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.

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

Hatice Aydoğdu
Nevşehir Hacı Bektaş Veli University
Bayram Kendigelen
bayramkendigelen@nevsehir.edu.tr

Nevşehir Hacı Bektaş Veli University
Merve Bilmez Selen
Inonu University
İbrahim Şevki Bayrakdar
Eskişehir Osmangazi University

Article

Keywords: Paediatric dentistry, Artificial intelligence, Deep learning, Periapical radiography, Treatment suggestion

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11664484/v1

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

Additional Declarations: No competing interests reported.

Abstract

To compare YOLOv11 and YOLO26 instance segmentation models for suggesting management categories from paediatric periapical radiographs. From a single centre, 3825 periapical radiographs of children aged 3–13 years were annotated with 8 management categories assignable from radiographs alone: restoration, restoration replacement, pulp capping, pulpotomy, root canal treatment, root canal retreatment, extraction and follow-up. Teeth with 2 equally acceptable categories received dual labels. Interobserver agreement was high (mean Cohen's kappa, 0.865). Confidence thresholds were selected on the validation set; the primary outcome, test-set macro-averaged F1-score, was compared with a paired bootstrap. The macro-averaged F1-score (95% confidence interval) was 0.610 (0.571–0.643) for YOLOv11 and 0.595 (0.553–0.632) for YOLO26, with no clear difference (− 0.014 [− 0.052 to 0.024]); YOLO26 had higher precision and lower recall. F1-scores were highest for extraction (0.805 and 0.780) and lowest for restoration replacement, pulp capping, pulpotomy and follow-up (0.43–0.56). For dual-labelled teeth, at least one acceptable category was predicted in 78.2% (YOLOv11) and 70.7% (YOLO26) of cases. Both architectures learned radiograph-based management categories with similar performance, lowest for categories whose distinction depends on clinical information. The models are a methodological step from diagnostic towards treatment-oriented tasks, not a substitute for clinical decision-making.

Introduction

Artificial intelligence (AI) has been increasingly investigated in paediatric dentistry, and systematic reviews report promising diagnostic performance across a range of applications, although real-world clinical translation remains limited [1, 2]. Deep learning models, particularly convolutional neural networks (CNNs) such as You Only Look Once (YOLO)-family detectors, have been applied to diverse tasks in paediatric dentistry. These include automated tooth numbering in mixed dentition using cone-beam computed tomography [3], detection of white spot lesions and molar incisor hypomineralisation from intraoral photographs [4, 5], and localisation of ectopic canines and molars [6]. Additional applications encompass dental age estimation [7], early childhood caries risk prediction [8], and mesiodens identification [9]. However, current AI research in paediatric dentistry has largely remained limited to diagnostic and predictive tasks, with most models designed as single-task systems targeting isolated pathologies [1, 2]. This creates a diagnostic-therapeutic gap — a disconnect between identification of findings and actionable treatment guidance.

Beyond diagnosis, systems that translate imaging findings into treatment recommendations remain scarce in dentistry, and their impact on treatment decisions has not yet been established [10]. In paediatric dentistry, a recent study mapped radiographic furcation-lesion categories of primary molars to root canal treatment or extraction on panoramic radiographs [11]; however, to our knowledge, the broader range of restorative, pulp-therapy, extraction and monitoring decisions has not been addressed on periapical radiographs. Learning these decisions directly from expert-labelled radiographs, rather than

through fixed rules, could serve as a methodological step towards extending dental AI research from diagnostic to treatment-oriented tasks.

To address this gap, the present study explores whether deep learning models can predict, from periapical radiographs alone, the treatment category that expert paediatric dentists would suggest on radiographic grounds, across 8 categories commonly encountered in paediatric practice. Because the approach is image-based, its outputs reflect radiographic findings only; the final treatment decision, which also depends on history and clinical examination, remains with the clinician. The model is conceived as the imaging component of future multimodal decision-support systems, such as those already explored for orthodontic extraction decisions [12], which may further incorporate clinical information.

YOLO-family instance segmentation models were selected because they localise and classify multiple regions in a single pass, which suits periapical radiographs that often show several teeth with different treatment needs, and because they are widely used in dental radiography, allowing comparison with previous studies [13–15]. Since architecture superiority has proven task-dependent in dental imaging [13, 15], 2 recent releases from the same framework, YOLOv11 and YOLO26 [16, 17], were trained with an identical pipeline and compared. This study aimed to develop and evaluate YOLOv11 and YOLO26 for radiograph-based treatment suggestion in paediatric dentistry and compare their performance. The null hypothesis was that YOLOv11 and YOLO26 would not differ in macro-averaged F1-score on the held-out test set.

Methods

This retrospective study was conducted at the Faculty of Dentistry, Nevşehir Hacı Bektaş Veli University (Nevşehir, Türkiye), and is reported in accordance with the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) guidelines [18] and the Standards for Reporting of Diagnostic Accuracy Studies-Artificial Intelligence (STARD-AI) reporting guideline [19]; the completed checklists are provided in Supplementary Tables S1 and S2. The study protocol was reviewed and approved by the Non-Interventional Clinical Research Publication Ethics Committee of Nevşehir Hacı Bektaş Veli University (decision no. 2025.08.01, 31 July 2025). Owing to the retrospective design and the complete anonymisation of all data, the ethics committee waived the requirement for informed consent, including informed consent from the participants' parents or legal guardians. All methods were carried out in accordance with relevant guidelines and regulations and with the 1964 Declaration of Helsinki and its later amendments.

Periapical radiographs were retrieved from the digital archive of a single institution, the Faculty of Dentistry, Nevşehir Hacı Bektaş Veli University. Radiographs of patients aged 3–13 years acquired during routine care between October 2024 and May 2026 were eligible. **All radiographs acquired during this period were screened consecutively.** A radiograph was included if it showed at least one primary or permanent tooth with a radiographic finding requiring a treatment or monitoring decision, such as dental

caries, a periapical lesion, root resorption, a defective restoration, or a previously treated tooth. More than one radiograph per patient could be included. Radiographs showing no such finding, for example those with only caries-free teeth, were excluded because they contain no target instances. Consequently, every included radiograph contained at least one annotated tooth.

