Supplementary Information

Deep learning-based treatment suggestion on periapical radiographs in paediatric dentistry: a comparative evaluation of YOLOv11 and YOLO26

Hatice Aydoğdu[1], Bayram Kendigelen[1*], Merve Bilmez Selen[2], İbrahim Şevki Bayrakdar[3]

[1]Department of Pediatric Dentistry, Faculty of Dentistry, Nevşehir Hacı Bektaş Veli University, Nevşehir, Türkiye

[2]Department of Pediatric Dentistry, Faculty of Dentistry, İnönü University, Malatya, Türkiye

[3]Department of Oral and Maxillofacial Radiology, Faculty of Dentistry, Eskişehir Osmangazi University, Eskişehir, Türkiye

[*]Corresponding author: bayramkendigelen@nevsehir.edu.tr

Contents

Completed reporting checklists: **Supplementary Table S1** (CLAIM 2024) and **Supplementary Table S2** (STARD-AI)

Performance on the validation set: **Supplementary Table S3** and **Supplementary Fig. S1**

Representative examples from the test set: **Supplementary Figs. S2–S4**

Completed reporting checklists: CLAIM 2024 and STARD-AI

The study is reported in accordance with the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) 2024 update [1] and the Standards for Reporting of Diagnostic Accuracy Studies–Artificial Intelligence (STARD-AI) reporting guideline [2]. Locations refer to section headings, paragraphs, tables and figures of the manuscript, because page and line numbers change during typesetting. Checklist items are quoted from the original publications.

Supplementary Table S1. CLAIM 2024 checklist (44 items) [1]

No.	Checklist item	Location in manuscript	How the item is reported
Title / Abstract			
1	Identification as a study of AI methodology, specifying the category of technology used (e.g., deep learning)	Title; Abstract	“Deep learning-based” in the title; YOLOv11 and YOLO26 instance segmentation models named in the abstract.
Abstract			
2	Summary of study design, methods, results, and conclusions	Abstract	Unstructured abstract, as required by the journal, covering the design, methods, results and conclusions
Introduction			
3	Scientific and/or clinical background, including the intended use and role of the AI approach	Introduction, paragraphs 1–3	AI applications in paediatric dentistry and the gap between diagnosis and treatment guidance; intended role as the imaging component of future decision-support systems, with the final decision left to the clinician.
4	Study aims, objectives, and hypotheses	Introduction, final paragraph	Aim: to develop, evaluate and compare YOLOv11 and YOLO26. Null hypothesis: no difference in macro-averaged F1-score on the held-out test set.
Methods – Study Design			
5	Prospective or retrospective study	Methods, opening paragraph	Retrospective study.

No.	Checklist item	Location in manuscript	How the item is reported
6	Study goal	Introduction, final paragraph; Methods – Performance evaluation and statistical analysis	Model development and head-to-head comparison of 2 architectures on a held-out test set.
Methods – Data			
7	Data sources	Methods, data source and eligibility paragraphs	Digital archive of a single institution (Faculty of Dentistry, Nevşehir Hacı Bektaş Veli University); radiographs acquired during routine care between October 2024 and May 2026.
8	Inclusion and exclusion criteria	Methods, data source and eligibility paragraphs; Fig. 1	Inclusion: age 3–13 years and at least one tooth with a finding requiring a treatment or monitoring decision. Exclusion: non-diagnostic quality or misleading artefacts, dental trauma, no consensus on the reference standard.
9	Data pre-processing	Methods – Development of the deep learning model	Conversion to 8 bit; resizing to 1024 × 1024 pixels with letterbox padding (aspect ratio preserved); on-the-fly augmentation with the settings listed.
10	Selection of data subsets	Methods – Development of the deep learning model; Methods – Reference standard and labelling protocol	Entire radiographs were used, resized with letterbox padding that preserved the aspect ratio; a random subset of 50 radiographs was used for the inter-rater agreement analysis.
11	De-identification methods	Methods, opening paragraph; Methods – Dataset	All data were completely anonymised before analysis; patient identity, sex and exact age were therefore not available.
12	How missing data were handled	Methods, data source and eligibility paragraphs; Methods – Dataset; Fig. 1	Radiographs of non-diagnostic quality or without consensus on the reference standard were excluded; no reference labels or model outputs were missing; unavailable demographic variables were not analysed.

