scientific reports

Check for updates

OPEN

Comparative performance of one-stage and two-stage deep learning models for instance segmentation of overhanging dental restorations on bitewing radiographs

Ömer Hatipoğlu[1]✉, Özge Başar[2], Güldane Mağat[3], Ali Altındağ[3] & Fatma Pertek Hatipoğlu[4]

Accurate detection of overhanging dental restorations on bitewing radiographs is clinically important but remains challenging due to subtle marginal discrepancies. This study aimed to develop and compare deep learning–based instance segmentation models for the automated detection and classification of overhanging restorations, with a specific focus on the relative performances of one-stage and two-stage architectures. A total of 1236 anonymized bitewing radiographs were retrospectively collected and manually annotated using polygon-based segmentation by specialist dentists. The restorations were classified as ideal or overhanging based on established radiographic criteria. One-stage YOLO11-based segmentation models (YOLO11n-, YOLO11s-, and YOLO11m-seg) were compared with two-stage architectures implemented in Detectron2, including the Mask R-CNN, Cascade Mask R-CNN, and PointRend. All models were trained and evaluated using identical train–validation–test splits. Performance was assessed on an independent test set using COCO-style instance segmentation metrics (AP50–95, AP50, AP75, and AR@100), reported both overall and for the clinically relevant overhang class. Uncertainty was quantified using non-parametric bootstrap resampling (B = 200) to derive the 95% confidence intervals. All models achieved high AP50 values, indicating effective coarse localization of the restorations. However, two-stage architectures showed consistently higher AP-based point estimates than the one-stage YOLO-based models at stricter overlap thresholds. Cascade Mask R-CNN achieved the highest overall performance (AP50–95 = 0.703), while PointRend yielded the highest overhang-specific AP-based segmentation performance (Overhang AP50–95 = 0.743). YOLO-based models demonstrated lower AP50–95 values despite relatively strong AP50 performance, reflecting a reduced boundary precision. Increasing the YOLO model capacity did not improve performance. Two-stage instance segmentation models, particularly those incorporating boundary-aware refinement, showed higher AP-based performance than one-stage approaches for detecting overhanging dental restorations on bitewing radiographs in this dataset. These findings highlight the importance of architectural design over model scale alone and suggest that two-stage frameworks may offer advantages for precise evaluation of restoration margins in a controlled experimental setting. However, external validation across centers and imaging systems is required before their suitability for routine clinical decision support applications can be established.

Keywords Artificial intelligence, Bitewing radiography, Deep learning, Dental restorations, Instance segmentation, Overhanging restorations

Accurate assessment of the marginal integrity of dental restorations is essential for maintaining periodontal health and ensuring long-term restorative success. Overhanging restorations, defined as marginal excess extending

[1]Department of Restorative Dentistry, Recep Tayyip Erdoğan University, Rize, Turkey. [2]Department of Endodontics, Faculty of Dentistry, Adnan Menderes University, Aydın, Turkey. [3]Department of Oral and Maxillofacial Radiology, Necmettin Erbakan University, Konya, Turkey. [4]Department of Endodontics, Recep Tayyip Erdoğan University, Rize, Turkey. ✉email: omerhtp@gmail.com

beyond the natural tooth contour, are associated with increased plaque accumulation, gingival inflammation, and attachment loss, ultimately predisposing patients to periodontal disease and secondary caries[1]. Despite their clinical relevance, overhanging restorations may remain undetected during routine clinical examinations, particularly when marginal discrepancies are subtle or located interproximally[2].

Bitewing radiography is the most widely used imaging modality for evaluating proximal tooth surfaces and restorative margins. Owing to its ability to visualize interproximal areas with relatively low radiation dose, bitewing radiographs are routinely employed for the detection of proximal caries, assessment of alveolar bone levels, and evaluation of restoration quality[3]. Nevertheless, the radiographic identification of overhanging restorations is inherently challenging and subject to inter-observer variability, as detection depends on image quality, projection geometry, and clinician experience. These limitations underscore the need for objective and reproducible tools that can assist clinicians in identifying marginal discrepancies on bitewing radiographs[4].

Recent advances in artificial intelligence (AI), particularly deep learning–based computer vision methods, have demonstrated promising performance in dental image analysis, including caries detection, periodontal bone loss assessment, and restoration evaluation[5,6]. Instance segmentation approaches are particularly relevant for restorative diagnostics because they enable the localization and classification of restorations and pixel-level delineation of their contours[7,8]. Such detailed geometric information is critical for detecting overhanging margins, which are often characterized by small protrusions or irregularities along restoration boundaries[9].

Deep learning–based instance segmentation models can be broadly categorized into one-stage and two-stage architectures[10]. One-stage models perform localization, classification, and mask prediction simultaneously in a single forward pass, offering high computational efficiency and faster inference. In contrast, two-stage models explicitly separate the generation of candidate regions from subsequent region-based classification and mask refinement, often achieving superior boundary accuracy at the cost of increased computational complexity[11]. Although both approaches have been extensively evaluated in general computer vision tasks, their relative performance in clinically relevant dental applications, particularly for subtle marginal defects such as overhanging restorations, remains insufficiently explored.

To date, most AI-based studies in dental radiology have focused on detection or classification tasks, with comparatively fewer investigations addressing instance-level segmentation of restorations[12]. Moreover, direct and systematic comparisons between one-stage and two-stage instance segmentation models for the detection of overhanging restorations on bitewing radiographs are lacking[9,13]. Such comparisons are essential to determine whether the improved boundary precision of two-stage architectures translates into clinically meaningful advantages over faster one-stage approaches in this specific diagnostic context.

To date, deep learning studies related to restorative findings on dental radiographs have shown promising results; however, the existing literature remains methodologically fragmented. In particular, previous studies on overhanging restorations mainly evaluated YOLO-based systems and demonstrated that artificial intelligence can detect these lesions with relatively high performance, but they did not provide a direct architecture-level comparison between one-stage and two-stage instance segmentation frameworks[9,13]. Likewise, broader studies on dental restorations have confirmed the feasibility of deep learning for restoration-related detection and segmentation; however, these studies were mostly conducted on panoramic radiographs or focused on general restorative findings rather than subtle proximal marginal defects on bitewing radiographs[14]. In addition, recent multiclass bitewing segmentation research has shown that overhanging fillings remain a relatively difficult category in more complex multi-label settings, suggesting that dedicated model development for this specific finding is still needed[15].

Beyond the dental field, the distinction between one-stage and two-stage segmentation strategies is clinically relevant for tasks requiring precise boundary delineation. In general, one-stage models are computationally efficient, whereas two-stage approaches often provide more accurate localization and mask refinement, especially when object boundaries are subtle or clinically important[11,16]. This distinction is particularly important in the detection of overhanging restorations, where diagnosis depends not only on identifying the presence of a restoration, but also on accurately delineating small marginal extensions. Therefore, the main novelty of the present study is that it does not merely apply deep learning to overhang detection, but systematically compares one-stage and two-stage instance segmentation architectures under identical experimental conditions to determine which design is more suitable for this boundary-sensitive diagnostic task.

Therefore, this study aimed to develop and evaluate deep learning–based instance segmentation models for the automated detection and classification of overhanging dental restorations on bitewing radiographs. Specifically, we compared the performance of a one-stage segmentation framework (YOLO-based models) with multiple two-stage architectures (Mask R-CNN, Cascade Mask R-CNN, and PointRend) using the same dataset split, standardized evaluation, and uncertainty estimation procedures. The null hypothesis (H_0) of this study was that one-stage and two-stage deep learning architectures would demonstrate comparable descriptive instance segmentation performance for the detection and classification of overhanging restorations on bitewing radiographs, both in overall metrics and overhang-specific performance.

Methodology

Study design

This retrospective, single-center, image-based observational study was conducted in accordance with the guidelines proposed by Schwendicke, et al.[17] for the appropriate development and application of AI in dental research. In addition, this study adhered to the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) guidelines[18] and followed the reporting standards of the Standards for Reporting of Diagnostic Accuracy Studies (STARD) checklist[19] to ensure methodological transparency, reproducibility, and robust reporting of the diagnostic performance. Bitewing radiographs were retrospectively obtained from the archives of the Department of Oral and Maxillofacial Radiology, Necmettin Erbakan University. Ethical approval for the

retrospective use of anonymized radiographic data was obtained from the Necmettin Erbakan University Ethics Committee (Approval No:2025/542).

Dataset

The bitewing radiographs were originally acquired for routine diagnostic purposes, including the detection of approximal caries, evaluation of alveolar bone levels, and assessment of existing dental restorations. Only high-quality bitewing radiographs were included in the dataset. Images presenting technical or positioning errors, such as incorrect wing biting, film bending, cone-cut artifacts, improper positioning or angulation, and motion-related distortions, were systematically excluded to ensure diagnostic reliability and image consistency.

