J Dent Mater Tech 2025; 15 (2): 88-97
DOI: 10.22038/jdmt.2026.94821.1982

J Dent Mater Tech

Original Article Open Access

Automated detection and severity classification of proximal dental caries on bitewing radiographs using Faster R-CNN and ResNet-50 deep learning models

Mohammad Javad Moghaddas [1], Arman Monajemi Mamaghani [2], Amin Amiri Tehranizadeh [3], Reza Izanlou [4], Salehe Sekandari [1,4], Samareh Mortazavi [5], Hoda Zare [6], Zahra KhandanKhadem-Reza [6], Fatemeh Baharvand Ahmadi [3], Farnaz Mohajertehran* [4,5]

Abstract

Objective: This study aimed to develop and evaluate a two-stage deep learning framework for automated localization and severity classification of proximal caries on bitewing radiographs according to the radiographic criteria of the International Caries Classification and Management System (ICCMS).

Methods: In this retrospective single-center model-development study, 400 bitewing radiographs were retrieved from the picture archiving and communication system of Mashhad University of Medical Sciences, Mashhad, Iran. After quality screening, 350 radiographs were included. Interproximal lesions were labeled by two restorative dentistry specialists, with disagreements resolved by an oral and maxillofacial radiologist. Lesion localization was performed using Faster R-CNN, and severity classification was performed using ResNet-50 V1 and V2. The dataset was divided into training, validation, and test sets containing 278, 36, and 36 radiographs, respectively. Models were evaluated in six-class ICCMS-based severity grading and three-class grouped severity classification. Performance was assessed using accuracy, precision, recall, specificity, F1-score, and mean average precision (mAP).

Results: In the three-class task, ResNet-50 V1 achieved the highest accuracy (0.92), while both models showed comparable F1-scores and mAP values. In the six-class task, ResNet-50 V2 outperformed V1, achieving accuracy, precision, recall, specificity, F1-score, and mAP values of 0.84, 0.60, 0.55, 0.93, 0.57, and 0.44, respectively. Performance was strongest for early lesions and lower for advanced classes with fewer annotated examples.

Conclusions: The Faster R-CNN and ResNet-50 framework showed preliminary feasibility for automated caries localization and ICCMS-based severity classification on bitewing radiographs. Larger, balanced, multi-center datasets with external validation are required before clinical implementation.

Keywords: Artificial intelligence, Bitewing radiography, Clinical decision support systems, Convolutional neural networks, Deep learning, Dental caries

Introduction

Dental caries is one of the most prevalent chronic oral diseases worldwide and remains a major cause of pain, infection, impaired oral function, and tooth loss if not detected and managed at an early stage (1-3). Early diagnosis is clinically important because non-cavitated and shallow lesions may be managed using preventive or minimally invasive approaches, whereas advanced dentinal lesions often require operative intervention. Accurate radiographic assessment is therefore essential for treatment planning, disease monitoring, and reducing unnecessary restorative procedures (4, 5).

Bitewing radiography is widely used for detecting proximal and occlusal caries because it provides detailed visualization of the crowns of posterior teeth and the interproximal surfaces (6). However, radiographic diagnosis remains inherently challenging. Lesion visibility may be affected by radiographic contrast and brightness, projection geometry, proximal overlap, restoration margins, and observer experience (7). Early enamel lesions, in particular, may be difficult to detect because radiographic changes become visible only after sufficient mineral loss has occurred. Therefore, bitewing

[1] Department of Restorative Dentistry, School of Dentistry, Mashhad University of Medical Sciences, Mashhad, Iran
[2] Faculty of Pharmacy, Mashhad University of Medical Sciences, Mashhad, Iran
[3] Department of Medical Informatics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran
[4] Dental Research Center, Mashhad University of Medical Sciences, Mashhad, Iran
[5] Oral and Maxillofacial Diseases Research Center, Mashhad University of Medical Sciences, Mashhad, Iran
[6] Department of Medical Physics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran

*Corresponding Author: Dr Farnaz Mohajertehran
Email: mohajertf@mums.ac.ir

Accepted: 19 May 2026. Submitted: 2 February 2026

radiographs should be interpreted as an adjunct to clinical examination rather than as an isolated diagnostic gold standard (8). Therefore, bitewing radiographs should be interpreted as an adjunct to clinical examination rather than as an isolated diagnostic gold standard.