No.	Checklist item	Location in manuscript	How the item is reported
13	Image acquisition protocol	Methods, data source and eligibility paragraphs	Planmeca ProX unit, photostimulable phosphor plates, bisecting-angle technique, 60–70 kV, 2 mA, 0.05 s; plates read with a Dürr VistaScan Mini Plus scanner.
Methods – Reference Standard			
14	Definition of method(s) used to obtain reference standard	Methods – Reference standard and labelling protocol; Table 1	Eight management categories with radiographic criteria based on AAPD and IAPD guidance; polygon annotation of the entire tooth; dual labels when 2 options were equally acceptable (Table 2).
15	Rationale for choosing the reference standard	Methods – Reference standard and labelling protocol	No standardised labelling scheme exists for this task; without clinical information, only the radiographic components of the guidelines could be applied; dual labels avoid arbitrary single choices.
16	Source of reference standard annotations	Methods – Reference standard and labelling protocol	Initial annotation by a paediatric dentistry research assistant (4 years of annotation experience); independent review by 2 paediatric dentistry specialists (≥ 3 years of clinical experience); adjudication by a third specialist; final review by an oral and maxillofacial radiologist (> 10 years); CranioCatch Annotation Tool.
17	Annotation of test set	Methods – Reference standard and labelling protocol	All radiographs, including those of the test set, were annotated with the same protocol before model development; the annotators had no access to clinical information or model outputs.
18	Measures of inter- and intra-rater variability of features described by the annotators	Methods – Reference standard and labelling protocol (inter-rater agreement paragraph)	Cohen’s κ on 50 radiographs (107 teeth) labelled independently by 2 specialists: weighted mean κ = 0.865; category-level κ 0.494–1.000. Intra-rater variability was not assessed.

No.	Checklist item	Location in manuscript	How the item is reported
Methods – Data Partitions			
19	How data were assigned to partitions	Methods – Dataset; Table 3; Fig. 1	Random image-level split without stratification into training (3060), validation (382) and test (383) radiographs, i.e., 80%, 10% and 10%.
20	Level at which partitions are disjoint	Methods – Dataset; Discussion, paragraph 13 (third limitation)	Partitions are disjoint at the radiograph level; because the data were anonymised, radiographs of the same patient may occur in different subsets.
Methods – Testing Data			
21	Intended sample size	Methods – Dataset	No a priori sample size calculation was performed; all eligible radiographs were included. The test set comprised 383 radiographs with 971 reference labels.
Methods – Model			
22	Detailed description of model	Methods – Development of the deep learning model; Discussion, paragraphs 3–5	YOLOv11l-seg and YOLO26l-seg instance segmentation models; input: periapical radiograph; output: tooth masks with management category and confidence score; architectural differences summarised in the Discussion.
23	Software libraries, frameworks, and packages	Methods – Development of the deep learning model	Ultralytics 8.4.14, Python 3.12.10 and PyTorch 2.12.0.
24	Initialization of model parameters	Methods – Development of the deep learning model	Official Ultralytics checkpoints pretrained on the COCO dataset.

No.	Checklist item	Location in manuscript	How the item is reported
Methods – Training			
25	Details of training approach	Methods – Development of the deep learning model	400 scheduled epochs, batch size 6, early stopping after 100 epochs without improvement, MuSGD optimiser (learning rate 0.01, momentum 0.9, weight decay 0.0005), random seed 0; augmentation settings; overlap_mask disabled so that both labels of a dual-labelled tooth were kept; computing hardware described.
26	Method of selecting the final model	Methods – Development of the deep learning model	Weights of the epoch with the highest validation fitness (box plus mask mAP50–95): epoch 102 for YOLOv11 and epoch 123 for YOLO26.
27	Ensembling techniques	Not applicable	No ensembling was used; each model was trained and evaluated separately.
Methods – Evaluation			
28	Metrics of model performance	Methods – Performance evaluation and statistical analysis	Primary: macro-averaged F1-score. Secondary: per-category precision, recall and F1-score, micro-averaged F1-score, mask mAP50 and mAP50–95, Dice and Jaccard coefficients. A true positive required the same category and a mask IoU ≥ 0.5.
29	Statistical measures of significance and uncertainty	Methods – Performance evaluation and statistical analysis	95% CIs by image-level bootstrap resampling (5000 iterations, percentile method); paired bootstrap for the between-model difference in macro-averaged F1-score, reported with its 95% CI.
30	Robustness or sensitivity analysis	Methods – Performance evaluation and statistical analysis; Results – Secondary analysis of dual-labelled teeth	Secondary analysis of dual-labelled teeth in which either acceptable category was credited; no robustness analysis across acquisition or image-quality conditions was performed.

No.	Checklist item	Location in manuscript	How the item is reported
31	Methods for explainability or interpretability	Not performed	No saliency or attribution methods were applied; limited interpretability is discussed (Discussion, paragraph 12); representative predictions are shown in Supplementary Figs. S2–S4.
32	Evaluation on internal data	Methods – Dataset; Results – Model comparison on the test set	Held-out internal test set (383 radiographs) that was not used for training, model selection or threshold selection.
33	Testing on external data	Not performed; Discussion, paragraph 14 (fourth limitation)	No external test set; external, multicentre validation is stated as required before clinical application.
34	Clinical trial registration	Not applicable	Retrospective study; not a clinical trial.
Results – Data			
35	Numbers of patients or examinations included and excluded	Fig. 1; Methods, data source and eligibility paragraphs	5186 radiographs screened, 1361 excluded and 3825 included; the number of patients is unknown because the data were anonymised.
36	Demographic and clinical characteristics of cases in each partition	Methods – Dataset; Methods, data source and eligibility paragraphs; Table 3	Age range 3–13 years; number of labels per category in each partition (Table 3); 88.8% primary teeth in the agreement subset; sex and exact age not available.
Results – Model Performance			
37	Performance metrics and measures of statistical uncertainty	Results – Model comparison on the test set; Results – Overall model performance on the validation set; Tables 4 and 5; Supplementary Table S3 and Supplementary Fig. S1	Macro- and micro-averaged F1-scores with 95% CIs; between-model difference with 95% CI; validation results with 95% CIs in Supplementary Table S3.