All included bitewing images depicted both dental arches. For the purpose of analysis, only restorations located on the crowns of clearly visible teeth were considered, while restorations on partially visualized or non-assessable teeth were disregarded.

During retrospective data extraction, temporary patient identifiers were used before anonymization to perform data partitioning in a patient-level non-overlapping manner. These identifiers were used only to group radiographs belonging to the same patient, identify duplicate or repeated radiographs, and ensure that all images from a given patient were assigned exclusively to either the training, validation, or test subset. Thus, the split was performed at the patient level before anonymization, and an initial image-level split followed by post hoc correction was not performed. After this verification and splitting step, all patient identifiers were permanently removed, and only anonymized image-level records were retained for model development and evaluation.

A total of 1,236 anonymized bitewing radiographs were collected from 2014 to 2025. All images were acquired using the Planmeca ProMax 2D intraoral digital imaging system (Planmeca, Helsinki, Finland). The imaging protocol followed the manufacturer's recommendations, employing two size-2 phosphor plates positioned bilaterally with standardized exposure parameters of 68 kVp, 16 mA, and 13 s. Subsequently, the bitewing radiographs were digitized and stored in Joint Photographic Experts Group (JPEG) format using a Soredex Digora Optime image scanner system (Finland).

Ground truth annotation

Ground-truth annotations were generated using anonymized dental bitewing radiographs of the patients. All images were evaluated and annotated by three dental specialists: two oral and maxillofacial radiology specialists (G. M., A. A.), each with 16 years of clinical experience, and one endodontist with five years of clinical experience (Ö.B.). Polygon-based annotations were created using the web-based annotation platform makesense.ai (https://www.makesense.ai), which allows precise freehand delineation of regions of interest on radiographic images.

During the annotation process, the initial polygonal segmentation of all visible direct restorations on bitewing radiographs was performed by an endodontist (Ö.B). Each restoration was manually outlined using polygonal boundaries and classified into one of two categories based on marginal adaptation: ideal or overhanging restorations. Overhanging restorations were identified according to established clinical criteria, whereby a restoration was considered overhanging if its marginal extension exceeded the proximal tooth contour by approximately 0.5 mm or more[20]. This approximately 0.5-mm threshold was applied as an expert consensus-based radiographic criterion rather than as a formally calibrated metric measurement. No calibration phantom, pixel-spacing correction, magnification adjustment, or known anatomical reference-based scaling was used during annotation. Instead, restorations were classified as overhanging when the annotators identified a visible marginal extension beyond the proximal tooth contour on the bitewing radiograph, consistent with the established radiographic criterion. In borderline or ambiguous cases, radiographic contrast, projection geometry, and the relationship between the restoration margin and adjacent tooth contour were considered during joint specialist review, and the final label was determined by consensus. Restorations without detectable marginal excess were classified as ideal restorations.

Following the initial annotation, both oral and maxillofacial radiology specialists (G. M. and A. A.) independently reviewed all polygon boundaries and class assignments. Any discrepancies were discussed and resolved through joint consensus among the three annotators to ensure complete agreement on both the localization and classification of each restoration type. The finalized polygon coordinates and corresponding binary class labels (ideal vs. overhanging) were subsequently exported in the JavaScript Object Notation (JSON) format and used as the reference standard for model development and evaluation.

Model architectures

YOLO-based one-stage segmentation (YOLO11-seg)

YOLO11-seg (Ultralytics framework) was selected as a one-stage, end-to-end deep learning architecture for binary instance segmentation of dental restorations on bitewing radiographs. In one-stage designs, candidate instances are generated and refined in a single forward pass without an explicit region proposal stage, enabling computationally efficient inference while maintaining a competitive performance. Conceptually, the model comprises (i) a convolutional feature extraction backbone, (ii) a multi-scale feature aggregation module that fuses low- and high-level representations to improve sensitivity to objects of varying sizes, and (iii) prediction heads that jointly output the instance localization and class assignment. This multiscale formulation is particularly relevant for bitewing radiographs, where restorations may present with variable extents and subtle marginal irregularities that require both local details and broader contextual cues.

For each predicted instance, YOLO11-seg produces a set of outputs that are directly applicable to the study aims: a localization component (bounding box), a classification component (ideal vs. overhang), and an instance-specific segmentation mask delineating the restoration boundary. The segmentation branch extends the detection pipeline by generating pixel-level delineations for each detected restoration, thereby allowing the model to capture clinically important morphological characteristics, such as marginal excess and contour

irregularities. Modern YOLO-seg designs achieve this efficiently by coupling instance predictions with a lightweight mask generation mechanism, enabling the production of per-instance masks without the heavy computational overhead typically associated with two-stage segmentation pipelines. YOLO11-seg supports the simultaneous execution of (a) restoration localization, (b) binary categorization of marginal status, and (c) fine-grained segmentation of restoration geometry within a unified architecture.

To investigate the influence of model capacity on segmentation and classification performance, multiple YOLO11-seg variants were evaluated (YOLO11n-seg, YOLO11s-seg, and YOLO11m-seg). These variants share the same overall architectural design and output structure but differ in parameter count and representational capacity, allowing for a systematic comparison under an identical task definition (binary instance segmentation: ideal vs. overhang). The training protocol, data split strategy, and evaluation procedures for the YOLO-based models are described in the following sections.

Detectron2-based two-stage instance segmentation

In addition to the one-stage YOLO-based models, two-stage instance segmentation architectures implemented in the Detectron2 framework (Meta AI Research) were evaluated for the same binary task (ideal vs. overhang) as the one-stage models. Two-stage designs explicitly separate the generation of candidate regions from subsequent classification and mask refinement steps. This formulation is commonly preferred in medical imaging tasks, where subtle morphological differences and boundary precision are critical, because it enables more targeted feature extraction at the region level and often yields more accurate localization and segmentation of fine structures.

The core two-stage baseline in this study was Mask R-CNN, which followed the Generalized R-CNN paradigm. First, a convolutional backbone extracts hierarchical feature maps from the input image, and a Feature Pyramid Network (FPN) aggregates multi-scale features to enhance the detection of objects with varying sizes and appearances. Next, a Region Proposal Network (RPN) generates candidate regions that are likely to contain restorations. These proposals are then processed using ROIAlign to extract fixed-size spatially aligned feature representations from the multi-scale feature maps. Finally, separate task-specific heads perform (i) classification of each instance into ideal or overhang, (ii) bounding box regression to refine localization, and (iii) pixel-level mask prediction to delineate restoration contours. This explicit region-focused processing is particularly relevant for bitewing radiographs, where the marginal excess can be subtle and depends on the precise contour of the restoration relative to the adjacent tooth structures.

To further explore architectures tailored to improved localization and boundary quality, the Cascade Mask R-CNN and PointRend were also included. Cascade Mask R-CNN extends the two-stage approach by applying a sequence of detection heads trained at progressively stricter IoU thresholds, enabling iterative refinement of bounding boxes and improving robustness against localization errors that can adversely affect mask quality. In contrast, PointRend enhances segmentation by refining the mask boundaries using an iterative point-based rendering strategy. Instead of predicting masks solely on a coarse grid, PointRend adaptively selects uncertain points, often concentrated near the edges, and predicts their labels using fine-grained features, resulting in sharper and more accurate boundaries. This boundary-aware refinement is particularly advantageous for the present task, as overhanging restorations are often characterized by small protrusions or irregular marginal extensions that can be missed by coarse-mask predictions.

Training

The training and inference procedures were implemented in Python using a PyTorch-based deep learning environment. All experiments were performed in Google Colab on a GPU-enabled runtime, specifically utilizing an NVIDIA L4 GPU to ensure adequate computational efficiency for model training and evaluation. All models were trained and evaluated using the same train–validation–test split and the same evaluation protocol to support reproducibility and standardized comparison. However, because the models were implemented in different frameworks and represented different architectural families, architecture- and framework-specific training configurations were used, including differences in input resolution, number of epochs, fine-tuning strategy, batch-size selection, and default augmentation pipelines.

The dataset was partitioned into three subsets using a fixed random seed: 80% for training (988 images), 10% for validation (123 images), and 10% for testing (125 images). The same split was consistently used for all models to enable direct performance comparisons under identical data conditions (Fig. 1). In terms of annotated instances, the dataset comprised 3936 annotated direct restoration instances across all subsets. Of these, 3194 annotations were included in the training set (1755 ideal and 1439 overhanging restorations), 352 annotations in the validation set (190 ideal and 162 overhanging), and 390 annotations in the test set (211 ideal and 179 overhanging).