Teeth without findings requiring management were left unlabelled; as radiographs typically also contained such teeth, these served as negative examples, and predictions on them were counted as false positives. Minor artefacts were accepted if they did not impair interpretation of the region of interest, whereas radiographs of non-diagnostic quality or with artefacts likely to cause misinterpretation or unreliable labelling, such as cone cut, marked elongation or foreshortening, motion blur, or phosphor plate scratches or bending, were excluded. Radiographs showing dental trauma and those for which the annotators could not reach consensus on the reference standard were also excluded. Of the 5186 radiographs screened, 1361 were excluded and 3825 were included in the analysis (Fig. 1). All radiographs were acquired with the same ProX intraoral X-ray unit (Planmeca Oy, Helsinki, Finland) using photostimulable phosphor plates and the bisecting-angle technique, at a tube voltage of 60–70 kV adjusted to patient size and tooth region, with a fixed tube current of 2 mA and an exposure time of 0.05 s. The plates were read with a VistaScan Mini Plus scanner (Dürr Dental, Bietigheim-Bissingen, Germany).

Reference standard and labelling protocol

In the absence of a standardised labelling scheme for radiograph-based treatment suggestion in paediatric dentistry, a labelling protocol was developed for this study. Each annotated tooth was assigned the management category that could be considered on the basis of its radiographic findings alone; because no clinical information was available, only the radiographic components of the current guidelines of the American Academy of Pediatric Dentistry (AAPD) [20–22] and the International Association of Paediatric Dentistry (IAPD) [23,24] were applied. The label therefore represents a radiographic suggestion rather than a definitive treatment decision. Eight management categories commonly encountered in paediatric practice were defined: restoration, restoration replacement, root canal treatment, root canal retreatment, pulpotomy, pulp capping, extraction, and follow-up. The categories were defined at the level of treatment type and applied to both primary and permanent teeth, except root canal retreatment, which was assigned only to permanent teeth; the dentition-specific technique (e.g., pulpectomy versus conventional root canal treatment, or the choice of pulpotomy agent) was left to the clinician and was not predicted by the model. The radiographic criteria used for each category are summarised in Table 1.

Table 1. Radiographic criteria used to assign each management category.

Category	Radiographic criteria	Dentition
Restoration	Carious lesion on an unrestored surface extending into dentine but not approaching the pulp; no radiographic signs of pulpal or periradicular pathology	Primary and permanent
Restoration replacement	Existing restoration with secondary caries or a marginal defect (gap, overhang, or fracture), without pulpal or periradicular involvement	Primary and permanent
Pulp capping	Deep carious lesion approaching the pulp, without radiographic signs of pulpal or periradicular pathology; encompasses indirect pulp treatment and direct pulp capping, the choice between them being made clinically	Primary and permanent
Pulpotomy	Carious lesion reaching or likely to expose the pulp, with no furcal or periapical radiolucency and no internal or pathological external root resorption; sufficient remaining root length in primary teeth	Primary and permanent
Root canal treatment	Furcal or periapical radiolucency, widened periodontal ligament space, or loss of lamina dura in a restorable tooth; in primary teeth, sufficient remaining root length and no involvement of the successor tooth germ	Primary and permanent
Root canal retreatment	Previously root-filled tooth with a persistent or new periapical radiolucency or an inadequate root filling	Permanent only
Extraction	Non-restorable crown destruction, extensive furcal or periapical pathology involving the successor tooth germ, or advanced pathological root resorption	Primary and permanent
Follow-up	Findings requiring monitoring rather than immediate intervention, such as non-cavitated lesions, adequately treated teeth without active pathology, or primary teeth with advanced physiological root resorption close to exfoliation	Primary and permanent

Annotations were created as polygons with the CranioCatch Annotation Tool (CranioCatch, Eskişehir, Türkiye). Each polygon was drawn along the contours of the entire tooth, including any associated radiographic finding such as a periapical or furcal lesion or an overhanging restoration. Only the management category was recorded for each polygon; the underlying diagnosis was not labelled separately.

Initial annotations were made by a dentist and research assistant in paediatric dentistry with 4 years of experience in medical image annotation. All annotations were then reviewed independently by 2 paediatric dentistry specialists, each with at least 3 years of clinical experience. Uncertain cases were reviewed jointly, remaining disagreements were referred to a third paediatric dentistry specialist, and radiographs that remained unresolved were excluded. Finally, all annotations were reviewed by an oral and maxillofacial radiologist with more than 10 years of experience with regard to the radiographic findings and polygon delineation, and any remaining disagreement was resolved by discussion. The annotators had no access to clinical information and assigned labels on the basis of the radiographs

alone. The reference standard was established before model development, and the annotators had no access to model outputs.

A tooth received 2 labels when the radiograph alone could not distinguish between 2 clinically acceptable management options, because the deciding information is clinical or intraoperative. Dual labels were assigned within the same annotation and review process, only in such cases and for no more than 2 categories per tooth; the tooth was then annotated with 2 polygons covering the same region, one for each category. The most frequent pairs reflected recognisable clinical boundaries: extraction versus follow-up (n = 298), typically for primary teeth approaching exfoliation, where the decision depends on mobility, symptoms and the eruption stage of the successor; pulpotomy versus root canal treatment (n = 297), where the condition of the radicular pulp is confirmed only intraoperatively by haemorrhage control [20–22]; follow-up versus root canal retreatment (n = 270), for previously treated teeth with an equivocal periapical status; and restoration versus pulp capping (n = 245), where lesion depth relative to the pulp is finally determined during caries removal. Rather than forcing a single label, which would have introduced arbitrary choices into the reference standard, both options were retained so that the ambiguity inherent in radiograph-only assessment was represented explicitly. Because each dual-labelled tooth carries 2 labels, the 9757 labels in the dataset correspond to 7932 teeth; 1825 of these teeth (23.0%) received dual labels, forming 22 category combinations (Table 2).

Table 2. The 10 most frequent category pairs among the 1825 dual-labelled teeth (22 combinations in total; top 10: 1686, 92.4%).