No.	Checklist item	Location in manuscript	How the item is reported
38	Estimates of diagnostic performance and their precision	Results – Per-category performance on the test set; Results – Segmentation quality; Tables 4 and 5	Per-category precision, recall and F1-score, with 95% CIs for the F1-scores; per-category Dice and Jaccard coefficients, with 95% CIs for their means.
39	Failure analysis of incorrect results	Results – Confusion matrix analysis; Fig. 2; Discussion, paragraph 6; Supplementary Figs. S2–S4	Missed labels, unmatched predictions and between-category confusions; clinically relevant errors (restoration predicted for teeth labelled for root canal treatment) discussed; examples of errors in Supplementary Figs. S2–S4.
Discussion			
40	Study limitations	Discussion, paragraphs 11–15 (limitations)	Radiograph-only reference standard; no separate diagnosis labels; image-level split; unavailable demographics; exclusion of trauma; single centre and imaging system; few test labels in some categories; single confidence threshold.
41	Implications for practice, including intended use and/or clinical role	Discussion, paragraph 2; Discussion, paragraphs 8–10; Conclusion	A methodological step from diagnostic towards treatment-oriented tasks, not a substitute for clinical decision-making; integration of clinical data and external validation are required.
Other Information			
42	Provide a reference to the full study protocol or to additional technical details	Supplementary Table S3 and Supplementary Figs. S1–S4; Data availability	No separate protocol was published; additional results (validation set) and examples are provided in Supplementary Table S3 and Supplementary Figs. S1–S4; code available on request.
43	Statement about the availability of software, trained model, and/or data	Data availability	Data and code are available from the corresponding author on reasonable request.

No.	Checklist item	Location in manuscript	How the item is reported
44	Sources of funding and other support; role of funders	Funding	Health Institutes of Türkiye (TÜSEB), project number 48468; the funder had no role in the study.

AAPD, American Academy of Pediatric Dentistry; AI, artificial intelligence; CI, confidence interval; COCO, Common Objects in Context; IAPD, International Association of Paediatric Dentistry; IoU, intersection over union; mAP, mean average precision.

Supplementary Table S2. STARD-AI checklist (48 items, including sub-items) [2]

No.	Checklist item	Location in manuscript	How the item is reported
Title or abstract			
1	Identification as a study reporting AI-centred diagnostic accuracy and reporting at least one measure of accuracy within title or abstract	Title; Abstract	Deep learning-based models named in the title; macro-averaged F1-scores with 95% CIs reported in the abstract.
Abstract			
2	Structured summary of study design, methods, results and conclusions (for specific guidance, please see STARD for Abstracts)	Abstract	Unstructured abstract, as required by the journal, covering the design, methods, results and conclusions
Introduction			
3	Scientific and clinical background, including the intended use of the index test, whether it is novel or an established index test, and its integration into an existing or new workflow, if applicable	Introduction, paragraphs 1–3	Novel index test (radiograph-based treatment suggestion); intended as the imaging component of future multimodal decision-support systems, with the final decision remaining with the clinician.
4	Study objectives and hypotheses	Introduction, final paragraph	Aim and null hypothesis (no difference in macro-averaged F1-score on the held-out test set).
Methods – Study design			
5	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	Methods, opening paragraph; Methods, data source and eligibility paragraphs	Retrospective study of archived routine radiographs.

No.	Checklist item	Location in manuscript	How the item is reported
Methods – Ethics			
6	Formal approval from an ethics committee. If not required, justify why	Methods, opening paragraph	Approved by the Non-Interventional Clinical Research Publication Ethics Committee of Nevşehir Hacı Bektaş Veli University (decision no. 2025.08.01, 31 July 2025); informed consent waived.
Methods – Participants			
7	Eligibility criteria: listing separate inclusion and exclusion criteria in the order that they are applied at both participant level and data level	Methods, data source and eligibility paragraphs; Fig. 1	Participant level: age 3–13 years. Data level: at least one tooth with a finding requiring a treatment or monitoring decision; diagnostic image quality; no dental trauma; consensus on the reference standard.
8	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	Methods, data source and eligibility paragraphs	Radiographs of children acquired during routine dental care and identified in the institutional archive; inclusion based on radiographic findings.
9	Where and when potentially eligible participants were identified (setting, location, and dates)	Methods, opening paragraph; Methods, data source and eligibility paragraphs	Faculty of Dentistry, Nevşehir Hacı Bektaş Veli University, Nevşehir, Türkiye; radiographs acquired between October 2024 and May 2026.
10	Whether participants formed a consecutive, random, or convenience series	Methods, data source and eligibility paragraphs	Consecutive series: all radiographs acquired during the study period were screened.
Methods – Dataset			
11	Source of the data and whether it has been routinely collected, specifically collected for the purpose of the study or acquired from an open-source repository	Methods, data source and eligibility paragraphs	Routinely collected clinical radiographs from a single institutional archive; not acquired for the study.