No task-specific or manually designed data-augmentation strategies were implemented in this study. Model training relied exclusively on the default on-the-fly augmentation pipelines provided by the respective deep-learning frameworks. For the YOLO-based models, the augmentation parameters were left unchanged within the Ultralytics training environment, allowing the standard augmentation behavior of the framework to be applied automatically during training. Similarly, Detectron2-based models were trained using the default data loading and preprocessing pipeline defined by the model zoo configurations without introducing any custom augmentation policies.

YOLO11-seg models

All YOLO-based instance segmentation experiments were conducted using the Ultralytics framework. Ground-truth annotations were converted from the COCO format to the YOLO segmentation format, in which each restoration instance was represented by a class label followed by a sequence of normalized polygon coordinates.

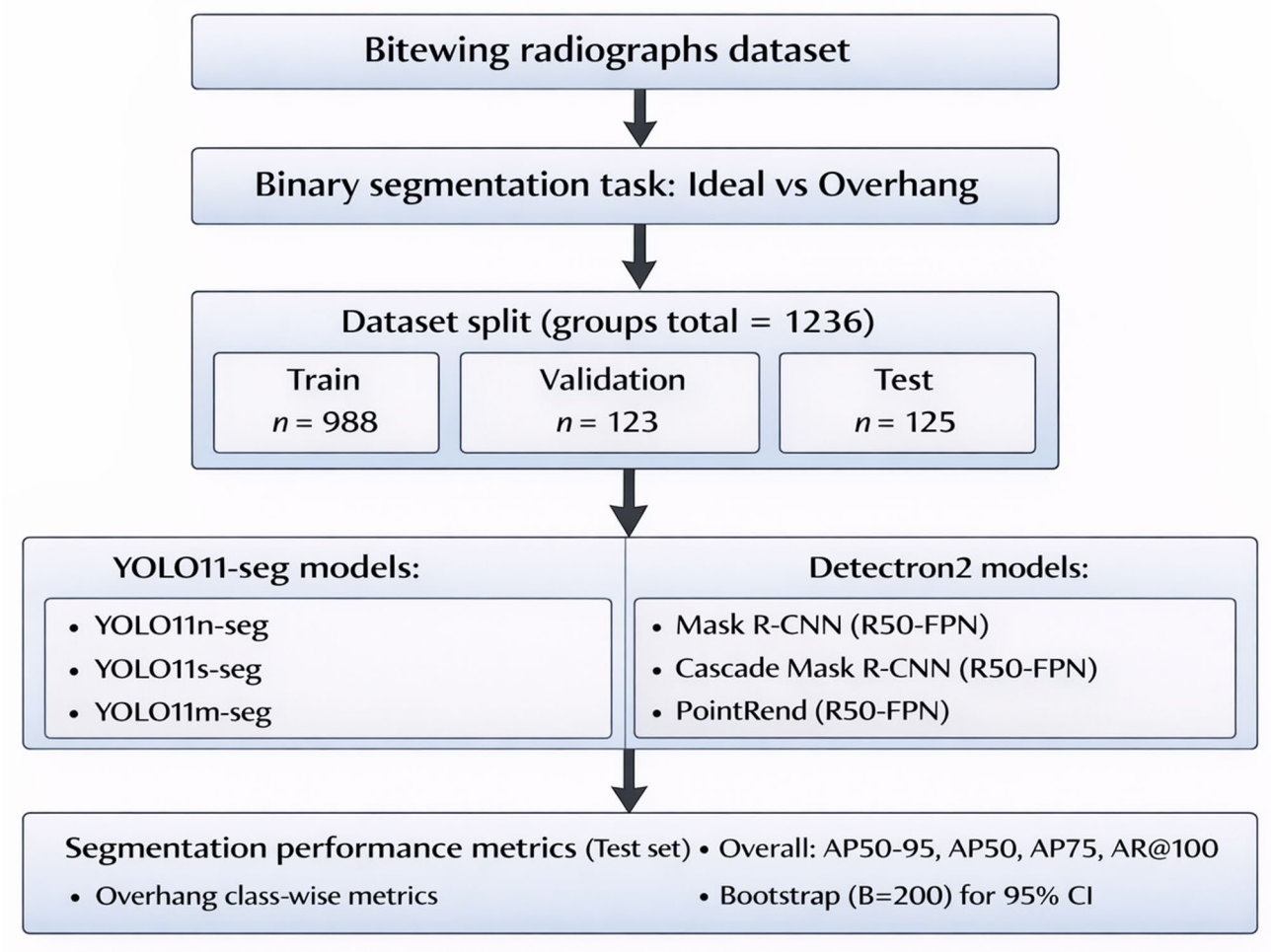

Fig. 1. Study workflow and experimental design for binary instance segmentation of overhanging restorations on bitewing radiographs.

The task was formulated as a binary instance segmentation with two classes (ideal and overhang). All YOLO11-seg models shared the same dataset definition, label structure, and train–validation–test split, allowing standardized comparison within the YOLO-based one-stage model family.

The dataset was partitioned into training, validation, and test subsets using a fixed random seed to ensure reproducibility. A single dataset configuration file was used to define the image and label directories and to consistently register class names across experiments. All YOLO11-seg variants were initialized with pre-trained weights and fine-tuned on the study dataset. Unless otherwise specified, the default Ultralytics optimization strategies and data augmentation pipelines were employed. Model checkpointing was performed automatically by the framework, and the best-performing weights on the validation set were retained for the final evaluation of the independent test dataset.

YOLO11n-seg YOLO11n-seg, the lightweight variant of the YOLO11 family, was trained end-to-end to establish a fast and computationally efficient baseline. The training was conducted for 100 epochs using an input resolution of 960 pixels and a batch size of 8. Owing to its reduced parameter count, no layer-freezing strategy was applied, and all layers were updated throughout the training.

YOLO11s-seg YOLO11s-seg represents an intermediate-capacity architecture with increased representational power compared to YOLO11n-seg. A two-phase fine-tuning strategy was adopted to stabilize training and mitigate overfitting on the medical imaging dataset. In the initial phase, the first 10 layers were frozen, and the model was trained for 30 epochs at an input resolution of 640 pixels. Subsequently, training was resumed from the last checkpoint, with all layers unfrozen, and continued for a total of 150 epochs. The batch size was automatically adjusted by the training engine based on the available GPU memory, and early stopping was enabled using a patience-based criterion.

YOLO11m-seg YOLO11m-seg, the highest-capacity model evaluated in this study, was trained using the same two-phase strategy as YOLO11s-seg. The initial phase consisted of 30 epochs, with the first 10 layers frozen, followed by full fine-tuning with all layers unfrozen. Owing to its larger parameter count, YOLO11m-seg was

trained for a longer total duration of 180 epochs to allow for adequate convergence. As with other YOLO variants, pretrained weights were used for initialization, the batch size was automatically selected, and early stopping was applied to prevent overfitting.

Detectron2-based two-stage instance segmentation
Two-stage instance segmentation models were trained using the Detectron2 framework (Meta AI Research). All annotations were prepared in the COCO format, and the dataset was registered in Detectron2 using separate training, validation, and test splits. Because the image file paths were stored as absolute paths within the COCO JSON files, the image root directory was specified as empty during the dataset registration. To ensure consistency across experiments, the class names were ordered according to their COCO category-id values and explicitly defined as ideal and overhang within the Detectron2 metadata catalog.

Three different two-stage architectures were evaluated: Mask R-CNN (ResNet-50 with Feature Pyramid Network), Cascade Mask R-CNN (ResNet-50-FPN), and PointRend (ResNet-50-FPN–based). All models were initialized with weights pre-trained on the COCO dataset and subsequently fine-tuned for binary instance segmentation. The final classification layers of all architectures were modified to accommodate two classes (ideal vs. overhang).

Mask R-CNN (R50-FPN) For the Mask R-CNN, the standard Detectron2 model zoo configuration based on a ResNet-50 backbone with a Feature Pyramid Network was used. The number of classes in the region of interest (ROI) heads was set to two. Training was performed using a mini-batch size of two images with an initial learning rate of 0.00025. The total training duration corresponded to 50 epochs, with the maximum number of iterations calculated based on the number of training images and the batch size (24,648 iterations). Learning rate step scheduling was disabled, and a warm-up phase was applied at the beginning of the training. The input images were resized to a maximum dimension of 1024 pixels for both training and inference. The model performance was monitored using the validation set, and the final evaluation was conducted exclusively on the independent test set using the COCO evaluation protocol for instance segmentation.

Cascade Mask R-CNN (R50-FPN) The Cascade Mask R-CNN architecture was trained using a ResNet-50-FPN backbone with a cascaded sequence of detection heads designed to progressively refine the localization quality. As with Mask R-CNN, the number of output classes was set to two. Training was performed with a mini-batch size of two images and an initial learning rate of 0.00025. The total training duration was extended to 80 epochs to allow for sufficient convergence across the cascade stages. The maximum number of iterations was computed in an epoch-based manner, and a learning rate decay strategy was applied at predefined proportions of the total training iteration (39,518 iterations). A warm-up phase was included at the beginning of the training, and intermediate evaluations were conducted periodically using the validation dataset. The final performance metrics were obtained from the held-out test set using a COCO-based instance segmentation evaluation.