#	Category Pair	n	%
1	Extraction + Follow-up	298	16.3%
2	Pulpotomy + Root canal treatment	297	16.3%
3	Follow-up + Root canal retreatment	270	14.8%
4	Restoration + Pulp capping	245	13.4%
5	Follow-up + Root canal treatment	176	9.6%
6	Restoration replacement + Follow-up	112	6.1%
7	Extraction + Root canal treatment	93	5.1%
8	Pulpotomy + Pulp capping	92	5.0%
9	Pulpotomy + Restoration	69	3.8%
10	Restoration replacement + Root canal treatment	34	1.9%
Top 10 Subtotal		**1686**	**92.4%**

Inter-rater agreement was assessed with Cohen's kappa on a randomly selected subset of 50 radiographs (107 labelled teeth), which the 2 paediatric dentistry specialists labelled independently from scratch, without access to the initial annotations. For each category, agreement was assessed on the presence or absence of the label on each tooth; when one rater assigned 2 labels and the other only one, the shared label was counted as agreement and the additional label as disagreement. Overall agreement was summarised as the mean of the category-level kappa values weighted by the number of teeth in each category, and kappa values were interpreted according to Landis and Koch [25]. Overall agreement was almost perfect ($\kappa = 0.865$). Category-level kappa values were 1.000 for restoration replacement, 0.876 for extraction, 0.865 for root canal treatment, 0.854 for restoration, 0.811 for pulp capping, 0.758 for pulpotomy, 0.658 for root canal retreatment and 0.494 for follow-up. The lower agreement for follow-up reflects the radiographically heterogeneous nature of this category.

Dataset

The 3825 included radiographs were randomly split at the image level, without stratification, into training (n = 3060), validation (n = 382), and test (n = 383) sets; the number of labels per category in each set is given in Table 3, and the category distribution was similar across the 3 sets. Because the radiographs were fully anonymised before analysis, the number of unique patients, their sex, and their exact age could not be determined, and radiographs from the same patient may therefore appear in more than one subset. No a priori sample size calculation was performed; all eligible radiographs were included.

Table 3. Number of labels per category in the training, validation and test sets.

ID	Category	Train	Validation	Test	Total
0	Restoration	2035	270	238	2543
1	Restoration replacement	426	49	61	536
2	Pulp capping	298	40	45	383
3	Pulpotomy	621	82	77	780
4	Root canal treatment	1636	194	212	2042
5	Root canal retreatment	252	33	24	309
6	Extraction	1691	215	218	2124
7	Follow-up	840	104	96	1040
	Total	**7799**	**987**	**971**	**9757**

Given the age range, most radiographs showed primary or mixed dentition. The dentition of each tooth was not recorded for the whole dataset; in the random subset of 50 radiographs used for the agreement analysis, 95 of 107 labelled teeth (88.8%) were primary teeth.

The categories were imbalanced, with restoration (n = 2543) and extraction (n = 2124) the most frequent and root canal retreatment (n = 309) and pulp capping (n = 383) the least frequent, an approximately 8:1 ratio. Oversampling and class-weighted loss, 2 common strategies for class imbalance [26], were explored in preliminary experiments but did not improve validation performance; the original class distribution was therefore retained for training.

Development of the deep learning model

YOLOv11l-seg and YOLO26l-seg instance segmentation models were initialised from the official Ultralytics checkpoints pretrained on the Common Objects in Context (COCO) dataset and trained with Ultralytics 8.4.14, Python 3.12.10, and PyTorch 2.12.0 [27]. Radiographs were converted to 8-bit and resized to an input size of 1024 × 1024 pixels with letterbox padding, preserving the aspect ratio. Both models used identical settings and on-the-fly augmentation: mosaic (probability 1.0), horizontal flipping (probability 0.5), translation (fraction 0.1), scaling (gain 0.5), and brightness jitter (value gain 0.4 in hue–saturation–value space), the only colour-space augmentation relevant to greyscale radiographs. Rotation, shear, perspective, vertical flipping, mixup, copy-paste, and cutmix were disabled.

Each label was treated as a separate instance, so both labels of a dual-labelled tooth (e.g., pulpotomy and root canal treatment) served as training targets. The overlap_mask option, which merges all instance masks of an image into a single mask in which overlapping pixels belong to only one instance, was therefore disabled; each instance kept its own binary mask in the segmentation loss, and neither polygon of a dual-labelled tooth was overwritten.

Training was scheduled for 400 epochs with a batch size of 6 and early stopping if validation fitness (box plus mask mean average precision [mAP] over intersection over union [IoU] thresholds of 0.50–0.95, mAP50–95) did not improve for 100 consecutive epochs. Validation fitness peaked at epoch 102 for YOLOv11 (training stopped at epoch 202) and at epoch 123 for YOLO26 (training stopped at epoch 223); the weights from these best epochs were used for evaluation. With the optimizer argument set to auto, Ultralytics selected MuSGD for the planned number of iterations (initial learning rate 0.01, momentum 0.9, weight decay 0.0005); the random seed was 0. Model training and data processing were performed on Dell PowerEdge T640 compute and graphics processing unit servers and a PowerEdge R540 storage server (Dell Inc., Round Rock, TX, USA).

Performance evaluation and statistical analysis

Both models were evaluated on the validation and test sets with the same automated procedure. For each model, the confidence threshold was set to the value that maximised the F1-score on the validation set and was kept fixed for the test set. A prediction was a true positive if its mask matched a ground-truth instance of the same category with an IoU of at least 0.5 (one-to-one matching); unmatched predictions were false positives, and unmatched ground-truth instances, including those predicted as another category, were false negatives. Each label of a dual-labelled tooth was a separate ground-truth instance. Confusion matrices were derived from the same matching; unmatched predictions and

reference labels were then matched regardless of category (mask IoU ≥ 0.5) to record between-category confusions, and those still unmatched were assigned to the background row or column.

The primary outcome was the macro-averaged F1-score, the unweighted mean of the 8 category-level F1-scores. Secondary outcomes were per-category precision, recall and F1-score, the micro-averaged F1-score, mask mAP50 and mAP50–95, computed with the Ultralytics validator from all predictions with a confidence of at least 0.001, and the Dice and Jaccard coefficients of true-positive masks. In a secondary analysis, each pair of overlapping reference labels of different categories was evaluated as a unit and considered correctly predicted when at least one of its 2 categories was predicted (same category, mask IoU ≥ 0.5); the proportion of pairs for which both categories were predicted was also recorded.

Ninety-five percent confidence intervals (CIs) were obtained by image-level bootstrap resampling of the validation and test sets (5000 iterations, percentile method), as patient identifiers were unavailable, and the macro-F1 difference between the models was estimated with a paired bootstrap, in which both models were evaluated on the same resampled images, and reported with its 95% CI. Analyses were performed in Python 3.12.10.

Results

The performance of YOLOv11 and YOLO26 is reported for the test set, on which the primary comparison of macro-averaged F1-scores was based, and for the validation set, which was used for model selection and threshold choice.

Selection of confidence thresholds

On the validation set, the macro-averaged F1-score was highest at a confidence threshold of 0.35 for YOLOv11 and 0.41 for YOLO26; these thresholds were used in all subsequent analyses.