No.	Checklist item	Location in manuscript	How the item is reported
12	Who undertook the annotations for the dataset (including experience levels and background) and how (within the same clinical context or in a post-hoc fashion), if applicable	Methods – Reference standard and labelling protocol	Initial annotation by a paediatric dentistry research assistant; review by 2 paediatric dentistry specialists, a third specialist for disagreements and an oral and maxillofacial radiologist, with experience levels stated; post hoc annotation of archived radiographs without clinical information.
13	Devices (manufacturer, model) that were used to capture data; software (with version number) used to engineer the index test, highlighting the intended use	Methods, data source and eligibility paragraphs; Methods – Reference standard and labelling protocol; Methods – Development of the deep learning model	Planmeca ProX unit, photostimulable phosphor plates, Dürr VistaScan Mini Plus scanner; CranioCatch Annotation Tool; Ultralytics 8.4.14, Python 3.12.10, PyTorch 2.12.0; intended use: radiograph-based treatment suggestion.
14	Data acquisition protocols (e.g. contrast protocol or reconstruction method for medical images) and details of data pre-processing in sufficient detail to allow replication	Methods, data source and eligibility paragraphs; Methods – Development of the deep learning model	Bisecting-angle technique, 60–70 kV, 2 mA, 0.05 s; conversion to 8 bit, letterbox resizing to 1024 × 1024 pixels and augmentation settings.
Methods – Test methods			
15a	Index test, in sufficient detail to allow replication	Methods – Development of the deep learning model; Methods – Performance evaluation and statistical analysis	Architectures, pretrained weights, software versions, hyperparameters, augmentation, model selection, confidence thresholds and matching rules.
15b	How the index test was developed, including any training, validation, testing and external evaluation, detailing sample sizes, when applicable	Methods – Dataset; Methods – Development of the deep learning model; Table 3; Fig. 1	Training 3060 radiographs (7799 labels), validation 382 (987) and test 383 (971); no external evaluation.

No.	Checklist item	Location in manuscript	How the item is reported
15c	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	Methods – Performance evaluation and statistical analysis; Results – Selection of confidence thresholds	Result categories: the 8 management categories. Confidence thresholds were set at the maximum of the validation F1–confidence curve (0.35 and 0.41) and fixed for the test set; a true positive required the same category and a mask IoU ≥ 0.5.
15d	The specified end user of the index test and the level of expertise required of users	Introduction, paragraph 3; Discussion, paragraph 2; Conclusion	Intended for clinicians, who interpret the suggestions together with the history and clinical examination; not intended for use without clinical assessment.
16a	Reference standard, in sufficient detail to allow replication	Methods – Reference standard and labelling protocol; Tables 1 and 2	Labelling protocol, radiographic criteria for each category, annotation and review process, and dual-label rules.
16b	Rationale for choosing the reference standard (if alternatives exist)	Methods – Reference standard and labelling protocol; Discussion, paragraph 11 (first limitation)	Expert radiograph-based labels were used because no standardised scheme or clinical information was available; validation against clinical records and treatment outcomes is identified as future work.
16c	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	Methods – Reference standard and labelling protocol; Table 1	Eight categories with radiographic criteria, pre-specified in the labelling protocol; dual labels for teeth with 2 equally acceptable options.
17a	Whether clinical information and reference standard results were available to the performers or readers of the index test	Methods – Development of the deep learning model; Methods – Performance evaluation and statistical analysis	The models received radiographs only; test-set reference labels were not used for training, model selection or threshold selection.
17b	Whether clinical information and index test results were available to the assessors of the reference standard	Methods – Reference standard and labelling protocol	Annotators had no clinical information and no access to model outputs; the reference standard was established before model development.