PointRend (R50-FPN–based) To further improve the boundary delineation of restorations, a PointRend-based instance segmentation model was trained on the ResNet-50-FPN backbone. In this configuration, the standard mask head of Mask R-CNN was replaced with a point-based rendering head that iteratively refines the mask predictions by focusing on uncertain regions, particularly along the object boundaries. The model was configured for two-class segmentation and initialized using COCO-pretrained weights. The training parameters were aligned with those used for the Cascade Mask R-CNN, including a mini-batch size of two, an initial learning rate of 0.00025, and a total training duration of 80 epochs. The learning rate scheduling and warm-up strategies mirrored those of the cascade model (39,518 iterations). Validation was performed during training, and the final evaluation was conducted on an independent test dataset using the COCO instance segmentation framework.

Model evaluation and performance metrics

The final performance of all models was assessed using a completely held-out test set that was not used during training or model selection. The "best" checkpoint was selected based on the validation performance and then evaluated once on the independent test split to provide an unbiased estimate of generalization.

The instance segmentation performance was quantified using COCO-style metrics, including mean average precision across IoU thresholds of 0.50–0.95 (AP50–95), AP at IoU = 0.50 (AP50), AP at IoU = 0.75 (AP75), and average recall at a maximum of 100 detections per image (AR@100). Metrics were reported overall and class-wise, with a particular focus on the overhang class owing to its clinical relevance.

The uncertainty of the primary outcomes was quantified using non-parametric bootstrap resampling of the test set at the image level (sampling with replacement; B = 200). For each bootstrap replicate, the COCO metrics were recomputed, and 95% confidence intervals were derived using percentile-based limits. Bootstrap results were reported for both the overall metrics and the overhang class. The bootstrap confidence intervals were used to quantify the uncertainty around individual model performance estimates. No formal paired statistical hypothesis testing or pairwise model-comparison analysis was performed; therefore, comparisons between architectures were interpreted descriptively based on point estimates, confidence intervals, and consistent performance patterns across metrics.

Inference latency was assessed descriptively using the framework-reported timing outputs obtained during model evaluation on the held-out test set. All latency measurements were performed in a Google Colab GPU-enabled runtime equipped with an NVIDIA L4 GPU. YOLO-based models were evaluated using the Ultralytics validation pipeline according to their respective model-specific evaluation configurations, whereas Detectron2-based models were evaluated using the Detectron2 inference procedure. For YOLO-based models, end-to-end latency was calculated as the sum of the framework-reported preprocessing, inference, and postprocessing times

Model	AP50-95	AP50	AP75	AR@100
YOLO11n-seg	0.634 (0.598–0.670)	0.826 (0.787–0.862)	0.726 (0.675–0.777)	0.724 (0.698–0.754)
YOLO11s-seg	0.615 (0.579–0.651)	0.813 (0.765–0.852)	0.733 (0.683–0.778)	0.738 (0.710–0.764)
YOLO11m-seg	0.599 (0.554–0.635)	0.805 (0.754–0.849)	0.696 (0.644–0.744)	0.709 (0.677–0.733)
Mask R-CNN (R50-FPN)	0.689 (0.651–0.720)	0.866 (0.836–0.897)	0.786 (0.743–0.830)	0.830 (0.805–0.851)
Cascade Mask R-CNN (R50-FPN)	0.703 (0.671–0.734)	0.867 (0.839–0.901)	0.799 (0.756–0.836)	0.843 (0.821–0.863)
PointRend (R50-FPN)	0.697 (0.662–0.733)	0.864 (0.837–0.891)	0.785 (0.741–0.825)	0.831 (0.808–0.853)

Table 1. Overall segmentation performance on the test set based on bootstrap analysis (B = 200). AP50–95: mean average precision averaged across intersection-over-union thresholds from 0.50 to 0.95; AP50: average precision at an IoU threshold of 0.50; AP75: average precision at an IoU threshold of 0.75; AR@100: average recall with a maximum of 100 detections per image; CI: confidence interval; IoU: intersection over union.

Model	Overhang AP50-95	Overhang AP50	Overhang AP75	Overhang AR@100
YOLO11n-seg	0.672 (0.628–0.720)	0.833 (0.784–0.879)	0.777 (0.727–0.831)	0.754 (0.712–0.795)
YOLO11s-seg	0.668 (0.619–0.711)	0.830 (0.778–0.875)	0.791 (0.731–0.842)	0.780 (0.739–0.815)
YOLO11m-seg	0.660 (0.611–0.703)	0.845 (0.795–0.891)	0.782 (0.719–0.841)	0.791 (0.757–0.819)
Mask R-CNN (R50-FPN)	0.729 (0.685–0.775)	0.882 (0.851–0.917)	0.828 (0.773–0.881)	0.864 (0.836–0.886)
Cascade Mask R-CNN (R50-FPN)	0.737 (0.697–0.784)	0.870 (0.832–0.912)	0.825 (0.776–0.879)	0.878 (0.852–0.899)
PointRend (R50-FPN)	0.743 (0.704–0.786)	0.878 (0.845–0.910)	0.820 (0.771–0.861)	0.869 (0.842–0.891)

Table 2. Class-wise segmentation performance for the Overhang class on the test set based on bootstrap analysis (B = 200). AP50–95: mean average precision averaged across intersection-over-union thresholds from 0.50 to 0.95; AP50: average precision at an IoU threshold of 0.50; AP75: average precision at an IoU threshold of 0.75; AR@100: average recall with a maximum of 100 detections per image; CI: confidence interval; IoU: intersection over union.

per image, whereas pure compute latency corresponded to the reported inference time. For Detectron2-based models, end-to-end latency corresponded to the reported total inference time per iteration, and pure compute latency corresponded to the reported pure compute time per iteration. No separate manually repeated latency benchmarking protocol or dedicated warm-up inference run was implemented beyond the timing procedures used by the respective evaluation frameworks; therefore, latency values should be interpreted as descriptive framework-level estimates under the stated hardware and evaluation conditions.

Results

The overall instance segmentation performance on the independent test set is summarized in Table 1. Among all evaluated models, Cascade Mask R-CNN (R50-FPN) achieved the highest AP-based segmentation performance, yielding the greatest AP50–95 (0.703; 95% CI 0.671–0.734), AP75 (0.799; 95% CI 0.756–0.836), and AR@100 (0.843; 95% CI 0.821–0.863). PointRend (R50-FPN) and Mask R-CNN (R50-FPN) demonstrated comparable performance, with slightly lower but overlapping confidence intervals for AP50–95 (0.697 [0.662–0.733] and 0.689 [0.651–0.720], respectively). Across the two-stage models, AP50 values were consistently high and similar (0.864–0.867), indicating a robust detection performance at moderate overlap thresholds. In comparison, the YOLO11-seg models showed lower overall precision at stricter IoU thresholds, with AP50–95 values ranging from 0.622 to 0.634 and AP75 values between 0.696 and 0.733. Nevertheless, the AP50 values for the YOLO-based models remained relatively high (0.805–0.826), reflecting effective coarse localization. AR@100 followed a similar trend, with two-stage models achieving higher recall (0.831–0.843) than the YOLO11-seg variants (0.709–0.738).

The class-wise instance segmentation performance for the clinically relevant overhang class is presented in Table 2. Overall, the two-stage architectures demonstrated higher class-specific AP-based segmentation performance compared to the one-stage YOLO-based models. The highest Overhang AP50–95 was achieved by PointRend (R50-FPN) (0.743; 95% CI 0.704–0.786), closely followed by Cascade Mask R-CNN (R50-FPN) (0.737; 0.697–0.784) and Mask R-CNN (R50-FPN) (0.729; 0.685–0.775). In contrast, the YOLO11-seg variants yielded lower AP50–95 values for the overhang class, ranging from 0.660 to 0.672. Across all models, Overhang AP50 values were consistently high, with two-stage models achieving values between 0.870 and 0.882, whereas YOLO11-seg models ranged from 0.830 to 0.845. A similar pattern was observed for overhang AP75, where two-stage models reached values above 0.820 (up to 0.828 for Mask R-CNN), compared with 0.777–0.791 for YOLO11-seg models. The highest overhang AR@100 was observed for Cascade Mask R-CNN (0.878; 0.852–0.899) and PointRend (0.869; 0.842–0.891), indicating improved recall for overhanging restorations, whereas YOLO11-seg models demonstrated recall values between 0.754 and 0.791.