Overall model performance on the validation set

At these thresholds, the macro-averaged F1-score, precision and recall were 0.578, 0.575 and 0.585 for YOLOv11 and 0.572, 0.635 and 0.528 for YOLO26. The mean Dice and Jaccard coefficients of true-positive masks were 0.952 and 0.910 for YOLOv11 and 0.950 and 0.906 for YOLO26. Per-category results and confusion matrices are provided in Supplementary Table S3 and Supplementary Fig. S1.

Model comparison on the test set

On the test set, the macro-averaged F1-score was 0.610 (95% CI 0.571–0.643) for YOLOv11 and 0.595 (95% CI 0.553–0.632) for YOLO26. The difference between the models (YOLO26 minus YOLOv11) was −0.014 (95% CI −0.052 to 0.024); the interval included zero.YOLO26 had higher macro-averaged precision (0.657 vs 0.605) but lower macro-averaged recall (0.555 vs 0.616). Mask mAP50 and mAP50–95 were 0.428 and 0.358 for YOLOv11 and 0.415 and 0.347 for YOLO26.

Per-category performance on the test set

Per-category results are given in Table 4. Extraction had the highest F1-score with both models (0.805 for YOLOv11 and 0.780 for YOLO26). Root canal treatment, restoration and root canal retreatment followed, with F1-scores between 0.65 and 0.69, whereas restoration replacement, pulp capping, pulpotomy and follow-up had the lowest F1-scores (0.43–0.56). Within each category, the differences between the 2 models were small, and their confidence intervals overlapped.

Table 4. Per-category performance of YOLOv11 and YOLO26 on the test set.

Category	n	YOLOv11			YOLO26		
		Precision	Recall	F1 (95% CI)	Precision	Recall	F1 (95% CI)
Pulpotomy	77	0.500	0.519	0.510 (0.407–0.603)	0.514	0.481	0.497 (0.388–0.589)
Extraction	218	0.837	0.775	0.805 (0.762–0.843)	0.857	0.716	0.780 (0.732–0.824)
Restoration	238	0.660	0.702	0.680 (0.630–0.727)	0.714	0.660	0.686 (0.635–0.733)
Restoration replacement	61	0.475	0.459	0.467 (0.344–0.573)	0.606	0.328	0.426 (0.289–0.547)
Follow-up	96	0.509	0.615	0.557 (0.471–0.635)	0.548	0.531	0.540 (0.447–0.621)
Pulp capping	45	0.535	0.511	0.523 (0.387–0.636)	0.615	0.356	0.451 (0.290–0.587)
Root canal retreatment	24	0.640	0.667	0.653 (0.500–0.784)	0.680	0.708	0.694 (0.528–0.826)
Root canal treatment	212	0.689	0.679	0.684 (0.627–0.735)	0.719	0.665	0.691 (0.633–0.744)
Macro average	**971**	**0.605**	**0.616**	**0.610 (0.571–0.643)**	**0.657**	**0.555**	**0.595 (0.553–0.632)**
Micro average	**971**	**0.655**	**0.665**	**0.660 (0.631–0.688)**	**0.702**	**0.613**	**0.655 (0.625–0.682)**

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

Confusion matrix analysis

Confusion matrices are shown in Fig. 2. For both models, most errors were missed reference labels or predictions without a matching reference label rather than confusions between categories. YOLO26 produced fewer predictions without a matching reference label (113 vs 198) but missed more reference labels (237 vs 182) than YOLOv11. When categories were confused, root canal treatment was most often involved, mainly with extraction and, less often, with restoration and pulpotomy. Representative examples with high, partial and low agreement are shown in Supplementary Figs. S2–S4.

Segmentation quality

For true-positive predictions, the mean Dice and Jaccard coefficients were 0.952 (95% CI 0.949–0.955) and 0.910 (95% CI 0.904–0.915), respectively, for both models. Across categories, the Dice coefficient ranged from 0.93 to 0.97 and the Jaccard coefficient from 0.88 to 0.94 (Table 5).

Table 5. Per-category Dice and Jaccard coefficients of true-positive masks for YOLOv11 and YOLO26 on the test set.

Category	YOLOv11		YOLO26	
	Dice	**Jaccard**	**Dice**	**Jaccard**
Pulpotomy	0.958	0.920	0.955	0.915
Extraction	0.935	0.881	0.935	0.879
Restoration	0.956	0.916	0.952	0.911
Restoration replacement	0.953	0.912	0.953	0.912
Follow-up	0.945	0.898	0.945	0.897
Pulp capping	0.962	0.927	0.967	0.936
Root canal retreatment	0.955	0.915	0.956	0.917
Root canal treatment	0.953	0.911	0.954	0.913
Mean (95% CI)	**0.952 (0.949–0.955)**	**0.910 (0.904–0.915)**	**0.952 (0.949–0.955)**	**0.910 (0.904–0.915)**

Mean, unweighted average of the 8 categories; CI, bootstrap confidence interval.

Secondary analysis of dual-labelled teeth

The test set contained 174 dual-label pairs, that is, 2 overlapping reference labels of different categories on the same tooth. YOLOv11 predicted at least one of the 2 categories for 136 pairs (78.2%) and both for 57 (32.8%); the corresponding values for YOLO26 were 123 (70.7%) and 39 (22.4%).

Discussion

In this study, YOLOv11 and YOLO26 were trained to suggest one of 8 management categories from periapical radiographs of children. On the test set, the 2 models reached similar macro-averaged F1-scores (0.610 and 0.595), and the 95% CI of their difference included zero. **Test performance was similar to validation performance, suggesting that the thresholds selected on the validation set transferred to unseen data.** Correctly categorised teeth were segmented with high overlap (mean Dice 0.95), and most errors were missed labels rather than confusions between categories.

Most AI studies on dental radiographs have addressed diagnostic tasks, such as caries detection, periapical lesion identification, marginal bone-loss detection and pulpitis diagnosis [28–31] and, in paediatric dentistry, tooth detection and numbering [3], trauma [13], and developmental anomalies [6,9]. Studies whose output is a treatment category rather than a diagnosis remain scarce [10,32,33]. The closest previous work in children detected furcation lesions in primary molars on panoramic radiographs and mapped them to 2 treatment options through fixed rules [11]; in the present study, 8 management categories were learned directly from expert-labelled periapical radiographs. Because a treatment decision also depends on information that a radiograph cannot provide, such as symptoms, clinical examination findings and the child's cooperation, these categories represent the management that could be considered on the basis of radiographic findings alone. The models should therefore be regarded as a methodological step from diagnostic towards treatment-oriented tasks rather than as a substitute for clinical decision-making.