No.	Checklist item	Location in manuscript	How the item is reported
Methods – Analysis			
18	Methods for estimating or comparing measures of diagnostic accuracy	Methods – Performance evaluation and statistical analysis	Label-level matching (same category, mask IoU ≥ 0.5); macro-averaged F1-score as the primary outcome; bootstrap 95% CIs and paired bootstrap comparison of the 2 models.
19	How indeterminate index test or reference standard results were handled	Methods – Reference standard and labelling protocol; Methods, data source and eligibility paragraphs	Radiographically ambiguous teeth received dual labels; radiographs without consensus were excluded; predictions below the confidence threshold were discarded.
20	How missing data on the index test and reference standard were handled	Methods, data source and eligibility paragraphs; Fig. 1	No index test outputs were missing; radiographs without an agreed reference standard were excluded.
21	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	Results – Per-category performance on the test set; Results – Secondary analysis of dual-labelled teeth; Discussion, paragraph 13 (third limitation)	Per-category performance (secondary outcome) and a secondary analysis of dual-labelled teeth; no subgroup analysis by dentition or tooth type.
22	Intended sample size and how it was determined	Methods – Dataset	No a priori sample size calculation was performed; all eligible radiographs were included.
23	Details of any performance error analysis, and algorithmic bias and fairness assessments if undertaken	Results – Confusion matrix analysis; Fig. 2; Discussion, paragraph 6; Supplementary Figs. S2–S4	Error analysis by confusion matrices and examples; algorithmic bias and fairness were not assessed because demographic data were unavailable.
Results – Participants and dataset			
24	Flow of participants, using a diagram	Fig. 1	Flow diagram of screening, exclusion and partitioning.
25	Baseline demographic, clinical and technical characteristics of training, validation and test set, if applicable	Methods – Dataset; Table 3; Methods, data source and eligibility paragraphs	Age range 3–13 years in all sets; category distribution per set (Table 3); identical device and acquisition protocol; sex and exact age unavailable.

No.	Checklist item	Location in manuscript	How the item is reported
26a	Distribution of severity of disease in those with the target condition	Tables 2 and 3	Severity was not graded; distribution of management categories per set (Table 3) and of dual-label combinations (Table 2).
26b	Distribution of alternative diagnoses in those without the target condition	Not applicable	Teeth without findings requiring management were not labelled and served as negatives; their diagnoses were not recorded.
27	Time interval and any clinical interventions between index test and reference standard	Not applicable	The index test and the reference standard were applied to the same radiograph.
28	Whether the datasets represent the distribution of the target condition that one would expect from the intended use population	Methods, data source and eligibility paragraphs; Table 3	Routine radiographs from a single centre; only radiographs with at least one finding requiring management were included, so the distribution reflects radiographs that require treatment decisions rather than unselected radiographs.
29	For external evaluation on an independent dataset, an assessment of how this differs from the training, validation and test sets	Not applicable	No external evaluation was performed (Discussion, paragraph 14).
Results – Test results			
30	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	Fig. 2; Supplementary Fig. S1	Confusion matrices of predicted versus reference categories, including missed labels and unmatched predictions.
31	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	Results – Model comparison on the test set; Tables 4 and 5; Supplementary Table S3	Macro-averaged, micro-averaged and per-category F1-scores and mean Dice and Jaccard coefficients with bootstrap 95% CIs; between-model difference with 95% CI.
32	Any adverse events from performing the index test or the reference standard	Not applicable	Retrospective analysis of existing radiographs; no additional procedures were performed.

No.	Checklist item	Location in manuscript	How the item is reported
Discussion			
33	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	Discussion, paragraphs 11–15 (limitations)	Reference standard, interpretability, image-level split with possible optimistic bias, unavailable demographics, single centre, wide confidence intervals for small categories, single threshold.
34	Implications for practice, including the intended use and clinical role of the index test	Discussion, paragraph 2; Discussion, paragraphs 8–10; Conclusion	A methodological step towards treatment-oriented tasks, not for direct clinical decision-making; external validation required.
35	Ethical considerations and adherence to ethical standards associated with the use of the index test and issues of fairness	Methods, opening paragraph; Discussion, paragraph 6; Discussion, paragraphs 8–10; Conclusion	Ethics approval and Declaration of Helsinki; risk of under-treatment from errors and the need for clinical oversight discussed; fairness across demographic groups not assessed (no demographic data).
Other information			
36	Registration number and name of registry	Not applicable	Retrospective study; not registered.
37	Where the full study protocol can be accessed	Not publicly available	The protocol approved by the ethics committee was not published; the methods are described in full in the manuscript and Supplementary Table S3 and Supplementary Figs. S1–S4.
38	Sources of funding and other support; role of funders	Funding	Health Institutes of Türkiye (TÜSEB), project number 48468; the funder had no role in the study.
39	Commercial interests, if applicable	Competing interests	İ.Ş.B. is the co-founder and CEO of CranioCatch, whose annotation tool was used in this study; the other authors declare no competing interests.
40a	Availability of datasets and code; detailing any restrictions on their reuse and repurposing	Data availability	Data and code are available from the corresponding author on reasonable request.

No.	Checklist item	Location in manuscript	How the item is reported
40b	Whether outputs are stored, auditable and available for evaluation, if necessary	Data availability	Model predictions for the test set are stored and available from the corresponding author on reasonable request.

CI, confidence interval; IoU, intersection over union.