One-stage YOLO-based models demonstrated substantially faster inference compared with two-stage architectures. End-to-end latency for YOLO11 variants ranged from 10.4 to 14.9 ms per image, with pure

Model	Framework	End-to-end latency (ms/image)	Pure compute latency (ms/image)
YOLO11n-seg	Ultralytics	10.5	5.1
YOLO11s-seg	Ultralytics	10.4	6.0
YOLO11m-seg	Ultralytics	14.9	10.7
Mask R-CNN (R50-FPN)	Detectron2	53.2	43.7
PointRend (R50-FPN)	Detectron2	59.7	50.5
Cascade Mask R-CNN (R50-FPN)	Detectron2	67.4	58.0

Table 3. Inference latency of one-stage and two-stage instance segmentation models.

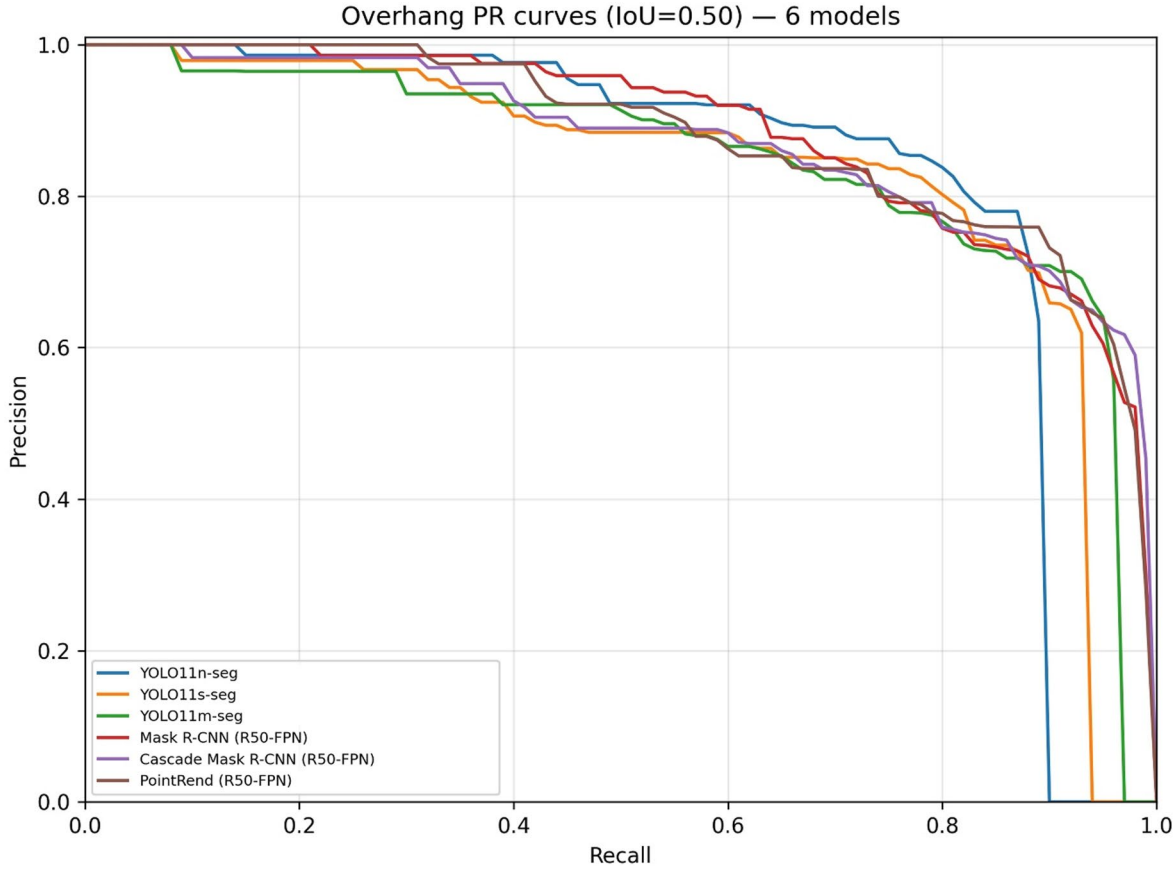

Fig. 2. Overhang class precision–recall curves at IoU = 0.50 for one-stage (YOLO11-seg) and two-stage (Detectron2) models.

compute times between 5.1 and 10.7 ms, reflecting their lightweight and fully integrated design. In contrast, two-stage models exhibited markedly higher latency, with end-to-end inference times of 53.2 ms for Mask R-CNN, 59.7 ms for PointRend, and 67.4 ms for Cascade Mask R-CNN. A similar trend was observed for pure compute latency, which exceeded 40 ms per image for all Detectron2-based architectures (Table 3).

The corresponding precision–recall curves at IoU = 0.50 (Fig. 2) illustrate the stability of the two-stage models at higher recall levels and the sharper precision drop observed in YOLO-based models as the recall approaches saturation.

Figure 3 provides a concise qualitative illustration of the performance characteristics of the PointRend (R50-FPN) two-stage instance segmentation model and helps to contextualize the quantitative results. In a clear-cut scenario without confounding structures (Fig. 3A), the model achieved near-perfect agreement with the ground truth in both classification and segmentation, accurately delineating overhanging margins when morphological expression was unambiguous. In a more complex multi-restoration setting (Fig. 3B), PointRend maintained precise boundary delineation across adjacent teeth and produced masks that remained confined to the restoration margins of interest, despite the presence of adjacent radiopaque materials with differing radiographic appearances related to endodontic treatment. The only discrepancy in this example occurred at the mandibular first molar, where the predicted mask remained close to the reference boundary, but the class assignment differed from the reference standard. In contrast, Fig. 3C highlights a limitation driven by radiographic contrast:

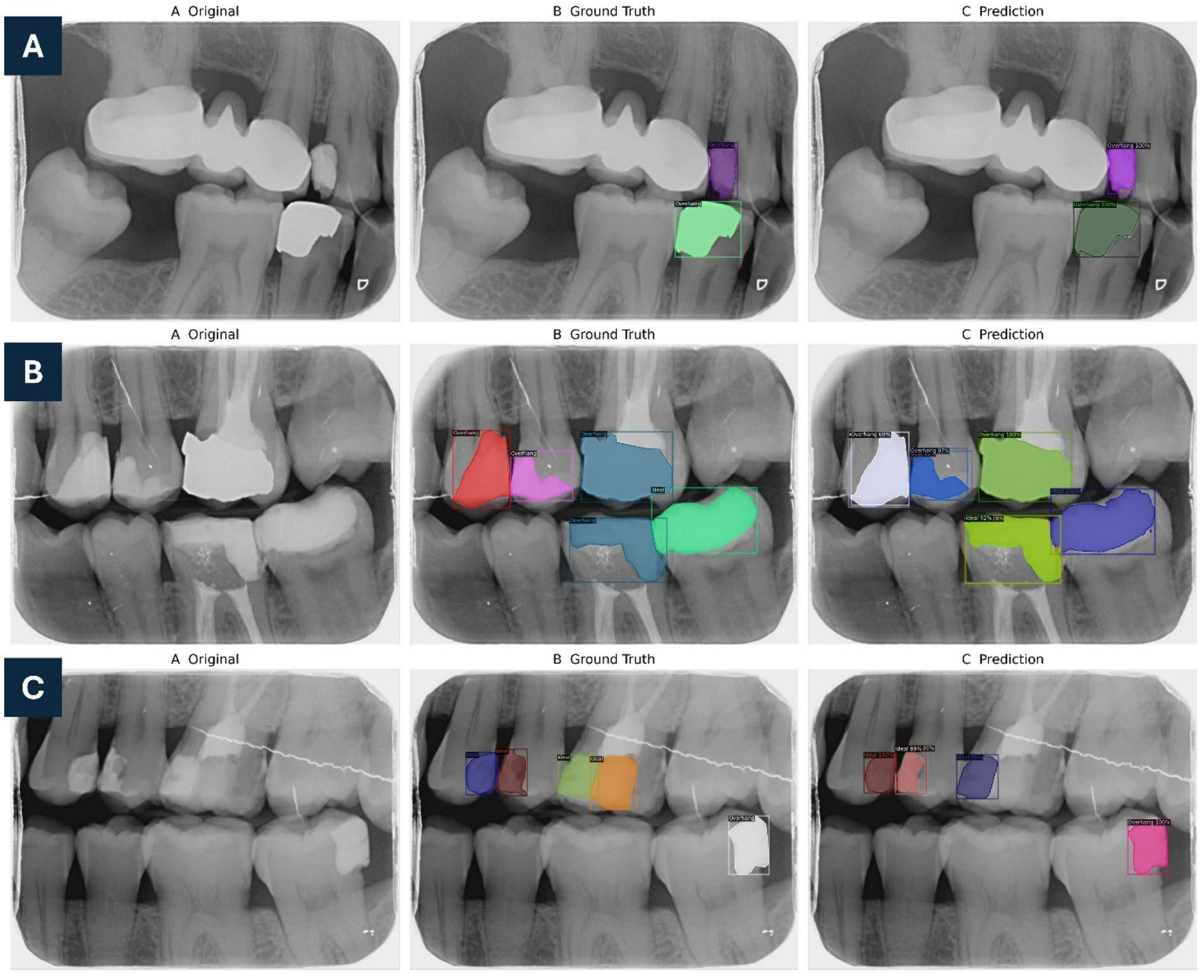

Fig. 3. Qualitative performance examples of the PointRend (R50-FPN) model: original images, ground truth annotations, and predictions.

a restoration on the maxillary first molar with radiopacity closely approximating enamel was inconsistently detected, likely due to reduced contrast at the tooth–restoration interface.