YOLOv11 and YOLO26 are the 2 most recent architectures developed by Ultralytics [16,34]. YOLOv11 refines the backbone and adds spatial attention to improve feature extraction [16], whereas YOLO26 performs end-to-end inference without non-maximum suppression and removes distribution focal loss [34]. Other recent variants, such as YOLOv12, were not included because they were **developed outside Ultralytics**; comparing YOLOv11 and YOLO26 within one framework and training pipeline allowed **2 successive architectures to be compared under identical training and evaluation conditions.**

Recent dental studies comparing YOLO generations have reported inconsistent results. Sarıoğlu et al. [13] compared YOLOv8, YOLOv11 and YOLOv12 for dental trauma segmentation on paediatric panoramic radiographs and reported multiclass F1-scores of 0.592, 0.653 and 0.683, respectively. Conversely, Mutlu et al. [15], comparing YOLOv11 and YOLOv12 for automated diagnostic charting of 13 dental conditions on panoramic radiographs, found YOLOv11 to be superior, with a mAP50 of 0.857 on the internal test set and 0.806 on an external dataset. These inconsistencies suggest that architecture superiority is task-dependent [35], underscoring the need for architecture-specific validation for each clinical application.

In the present study, the 2 architectures performed similarly: the difference in macro-averaged F1-score on the test set was small, its 95% CI included zero, and the per-category confidence intervals overlapped. The models differed mainly in their operating characteristics. YOLO26 produced fewer predictions without a matching reference label but missed more reference labels, resulting in higher precision and

lower recall. This may reflect its end-to-end one-to-one assignment, which yields fewer redundant predictions [34], as well as its higher selected threshold (0.41 vs 0.35). Successive YOLO generations have likewise yielded only incremental gains in other dental imaging tasks [13–15], suggesting that further improvement for this task is more likely to depend on the data and the reference standard than on the choice between recent architectures.

Performance differed between categories but did not consistently follow category frequency: root canal retreatment, the least frequent category, reached F1-scores similar to those of restoration and root canal treatment, whereas pulpotomy and follow-up, which were more frequent, were among the weakest categories. The weakest categories (restoration replacement, pulp capping, pulpotomy and follow-up) are those whose distinction from neighbouring options depends most on clinical information or on judgement about lesion progression. The confusion matrices showed a consistent pattern: confusions mostly involved root canal treatment and extraction, restoration or pulpotomy, corresponding to decisions on restorability and the extent of pulpal involvement.**Of these,** the confusions of root canal treatment with extraction and with pulpotomy also occurred as dual-label combinations**(n = 93 and 297; Table 2), suggesting that they reflect,** at least in part, the limits of radiograph-only assessment rather than model error alone. By contrast, the confusion with restoration was **rare among dual labels; most of these errors were restorations predicted for teeth labelled for root canal treatment (15 of 21 for YOLOv11 and 12 of 19 for YOLO26), corresponding to** missed periradicular pathology that could lead to under-treatment.

In conventional lesion segmentation, IoU serves as both the matching criterion and the performance target [10,36]. In the present study, the IoU threshold of 0.5 served mainly to establish that a prediction referred to the same tooth region, whereas its correctness depended on the assigned category. Boundary accuracy was assessed separately with Dice and Jaccard coefficients, which were high in all categories (Dice 0.93–0.97). Because these coefficients were computed for true-positive masks, they describe correctly categorised teeth only; nonetheless, they indicate that once the correct category was identified, the predicted masks closely matched the expert annotations, and that the limiting factor was the category decision rather than segmentation.

A dual label was assigned when 2 management categories were equally acceptable on radiographic evidence alone, which applied to 1825 teeth in the dataset. Most dental AI datasets assign a single reference label to each finding, implicitly assuming that one answer can be derived from the image. For treatment categories, this assumption is often not met, because the final choice depends on information that a radiograph cannot provide, such as pulp vitality and symptom history. The frequency of dual labels therefore quantifies, rather than conceals, the limits of radiograph-based treatment suggestions; forcing a single label in these cases would have introduced an arbitrary choice into the reference standard.

Because each dual label was trained as a separate instance with its own mask, the models could predict both categories for the same tooth. In practice, they predicted at least one of the 2 acceptable categories for most dual-labelled teeth, whereas both categories were predicted less often. Under label-

level evaluation, however, the unselected alternative counts as a false negative, which accounted for nearly a quarter of all false negatives (79 of 325 for YOLOv11 and 84 of 376 for YOLO26). Thus, label-level metrics may underestimate the models' ability to identify an acceptable management option for teeth with dual labels. In this respect, the dual-label strategy may represent radiographic ambiguity more faithfully than a single-label approach, which would treat this ambiguity as label noise.

Nevertheless, dual labelling remains an exploratory approach to radiographic ambiguity, not an established standard. It does not provide additional credit in the primary label-level analysis, because each reference label must be predicted separately. Rather, it records that 2 management options are acceptable based on the radiograph, while leaving the final choice to clinical examination. Other approaches, such as soft or ranked labels, may represent this ambiguity better and should be compared in future studies. The present results therefore show that a treatment-suggestion model can be trained on such a reference standard, not that its outputs can be used directly for clinical decision-making.

This study has several limitations. First, the reference standard was based on radiographs alone, without symptoms, clinical examination or pulp vitality testing. The labels therefore represent the management that experts considered on radiographic grounds, not the treatment that was actually performed**. Some decisions cannot be finalised on a radiograph;** for example, the choice between pulpotomy and root canal treatment depends partly on pulpal bleeding and symptoms [20]. This limitation is inherent to radiograph-only assessment and applies equally to human interpretation without clinical information; validation against clinical records and treatment outcomes is needed.

Second, diagnoses were not annotated separately; the categories were assigned directly from the radiographic appearance according to the criteria in Table 1. The models therefore learn a direct image-to-management mapping without an explicit diagnostic step, which limits interpretability.

Third, the dataset was split at the image level. **Although no radiograph appeared in more than one subset, different radiographs of the same patient may have been assigned to different subsets, so a slight optimistic bias in the performance estimates cannot be excluded.** Patient age and sex were not available for each image, and performance was not analysed separately for primary and permanent teeth or for tooth types. In addition, radiographs of traumatised teeth were excluded, because their management falls outside the 8 categories studied; the models therefore do not apply to dental trauma.