References

1. Tejani, A. S. et al. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. Radiol. Artif. Intell. 6, e240300 (2024).

2. Sounderajah, V. et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat. Med. 31, 3283–3289 (2025).

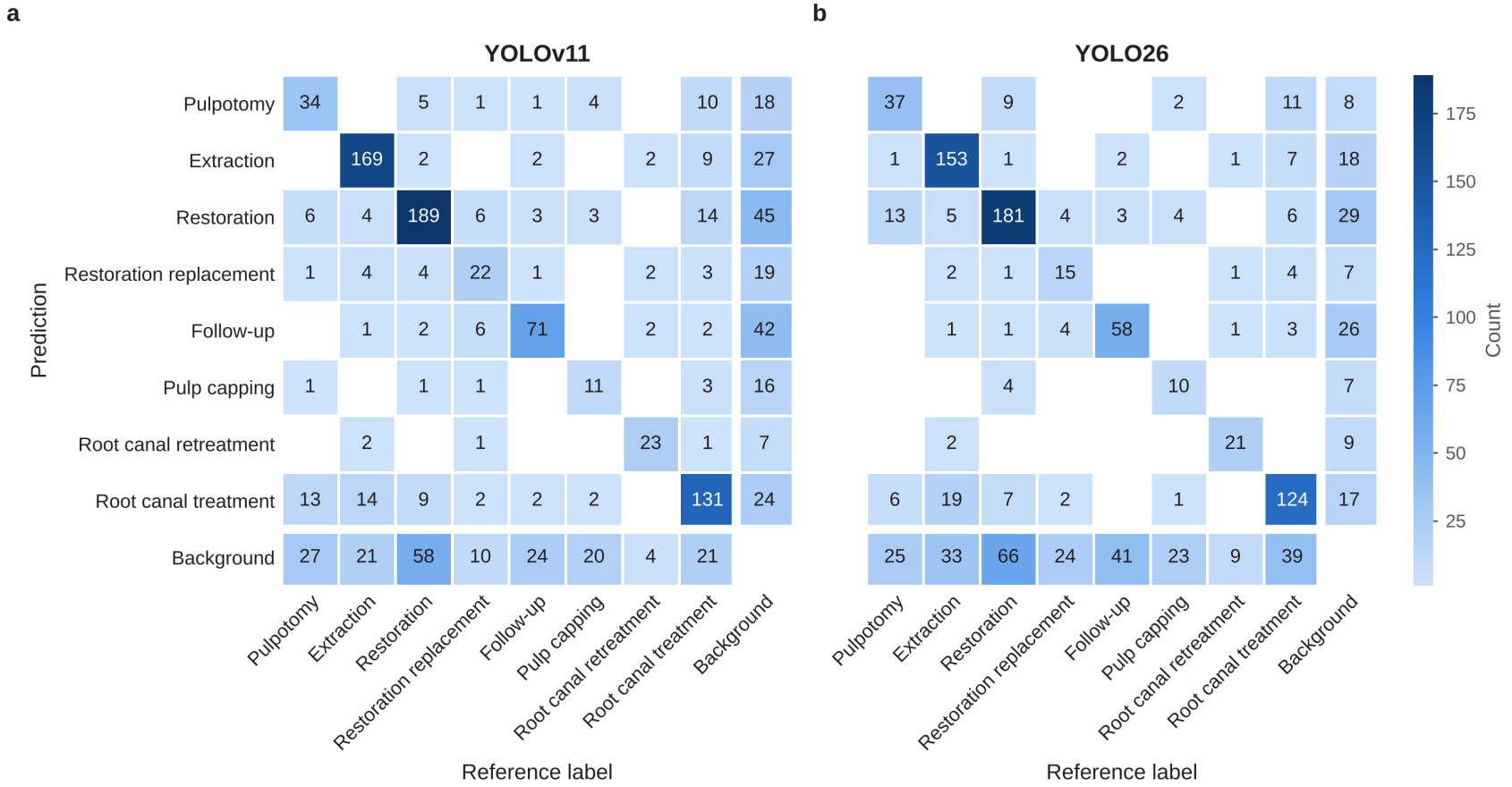
Performance on the validation set

Supplementary Table S3. Per-category performance of YOLOv11 and YOLO26 on the validation set.

Category	n	YOLOv11			YOLO26		
		Precision	Recall	F1 (95% CI)	Precision	Recall	F1 (95% CI)
Pulpotomy	82	0.466	0.415	0.439 (0.338–0.531)	0.552	0.451	0.497 (0.384–0.597)
Extraction	215	0.801	0.786	0.793 (0.749–0.834)	0.836	0.712	0.769 (0.719–0.813)
Restoration	270	0.700	0.700	0.700 (0.655–0.742)	0.739	0.670	0.703 (0.655–0.748)
Restoration replacement	49	0.393	0.449	0.419 (0.289–0.532)	0.500	0.306	0.380 (0.233–0.517)
Follow-up	104	0.563	0.683	0.617 (0.539–0.690)	0.617	0.558	0.586 (0.497–0.667)
Pulp capping	40	0.333	0.275	0.301 (0.161–0.435)	0.476	0.250	0.328 (0.174–0.478)
Root canal retreatment	33	0.676	0.697	0.687 (0.542–0.805)	0.656	0.636	0.646 (0.500–0.769)
Root canal treatment	194	0.665	0.675	0.670 (0.612–0.723)	0.705	0.639	0.670 (0.608–0.729)
Macro average	**987**	**0.575**	**0.585**	**0.578 (0.539–0.612)**	**0.635**	**0.528**	**0.572 (0.530–0.611)**
Micro average	**987**	**0.650**	**0.659**	**0.654 (0.625–0.682)**	**0.706**	**0.607**	**0.653 (0.623–0.683)**

n, number of reference labels; CI, bootstrap confidence interval.