Figure 4 provides additional qualitative insights into the model behavior under radiographically challenging conditions. As shown in Fig. 4A, the model delineated restorations with differing radiographic appearances and material-related radiopacity patterns, while generally maintaining the predicted masks within the relevant restoration boundaries under complex radiographic conditions. Segmentation and classification errors occurred at the maxillary first premolar, where restorations on adjacent teeth appeared merged owing to overlapping radiopacity and projection geometry. In Fig. 4B, a full-coverage restoration on the maxillary first molar was incorrectly predicted as an overhanging restoration, whereas Fig. 4C illustrates another instance in which a restoration on the mandibular second premolar, annotated as ideal, was labeled as overhanging.

Discussion

The present study systematically compared one- and two-stage deep learning–based instance segmentation architectures for the detection of overhanging dental restorations on bitewing radiographs, with particular emphasis on boundary accuracy and class-specific performance. Taken together, the results demonstrate a consistent performance pattern within this experimental setting: while all evaluated models were generally capable of localizing restorations, their ability to accurately delineate restoration margins, especially under stricter overlap criteria, varied substantially according to the architectural design. Specifically, the uniformly high AP50 values observed across both the one-stage and two-stage models indicate that the coarse detection and localization were generally successful across the evaluated models in this dataset. However, the divergence observed at higher IoU thresholds (AP75 and AP50–95) suggests that precise boundary delineation, which is critical for identifying subtle marginal discrepancies such as overhangs, was more challenging for one-stage architectures. Notably, the two-stage models showed consistently higher AP-based point estimates across the AP50–95 metric range, suggesting better boundary refinement under the present experimental conditions. Accordingly, the descriptive performance pattern observed in this study did not support the null hypothesis that

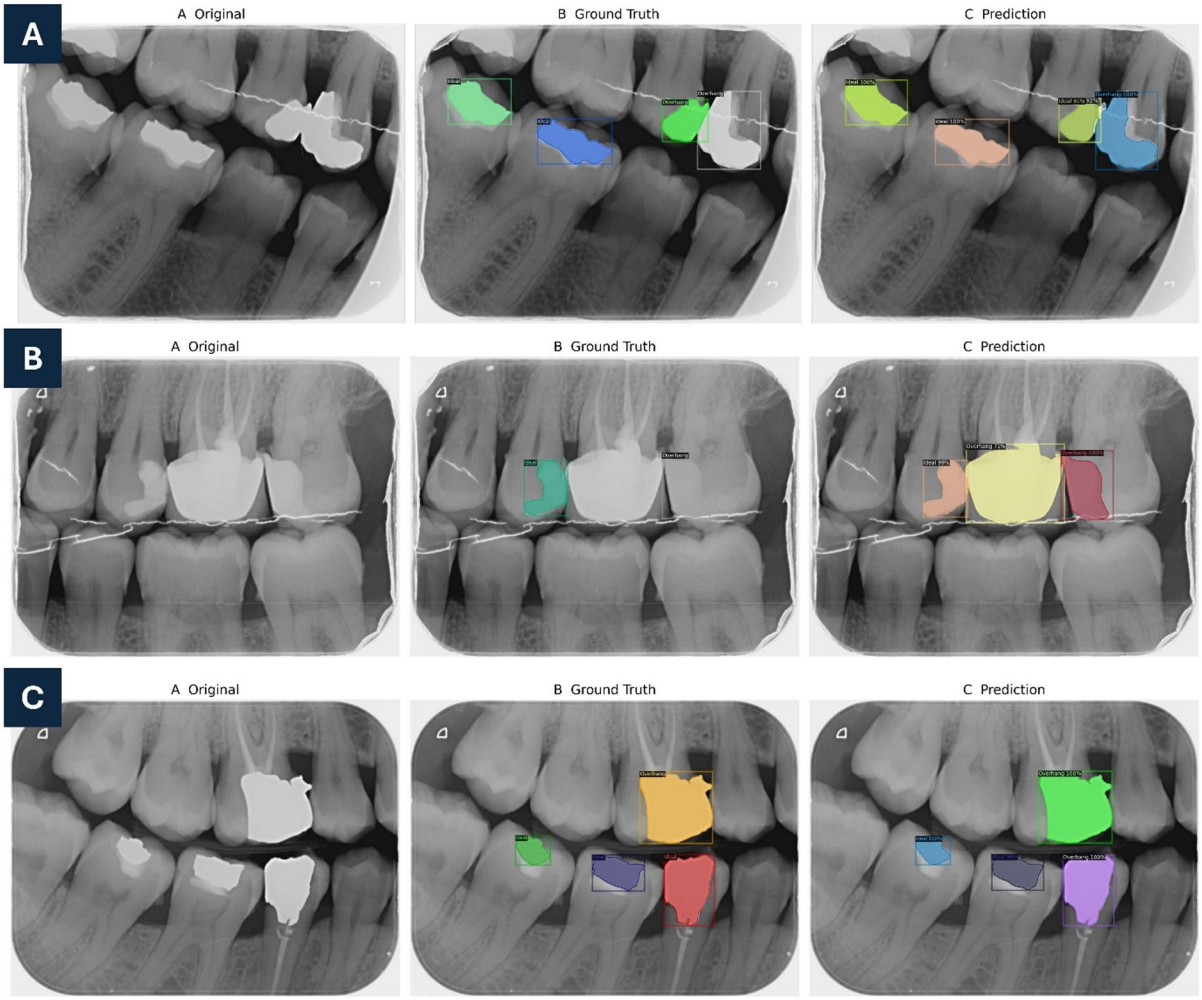

Fig. 4. Additional qualitative cases highlighting model behavior under challenging radiographic conditions (PointRend, R50-FPN).

one- and two-stage instance segmentation architectures would demonstrate comparable performance across the evaluated metrics.

The observed performance gap between one-stage and two-stage instance segmentation models can be primarily attributed to the fundamental architectural differences in the handling of object localization and boundary refinement. One-stage models, such as YOLO-based architectures, jointly perform localization, classification, and mask prediction in a single forward pass. Although this design confers substantial computational efficiency and enables strong performance in coarse localization, as reflected by high AP50 values, it inherently limits the degree of region-specific feature refinement available for precise boundary delineation[21,22]. In the context of bitewing radiographs, where overhanging restorations often manifest as small marginal protrusions or subtle contour irregularities, this limitation becomes particularly relevant. Minor inaccuracies along the restoration boundaries are disproportionately penalized at higher IoU thresholds, explaining the comparatively lower AP75 and AP50–95 values observed for the one-stage models[23]. From a clinical perspective, improved AP75 and AP50–95 values may be relevant because the diagnosis of an overhanging restoration depends on accurate delineation of the restoration margin rather than coarse localization alone. More precise segmentation may reduce confusion between true marginal excess and visually similar radiographic appearances, thereby supporting more consistent clinician review of suspected overhangs on bitewing radiographs. In this sense, the potential translational value of higher boundary accuracy lies in improving margin-focused assessment and case triage, rather than in making autonomous treatment decisions.

In contrast, two-stage architectures explicitly decouple candidate region generation from subsequent region-based classification and mask refinement[24]. The use of a Region Proposal Network, followed by ROI-aligned feature extraction, allows these models to focus computational resources on localized areas likely to contain restorations, thereby enhancing sensitivity to fine-grained geometric details[25]. Moreover, advanced two-stage variants further amplify this advantage using iterative or boundary-aware refinement mechanisms. Cascade Mask R-CNN progressively refines bounding boxes across multiple IoU thresholds, reducing localization error that would otherwise propagate to mask prediction[26], while PointRend selectively refines uncertain points

(predominantly along object edges) using high-resolution features[27]. These design choices are well aligned with the morphological characteristics of overhanging restorations and may partly explain the comparatively better performance of the two-stage models at stricter overlap criteria and in overhang-specific metrics in this dataset[24].