Fourth, the data came from a single centre and a single imaging system, and some categories had few test labels, such as root canal retreatment (n = 24) and pulp capping (n = 45), which resulted in wide confidence intervals. A single confidence threshold was applied to all categories; category-specific thresholds might improve the balance between precision and recall for less frequent categories. External, multicentre validation is required before any clinical application.

Future studies could combine the radiographic model with a diagnostic module, such as a caries-depth or periapical index classifier, in a hierarchical framework that mirrors clinical reasoning and improves interpretability. Incorporating clinical variables such as age, pain characteristics and pulp vitality is

another logical step, given that combining complementary data sources has improved performance in other dental applications [12]. The present work provides a methodological basis for such extensions.

Conclusion

YOLOv11 and YOLO26 learned expert-defined management categories from paediatric periapical radiographs with similar performance, and no clear difference was found between the 2 architectures. Performance was lowest for categories whose distinction depends on clinical information, and dual labelling offered a transparent way to represent teeth for which 2 management options were acceptable on the radiograph. As the models rely on radiographic findings alone, they are not intended to replace clinical decision-making but represent a methodological step from diagnostic towards treatment-oriented tasks. Integration of clinical data, external multicentre validation and comparison with actual treatment decisions are needed before any clinical use.

Declarations

Competing interests

Dr. İbrahim Şevki Bayrakdar is the co-founder and CEO of CranioCatch, a company that develops artificial intelligence solutions for dentistry, where he oversees company operations, strategic planning, and product development. Given this role and his business interest in CranioCatch, a potential conflict of interest relating to the content of this article exists; while every effort was made to ensure an objective and balanced perspective, readers are advised to consider this potential influence. H.A., B.K. and M.B.S. declare no competing interests.

Ethics approval

This retrospective study was approved by the Non-Interventional Clinical Research Publication Ethics Committee of Nevşehir Hacı Bektaş Veli University (decision no. 2025.08.01, 31 July 2025). Owing to the retrospective design and the complete anonymisation of all data, the ethics committee waived the requirement for informed consent, including informed consent from the participants' parents or legal guardians. All methods were carried out in accordance with relevant guidelines and regulations and with the 1964 Declaration of Helsinki and its later amendments.

Funding

This study was funded by the Health Institutes of Türkiye (Türkiye Sağlık Enstitüleri Başkanlığı, TÜSEB; project number 48468, to H.A. and B.K.). The funder had no role in the study design, data collection, analysis, interpretation of the results or writing of the manuscript.

Author Contribution

H.A. contributed to the conception and design, methodology, annotation review and validation, interpretation, provision of resources, and project administration; drafted and critically revised the manuscript.B.K. contributed to the conception and design, methodology, data curation and image annotation, analysis, interpretation, provision of resources, and project administration; critically revised the manuscript.M.B.S. contributed to annotation review and validation, and interpretation; drafted and critically revised the manuscript.İ.Ş.B. contributed to the supervision, confirmation of the reference standard, analysis, and interpretation; critically revised the manuscript.All authors gave final approval and agreed to be accountable for all aspects of the work.

Acknowledgement

Preliminary results of this study, based on a smaller dataset, were presented orally at the 2nd Scientific Congress of the Pediatric Dentists Association (Çocuk Diş Hekimleri Derneği), Istanbul, Türkiye, May 8–10, 2026, where the presentation received the first prize in the oral presentation category, and were published in the congress abstract book (Contemp Pediatr Dent. 2026;7(suppl 1):28-30. doi:10.51463/cpd.2026.S1).

Data Availability

The radiographs analysed during the current study are not publicly available owing to patient privacy restrictions but are available from the corresponding author on reasonable request. The model predictions for the test set and the code are also available from the corresponding author on reasonable request.

References

1. Mahajan, K. et al. Artificial intelligence in pediatric dentistry-A systematic review. *J. Dent. Res. Reviews*. **10**, 7–12 (2023).
2. Karamüftüoğlu, N. et al. Artificial Intelligence in Pediatric Dentistry: A Systematic Review and Meta-Analysis. *Children* **13**, 152 (2026).
3. Ozudogru, S. et al. Artificial intelligence system for automatic tooth detection and numbering in the mixed dentition in CBCT. *Eur J. Paediatr. Dent* **26**, (2025).
4. Ozsunkar, P. S. et al. Detecting white spot lesions on post-orthodontic oral photographs using deep learning based on the YOLOv5x algorithm: a pilot study. *BMC Oral Health*. **24**, 490 (2024).
5. Alevizakos, V., Bekes, K., Steffen, R. & von See, C. Artificial intelligence system for training diagnosis and differentiation with molar incisor hypomineralization (MIH) and similar pathologies. *Clin. Oral Investig*. **26**, 6917–6923 (2022).