Supplementary Figure S1. Confusion matrices of YOLOv11 (a) and YOLO26 (b) on the validation set. Rows are predictions and columns are reference labels; the background row and column contain unmatched reference labels and unmatched predictions, respectively. Cell values are label counts; darker shading indicates higher counts.



Representative examples from the test set

To support the quantitative results, this supplementary material presents representative examples from the test set with high, partial and low agreement between the reference standard and the model predictions (Supplementary Figs. S2–S4).

The panels are organised as follows.

Colour coding: Each of the 8 management categories is shown in a distinct colour, which is identical in all panels.

Panel layout: Each example is shown in 3 panels:

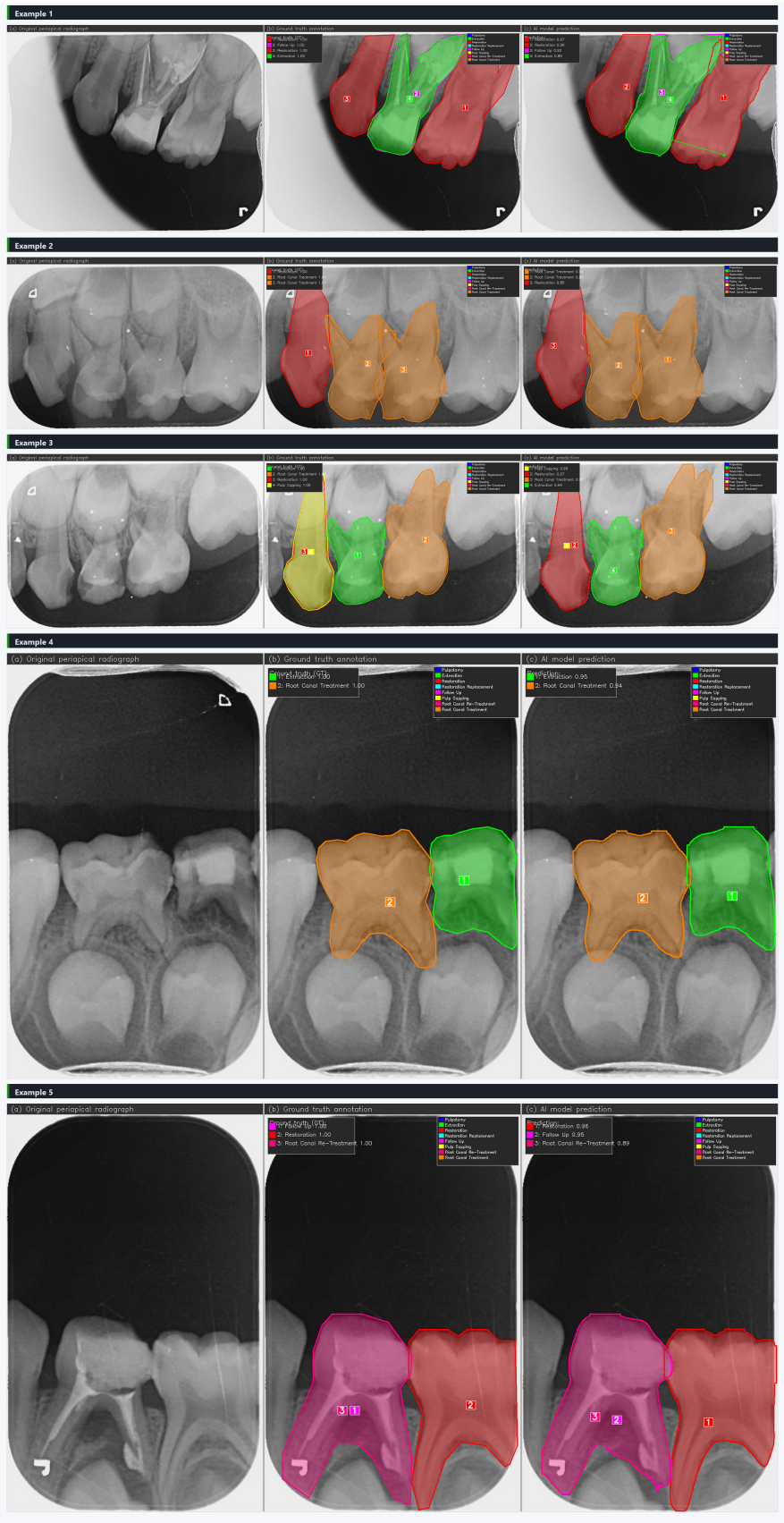
(a) Original periapical radiograph: the unannotated radiograph.

