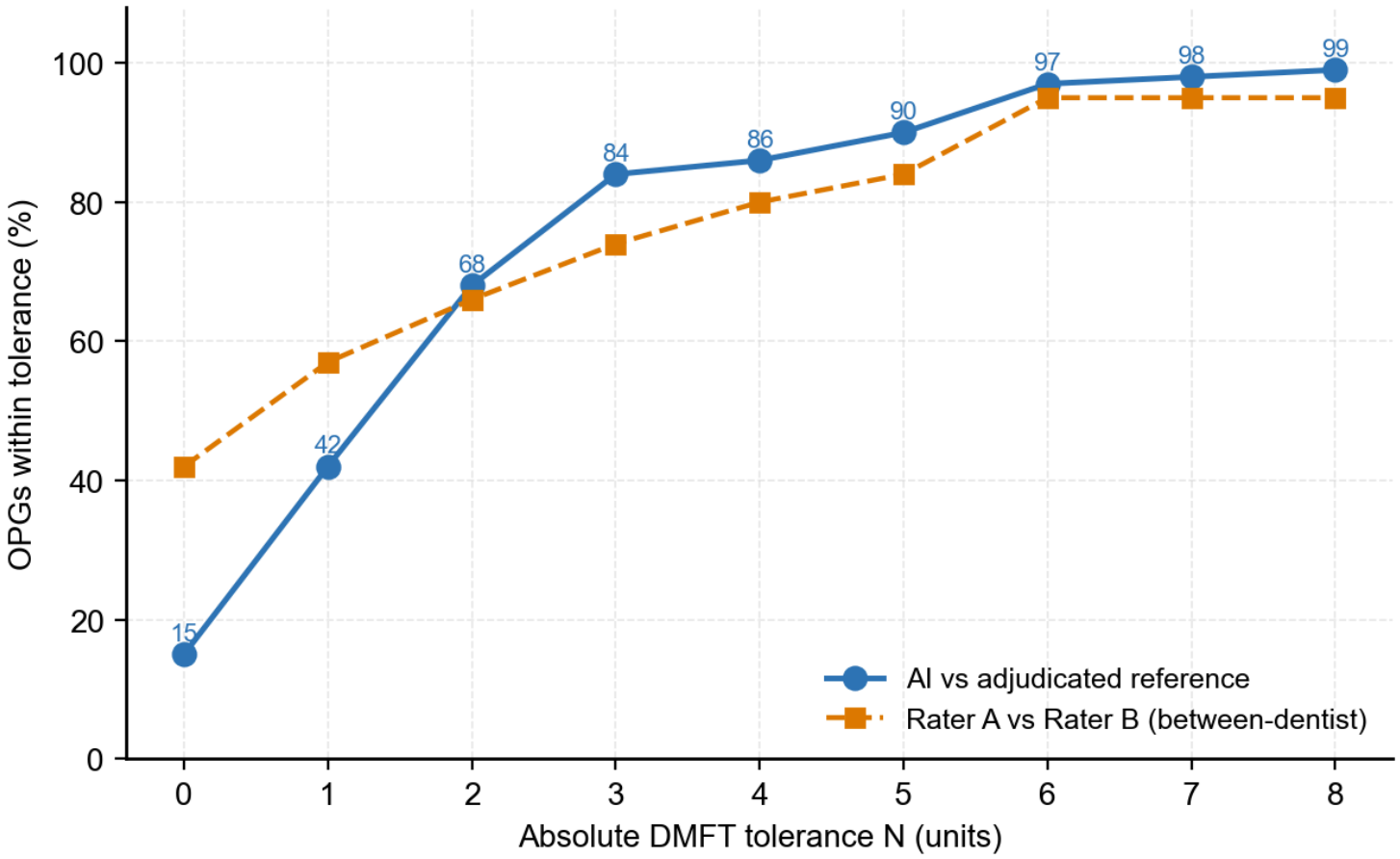


Figure 4

Cumulative agreement on the external Düsseldorf cohort. Each point indicates the fraction of OPGs for which the AI-predicted DMFT (blue) or the second dentist DMFT (orange; Rater A vs Rater B) falls within a given absolute tolerance N of the reference. The between-dentist curve was higher at exact agreement, whereas the AI curve was comparable or higher at clinically relevant tolerances; at a tolerance of 3 DMFT units, AI agreement with the adjudicated reference was 84% of OPGs, versus 74% between the two dentists.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- OnlineResourceSupplementary.docx