6. Gülşen, E. et al. Deep learning detection of ectopic canines and molars in mixed dentition. *Sci. Rep.* https://doi.org/10.1038/s41598-026-45912-4 (2026).
7. Bunyarit, S. S. et al. Dental age estimation of Malaysian Chinese children and adolescents: Chaillet and Demirjian's method revisited using artificial multilayer perceptron neural network. *Australian J. Forensic Sci.* **52**, 681–698 (2020).
8. Park, Y. H., Kim, S. H. & Choi, Y. Y. Prediction models of early childhood caries based on machine learning algorithms. *Int. J. Environ. Res. Public. Health*. **18**, 8613 (2021).
9. Ha, E. G., Jeon, K. J., Kim, Y. H., Kim, J. Y. & Han, S. S. Automatic detection of mesiodens on panoramic radiographs using artificial intelligence. *Sci. Rep.* **11**, 23061 (2021).
10. Alabdulkareem, M., Atieh, M., AbuMostafa, A., Aldalaan, K. & Alturki, N. Artificial Intelligence in Dental Treatment Planning and Diagnostic Decision-Making: A Systematic Review and Meta-Analysis. *Clin. Exp. Dent. Res.* **12**, e70343 (2026).
11. Karamüftüoğlu, N., Bulut, A., Akın, M. & Sağıroğlu, Ş. Panoramic radiograph-based deep learning models for diagnosis and clinical decision support of furcation lesions in primary molars. *Children* **12**, 1517 (2025).
12. Ahmad, S. I. et al. A Novel Multimodal Deep Image Analysis Model for Predicting Extraction/Non-Extraction Decision. *Orthod. Craniofac. Res.* **29**, 101–111 (2026).
13. Sarıoğlu, D. et al. Comparative Performance of YOLO Models in the Automated Segmentation of Dental Traumas on Panoramic Radiographs. *Dent. Traumatol.* **42**, 363–376 (2026).
14. Saber, S., Nasr, A. E., Torky, H., Saif, N. & A. A. & Automated assessment of periapical health based on the radiographic periapical index using YOLOv8, YOLOv11, and YOLOv12 one-stage object detection algorithms. *Sci. Rep.* **15**, 36487 (2025).
15. Mutlu, O., Aslan, E. & Mert, A. Deep Learning–Based Automated Diagnostic Charting on Panoramic Radiography: Comparison of YOLOv11 and YOLOv12. *Odontology*1–11 (2026). 10.1007/s10266-026-01333-3
16. Khanam, R. & Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725* (2024). https://doi.org/10.48550/arXiv.2410.17725 doi:10.48550/arXiv.2410.17725.
17. Hidayatullah, P. & Tubagus, R. YOLO26: A Comprehensive Architecture Overview and Key Improvements. *arXiv preprint arXiv:2602.14582* (2026). https://doi.org/10.48550/arXiv.2602.14582 doi:10.48550/arXiv.2602.14582.
18. Tejani, A. S. et al. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. *Radiol. Artif. Intell.* **6**, e240300 (2024).
19. Sounderajah, V. et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nat. Med.* **31**, 3283–3289 (2025).
20. American Academy of Pediatric Dentistry. Pulp therapy for primary and immature permanent teeth. in The Reference Manual of Pediatric Dentistry 399–407 (American Academy of Pediatric Dentistry, Chicago, IL, (2021).

21. American Academy of Pediatric Dentistry. Pediatric restorative dentistry. in The Reference Manual of Pediatric Dentistryvol. 371–383 (American Academy of Pediatric Dentistry, Chicago, IL, 2020). (2021).
22. American Academy of Pediatric Dentistry. Caries-risk assessment and management for infants, children, and adolescents. in The Reference Manual of Pediatric Dentistry 66–72 (American Academy of Pediatric Dentistry, Chicago, IL, (2020).
23. Dhar, V. et al. Consensus-Based Recommendations on Pulp Therapies in Primary and Permanent Teeth: IAPD Porto Forum. *Int. J. Paediatr. Dent.* https://doi.org/10.1111/ipd.70068 (2026).
24. Tinanoff, N. et al. Principles and care pathways for caries management in children: IAPD Rome forum. *Int. J. Paediatr. Dent.* **34**, 692–699 (2024).
25. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* 159–174. 10.2307/2529310 (1977).
26. Johnson, J. M. & Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *J. Big Data*. **6**, 27 (2019).
27. Paszke, A. et al. Pytorch: An imperative style, high-performance deep learning library. *Adv Neural Inf. Process. Syst* **32**, (2019).
28. Azhari, A. A., Helal, N., Sabri, L. M. & Abduljawad, A. Artificial intelligence (AI) in restorative dentistry: Performance of AI models designed for detection of interproximal carious lesions on primary and permanent dentition. *Digit. Health*. **9**, 20552076231216681 (2023).
29. Alaqla, A., Khanagar, S. B., Albelaihi, A. I., Singh, O. G. & Alfadley, A. Application and performance of artificial intelligence-based models in the detection, segmentation and classification of periapical lesions: a systematic review. *Front. Dent. Med.* **6**, 1717343 (2025).
30. Ibraheem, W. I. et al. Assessment of the diagnostic accuracy of artificial intelligence software in identifying common periodontal and restorative dental conditions (marginal bone loss, periapical lesion, crown, restoration, dental caries) in intraoral periapical radiographs. *Diagnostics* **15**, 1432 (2025).
31. Zheng, L. et al. Artificial intelligence in digital cariology: a new tool for the diagnosis of deep caries and pulpitis using convolutional neural networks. *Ann. Transl Med.* **9**, 763 (2021).
32. Dua, B. et al. Redefining oral healthcare through artificial intelligence: a review of current applications and a roadmap for the future of dentistry. *BMC Artif. Intell.* **1**, 13 (2025).
33. Tyagi, M. et al. Artificial intelligence tools in dentistry: A systematic review on their application and outcomes. *Cureus* **17**, (2025).
34. Sapkota, R., Karkee, M. & Ultralytics, Y. O. L. O. evolution: An overview of YOLO26, YOLO11, YOLOv8 and YOLOv5 object detectors for computer vision and pattern recognition. *arXiv preprint arXiv:2510.09653* (2025). https://doi.org/10.48550/arXiv.2510.09653 doi:10.48550/arXiv.2510.09653.
35. Sapkota, R. et al. YOLO advances to its genesis: A decadal and comprehensive review of the You Only Look Once (YOLO) series. *Artif. Intell. Rev.* **58**, 274 (2025).

36. Çelik, B., Savaştaer, E. F., Kaya, H. I. & Çelik, M. E. The role of deep learning for periapical lesion detection on panoramic radiographs. *Dentomaxillofacial Radiol.* **52**, 20230118 (2023).