(b) Ground truth annotation: the reference standard, that is, the polygons and management categories assigned according to the labelling protocol described in the main text. Two overlapping polygons of different colours on the same tooth indicate a dual label.

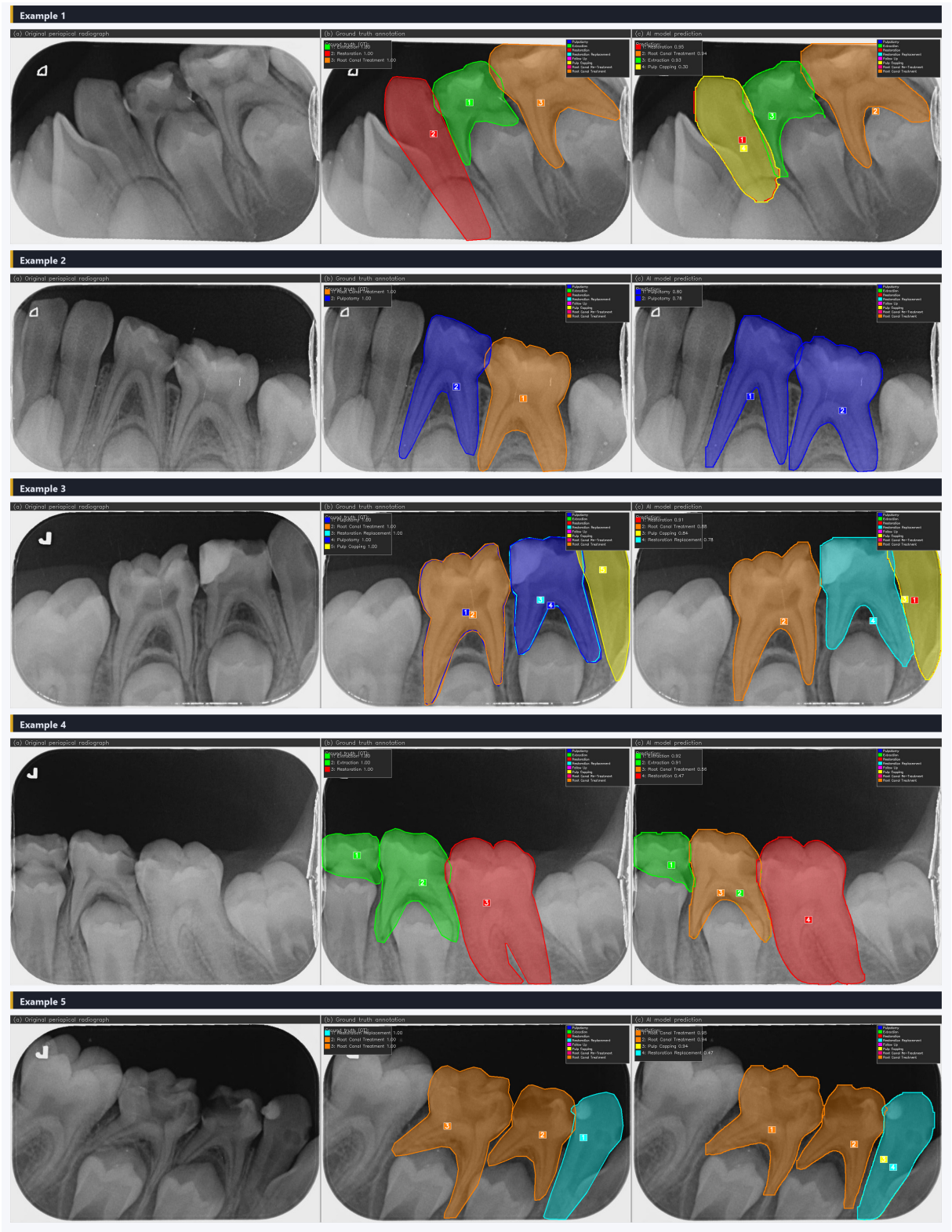
(c) AI model prediction: the instance segmentation output of YOLO26, with the predicted management category and the confidence score of each instance. The panels also show some predictions with confidence scores below the operating threshold used in the quantitative analysis.

Selection of examples: The examples were selected by the authors to illustrate correct predictions, partial agreement and errors, including missed labels and misclassifications. They are not a random sample and do not represent the overall performance, which is reported in the main text.

Supplementary Figure S2. Examples with high agreement between the reference standard and the predictions.



Supplementary Figure S3. Examples with partial agreement between the reference standard and the predictions.



Supplementary Figure S4. Examples with low agreement, including missed labels and misclassifications.