In addition to architectural considerations, the interaction between model capacity and dataset characteristics may also contribute to the observed differences. One-stage models with higher parameter counts did not consistently outperform their lighter counterparts, suggesting that an increased capacity alone does not guarantee improved boundary accuracy for this task. In relatively limited or specialized medical imaging datasets, larger models may be more susceptible to overfitting or may fail to fully exploit their representational power without extensive task-specific augmentation or optimization[28]. Conversely, the inductive biases embedded in two-stage designs, particularly their emphasis on region-level processing and boundary refinement, appear to be more directly suited to detecting marginal defects defined by small spatial deviations rather than gross object presence[29].

When YOLO-based one-stage models were examined internally, increasing the model capacity did not lead to a clear or statistically robust improvement in the segmentation performance. Bootstrap-based confidence intervals showed substantial overlap among YOLO11n-seg, YOLO11s-seg, and YOLO11m-seg across both overall and overhang-specific metrics, indicating that no definitive performance separation could be established within the YOLO family of models. Nevertheless, YOLO11n-seg demonstrated a consistent numerical trend toward higher overall AP50–95 and comparable or slightly higher overhang-specific performance than the larger YOLO variants. This observation contrasts with much of the existing literature, where YOLO11m-seg typically achieves superior performance on general object detection and segmentation benchmarks[30–32]. Recent dental radiographic applications have also reported promising results for YOLO11-based frameworks. For example, Ayhan et al.[33] proposed RCFLA-YOLO for the automated assessment of root canal filling quality on periapical radiographs, further supporting the potential of YOLO11-based models for objective radiographic quality assessment in dentistry. The discrepancy likely reflects the task-specific nature of overhanging restoration detection, which relies on identifying subtle marginal protrusions and fine boundary irregularities rather than large, high-contrast objects. In such settings, increased model capacity may not translate into improved boundary sensitivity and may instead amplify dataset-specific textures or noise, particularly in the absence of extensive task-specific augmentation[14]. Moreover, the tightly coupled localization and segmentation processes inherent to one-stage architectures can limit the optimization of fine-grained mask quality, allowing localization errors to propagate directly to mask prediction regardless of model size. Collectively, these findings suggest that, contrary to trends reported in the broader computer vision literature, scaling YOLO models did not confer a clear advantage in the present dataset and training setting, and that alignment between task characteristics and architectural inductive bias is more critical than increased capacity alone.

When the two-stage architectures were compared internally, a clear hierarchy emerged between the standard Mask R-CNN, Cascade Mask R-CNN, and PointRend, reflecting the progressive incorporation of the localization and boundary-refinement mechanisms. Among these models, the Cascade Mask R-CNN demonstrated the strongest overall performance across global metrics (AP50–95, AP75, and AR@100), consistent with previous studies[34–36]. This may indicate its superior robustness for instance localization and segmentation under stricter overlap criteria. This advantage can be attributed to its cascaded design, in which the bounding box regression and classification are iteratively refined across increasing IoU thresholds. By reducing the localization error at each stage, the Cascade Mask R-CNN minimizes error propagation to the mask head, resulting in more stable and accurate instance segmentation across a wide range of IoU levels[37]. In contrast, PointRend achieved the highest class-specific performance for the clinically critical overhang class, particularly in terms of AP50–95. Unlike Cascade Mask R-CNN, which primarily improves the global localization quality, PointRend directly targets boundary precision by selectively refining uncertain points along object edges using high-resolution features[38]. This boundary-aware rendering strategy is particularly advantageous for detecting overhanging restorations, which are often characterized by small marginal extensions rather than large geometric deviations[39]. Consequently, PointRend may sacrifice some global consistency in favor of sharper and more accurate delineation of subtle marginal protrusions, which may explain its class-specific numerical advantage despite slightly lower overall metrics compared with the cascade architecture.

Qualitative analysis of representative cases (Figs. 3 and 4) underscores that the primary strengths of the PointRend model lie in its ability to achieve precise boundary refinement and robust instance separation, even under radiographically complex conditions. The few observed misclassifications were primarily associated with borderline or ambiguous marginal definitions, reduced radiographic contrast, and projection artifacts rather than true segmentation failures. In particular, errors involving the misclassification of full-coverage or ideal restorations as overhanging appear to stem from visual patterns in the training data, such as radiolucent gaps or radiopacity overlaps, that mimic the appearance of marginal excess. This suggests that the model behavior may reflect the contextual biases inherent to radiographic morphology rather than the architectural limitations of the segmentation framework itself. Taken together, these findings suggest that some of the remaining errors may be related to image-level confounders and annotation subjectivity, although the contribution of model architecture cannot be fully excluded. Future improvements may therefore benefit more from dataset diversity, multiview training, and contrast normalization than from further model modification.

The present study has several limitations. First, the dataset was derived from a single academic center and acquired using a single imaging system, which may limit the generalizability of the findings to other clinical settings, radiographic devices, and acquisition protocols. Although strict image quality criteria were applied to ensure reliable ground truth, this selection may not fully reflect the variability encountered in routine clinical practice. Second, no task-specific data augmentation strategies were implemented beyond the default pipelines of the respective frameworks. While this choice supported standardized model comparison, it may have constrained the performance of certain architectures, particularly higher-capacity one-stage models,

which could potentially benefit from tailored augmentation or optimization. In addition, although the dataset split and evaluation protocol were standardized across all models, the training configurations differed between architectures because model-specific and framework-specific settings were used. These differences included input resolution, number of epochs, fine-tuning strategy, batch-size selection, and default augmentation pipelines. Therefore, the present findings should be interpreted as a comparison under standardized data and evaluation conditions rather than as a fully hyperparameter-controlled comparison. Third, this study focused exclusively on bitewing radiographs and the binary classification of restoration margins, without considering other clinically relevant restoration defects or multi-class marginal conditions, which may limit the scope of applicability. Fourth, although polygon-based annotations provided high-fidelity reference masks, the definition of overhanging restorations was based on a radiographic threshold and expert consensus rather than direct clinical or histological validation, introducing potential subjectivity into the reference standard. Moreover, the approximately 0.5-mm radiographic threshold was not supported by formal metric calibration using pixel spacing, calibration phantoms, magnification correction, or anatomical reference-based scaling. Therefore, subtle overhangs may have been influenced by image resolution, beam angulation, projection geometry, and radiographic contrast. In addition, subtle or borderline overhanging margins on bitewing radiographs may be challenging even for expert annotators, particularly when marginal discrepancies are small, radiographic contrast is limited, or projection geometry is unfavorable. Although a specialist-based consensus annotation workflow was used to reduce individual observer bias, formal inter-rater reliability statistics were not calculated because independently retained annotation sets were not available. Therefore, annotation uncertainty may have partly influenced model performance, especially in cases involving the distinction between ideal and overhanging restorations. Finally, although inference latency was reported to provide a general indication of computational efficiency, deployment-related performance has not been evaluated under real clinical workflow conditions or across different hardware environments. Therefore, the practical feasibility of routine implementation warrants further investigation.

The present study also offers several methodological advantages over much of the previous literature. First, all evaluated models were assessed using the same dataset split, annotation framework, and evaluation protocol, allowing a standardized comparison of one-stage and two-stage architectures under shared data and evaluation conditions. Second, performance was not summarized only by a single detection threshold, but was assessed using COCO-style instance segmentation metrics across multiple IoU levels, which is especially relevant for overhanging restorations because clinically meaningful errors often arise from boundary imprecision rather than complete detection failure. Third, uncertainty was quantified using bootstrap-based confidence intervals, which provides a more robust estimate of comparative model behavior than reporting point estimates alone. Taken together, these features clarify that the main contribution of the present work is not only the demonstration of AI feasibility, but the identification of architecture-dependent performance differences for a boundary-critical diagnostic task on bitewing radiographs.

Conclusion

This study demonstrated that deep-learning-based instance segmentation models can achieve promising performance in detecting overhanging dental restorations on bitewing radiographs in a controlled experimental setting. Although both one-stage and two-stage architectures achieved strong coarse localization, two-stage models showed consistently higher AP-based point estimates than YOLO-based approaches under stricter IoU criteria. Within this framework, Cascade Mask R-CNN showed the highest overall AP50–95 point estimate, whereas PointRend showed the highest overhang-specific AP50–95 point estimate. Overall, these findings suggest that architectural design, particularly region-level processing and boundary-focused refinement, may be more influential than model scale alone for accurate margin assessment in this dataset. However, these results should be considered preliminary, and external multicenter validation is required before conclusions regarding routine clinical applicability or decision support integration can be drawn.

Data availability

The datasets generated and/or analyzed during the current study are not publicly available because individual consent for public data sharing was not obtained and the data consist of clinical radiographic images. However, anonymized data may be made available from the corresponding author upon reasonable request and subject to institutional and ethical approval.