Figures

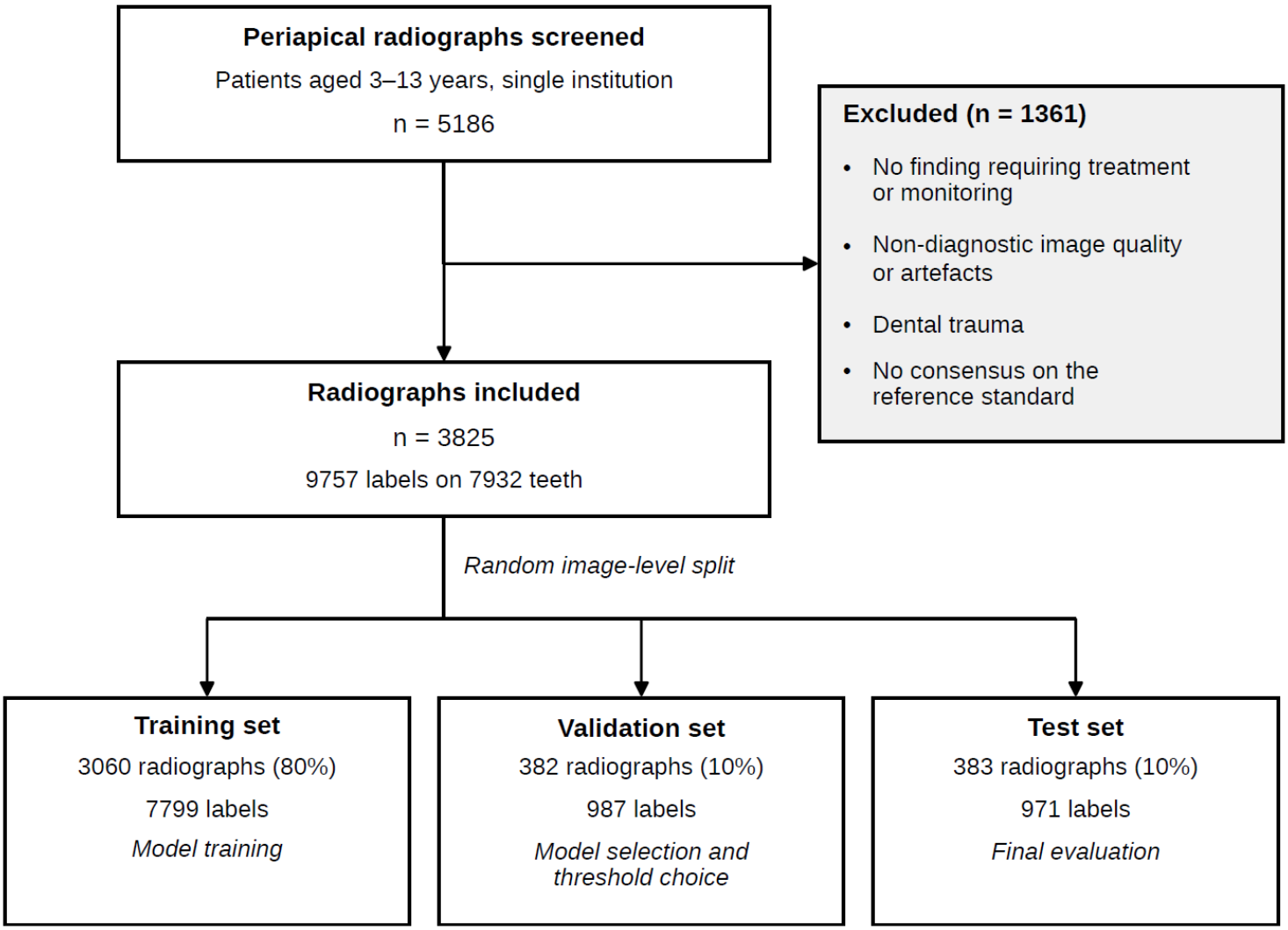

Figure 1

Flow diagram of radiograph selection and data partitioning.

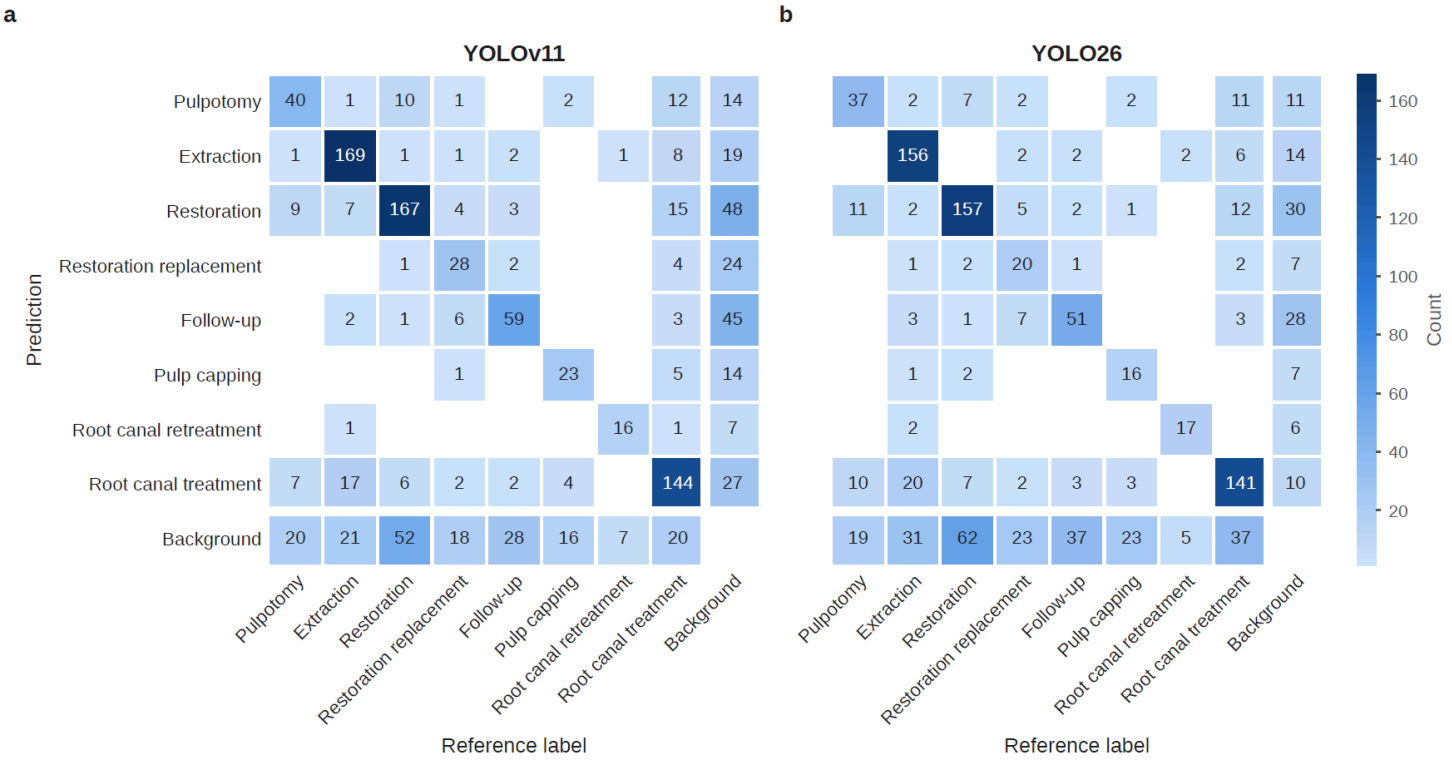

Figure 2

Confusion matrices of YOLOv11 (a) and YOLO26 (b) on the test 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.

Supplementary Files

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

- SupplementaryInformation.pdf