Received: 20 January 2026; Accepted: 8 June 2026
Published online: 11 June 2026

References

1. Indurkar, M. S., Dhumal, K. S. & Verma, R. Quantitative evaluation of bone loss around carious and restored teeth: A CBCT study. *J. Med. Sci. Clin. Res.* https://doi.org/10.18535/jmscr/v7i9.100 (2019).
2. Muryani, A., Amaliya, A., Garna, D. F., Oscandar, F. & Sukartini, E. Overhanging approximal restoration: clinical and radiography features at Tarogong Public Health Service Indonesia. *Padjadjaran J. Dent* **28**(2), 85–88 (2016).
3. Litzenburger, F. et al. Inter-and intraexaminer reliability of bitewing radiography and near-infrared light transillumination for proximal caries detection and assessment. *Dentomaxillofacial Radiol.* **47**(3), 20170292 (2018).
4. Liedke, G., Spin-Neto, R., Da Silveira, H. & Wenzel, A. Radiographic diagnosis of dental restoration misfit: a systematic review. *J. Oral Rehabil.* **41**(12), 957–967 (2014).
5. Khubrani, Y. H., Thomas, D., Slator, P. J., White, R. D. & Farnell, D. J. 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**(2), 89–108 (2025).

6. Mema, H. et al. Application of AI-driven software diagnocat in managing diagnostic imaging in dentistry: A retrospective study. *Appl. Sci.* **15**(17), 9790 (2025).
7. Rohrer, C. et al. Segmentation of dental restorations on panoramic radiographs using deep learning. *Diagnostics* **12**(6), 1316 (2022).
8. Luo, D., Zeng, W., Chen, J. & Tang, W. Deep learning for automatic image segmentation in stomatology and its clinical application. *Front. Med. Technol.* **3**, 767836 (2021).
9. Magat, G. et al. Automatic deep learning detection of overhanging restorations in bitewing radiographs. *Dentomaxillofacial Radiol.* **53**(7), 468–477 (2024).
10. Fatima, A. et al. Deep learning-based multiclass instance segmentation for dental lesion detection. *Healthcare* **11**(3), 347. https://doi.org/10.3390/healthcare11030347 (2023).
11. Bae, B., Choi, Y., Jung, H. & Ahn, J. Tunnel lining segmentation from ground-penetrating radar images using advanced single- and two-stage object detection and segmentation models. *Computer-Aided Civil Infrastruct. Eng.* **40**(22), 3580–3592 (2025).
12. Katsumata, A. Deep learning and artificial intelligence in dental diagnostic imaging. *Jpn. Dental Sci. Rev.* **59**, 329–333 (2023).
13. Akgül, N. et al. A YOLO-V5 approach for the evaluation of normal fillings and overhanging fillings: an artificial intelligence study. *Braz. Oral Res.* **38**, e098 (2024).
14. Çelik, B. & Çelik, M. E. Automated detection of dental restorations using deep learning on panoramic radiographs. *Dentomaxillofacial Radiol.* **51**(8), 20220244 (2022).
15. Bayırlı, A. B. et al. Segmentation-based multi-class detection and radiographic charting of periodontal and restorative conditions on bitewing radiographs using deep learning. *Diagnostics* **16**(2), 322 (2026).
16. Bi, H., Wen, V. & Xu, Z. Comparing one-stage and two-stage learning strategy in object detection. *Appl. Comput. Eng.* **5**(1), 171–177 (2023).
17. Schwendicke, F. et al. Artificial intelligence in dental research: Checklist for authors, reviewers, readers. *J. Dent.* **107**, 103610 (2021).
18. Mongan, J., Moy, L. & Kahn, C. E. Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A guide for authors and reviewers. *Radiol. Artif. Intell.* **2**(2), e200029. https://doi.org/10.1148/ryai.2020200029 (2020).
19. Bossuyt, P. M. et al. STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *Radiology* **277**(3), 826–832 (2015).
20. Claman, L. J., Koidis, P. T. & Burch, J. G. Proximal tooth surface quality and periodontal probing depth. *J. Am. Dent. Assoc.* **113**(6), 890–893 (1986).
21. Zhang, Q.-L. & Yang, Y.-B. A boundary-preserving conditional convolution network for instance segmentation. *Pattern Recogn. Lett.* **163**, 1–9 (2022).
22. Wen, R., Xie, W., Fan, Y. & Shen, L. SABE-YOLO: structure-aware and boundary-enhanced YOLO for weld seam instance segmentation. *J. Imaging* **11**(8), 262 (2025).
23. Ribeiro, R.P.S. & von Wangenheim, A. Instance segmentation in medical imaging: A comparative study of CNN and transformer-based models in a teledermatology study-case. In: Anais do Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS), 819–827. (Sociedade Brasileira de Computação, 2025). https://doi.org/10.5753/sbcas.2025.7814.
24. Bae, B., Choi, Y., Jung, H. & Ahn, J. Tunnel lining segmentation from ground-penetrating radar images using advanced single-and two-stage object detection and segmentation models. *Comput. Aided Civil Infrastruct. Eng.* (2025).
25. Rossi, L., Karimi, A. & Prati, A. A novel region of interest extraction layer for instance segmentation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 2203–2209 (2021). https://doi.org/10.1109/ICPR48806.2021.9412258.
26. Cai, Z. & Vasconcelos, N. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(5), 1483–1498 (2019).
27. Kirillov, A., Wu, Y., He, K. & Girshick, R. PointRend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020). https://doi.org/10.1109/CVPR42600.2020.00982.
28. Karaman, F. & Gümüş, F. Exploring the impact of model capacity and parameter tuning on 3D semantic segmentation. *Turk. J. Sci. Technol.* **20**(1), 327–337 (2025).
29. Abudukelimu, H. et al. DVF-YOLO-Seg: A two-stage breast mass segmentation model with enhanced feature extraction and small lesion detection. *Digital Health* **11**, 20552076251374190 (2025).
30. Sahoo, A. R., Sahoo, S. S. & Chakraborty, P. Polyp detection in colonoscopy images using YOLOv11. arXiv preprint arXiv:2501.09051 (2025).
31. Stepanenko, S. et al. Substantiating the YOLO11 architecture for determining the fractional composition of winter wheat grain mixtures. *Eastern-Eur. J. Enterp. Technol.* **4**(2(136)), 81–92. https://doi.org/10.15587/1729-4061.2025.338124 (2025).
32. Sapkota, R. & Karkee, M. Comparing YOLOv11 and YOLOv8 for instance segmentation of occluded and non-occluded immature green fruits in complex orchard environment. arXiv preprint arXiv:2410.19869 (2024).
33. Ayhan, M., Kayadibi, İ & Aykanat, B. RCFLA-YOLO: a deep learning-driven framework for the automated assessment of root canal filling quality in periapical radiographs. *BMC Med. Educ.* **25**(1), 894. https://doi.org/10.1186/s12909-025-07483-2 (2025).
34. Oh, H. Y., Khan, M. S., Jeon, S. B. & Jeong, M.-H. Automated detection of greenhouse structures using cascade mask R-CNN. *Appl. Sci.* **12**(11), 5553 (2022).
35. Yu, C., Sun, Y., Cao, Y., Liu, L. & Zhou, X. Log volume measurement and counting based on improved cascade mask R-CNN and deep SORT. *Forests* **15**(11), 1884 (2024).
36. Yang, Z., Dong, R., Xu, H. & Gu, J. Instance segmentation method based on improved mask R-CNN for the stacked electronic components. *Electronics* **9**(6), 886 (2020).
37. Wang, W., Xu, X. & Yang, H. Intelligent detection of tunnel leakage based on improved mask R-CNN. *Symmetry* **16**(6), 709 (2024).
38. Yin, H. et al. AC R-CNN: Pixelwise instance segmentation model for agrocybe cylindracea cap. *Agronomy* **14**(1), 77 (2023).
39. Zhu, B. et al. AP-PointRend: An improved network for building extraction via high-resolution remote sensing images. *Remote Sens.* **17**(9), 1481 (2025).

Author contributions

Ö.H. conceived and designed the study, supervised the overall research process, performed the data analysis, and drafted the manuscript. Ö.B., G.M, and A.A. contributed to data collection and investigation, including image evaluation and annotation processes. F.P.H. critically revised the manuscript for important intellectual content. All authors read and approved the final version of the manuscript.

Funding

No funding was received.

Declarations

Competing interests
The authors declare no competing interests.

Ethics approval and consent to participate
Ethical approval for the retrospective use of anonymized radiographic data was obtained from the Necmettin Erbakan University Ethics Committee (Approval No:2025/542).

Additional information
Correspondence and requests for materials should be addressed to Ö.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.