Artificial intelligence, particularly convolutional neural networks (CNNs), has increasingly been investigated as a tool for detecting dental caries on radiographic images (9). Previous studies have evaluated a range of deep learning architectures for caries detection and classification on bitewing radiographs, including conventional CNNs, ResNet-based classifiers, VGG-16, U-Net, Faster R-CNN, RetinaNet, EfficientDet, and YOLO-based detectors (9-13). These studies have demonstrated the potential of deep learning to assist in caries detection, lesion localization, severity classification, and radiographic image interpretation based on bitewing radiographs (9, 11, 14, 15).

However, many published models have focused on binary classification, such as caries versus no caries, or on lesion detection without clinically detailed severity grading (13-16). Such outputs may be insufficient for clinical decision-making because treatment decisions depend not only on whether a lesion is present, but also on its radiographic depth and severity.

The International Caries Classification and Management System (ICCMS) provides a structured framework for classifying caries severity and guiding caries management decisions. Incorporating ICCMS-based grading into artificial intelligence systems may make automated radiographic interpretation more clinically useful by allowing the model to distinguish early enamel lesions from progressively deeper dentinal lesions. This distinction is important because early lesions may be managed non-operatively, whereas deeper dentinal lesions are more likely to require restorative treatment.

Nevertheless, multi-class severity classification is more difficult than binary caries detection. Adjacent ICCMS categories may show subtle radiographic differences, class distributions are often imbalanced, and advanced untreated lesions may be underrepresented in routine datasets because such lesions are more likely to be clinically evident, symptomatic, or treated before being captured in screening-oriented radiographic archives (17). These factors may reduce model performance, particularly for minority classes and severe lesion categories.

To address this gap, the present study developed and internally evaluated a two-stage deep learning framework for automated caries assessment on bitewing radiographs. In the first stage, Faster Region-based Convolutional Neural Networks were used to localize suspected carious regions. In the second stage, ResNet-50 V1 and ResNet-50 V2 were used to classify the detected regions according to ICCMS-based radiographic severity categories. The present study aimed to assess the performance of this framework in detecting proximal caries on bitewing radiographs using both detailed six-class severity grading and broader three-class classification, and to compare the performance of two ResNet-50 variants across both classification tasks.

Materials and methods

Study design, setting, and ethical considerations

This retrospective single-center diagnostic model-development study was conducted using bitewing radiographs obtained from the Faculty of Dentistry, Mashhad University of Medical Sciences, Mashhad, Iran. The study was approved by the ethics committee of Mashhad University of Medical Sciences (IR.MUMS.DENTISTRY.REC.1403.050). All radiographs were de-identified before image processing and model development.

Image dataset and eligibility criteria

A total of 400 bitewing radiographs of permanent teeth were retrospectively retrieved from the picture archiving and communication system (PACS) of the Faculty of Dentistry, Mashhad University of Medical Sciences, Mashhad, Iran, between July 2022 and September 2024. All radiographs were acquired using a Sirona Dental Systems unit (Dentsply Sirona, Bensheim, Germany) at 60 kVp and 7 mA. Exposure time was adjusted according to the manufacturer's standard protocol and the patient's physical characteristics.

Radiographs were eligible if they showed permanent teeth and had sufficient diagnostic quality for assessment of interproximal caries. Radiographs were excluded if they included primary teeth, inadequate contrast, severe image degradation, major positioning errors, or interproximal overlap exceeding one-third of the enamel thickness. After quality screening by an independent dentist, 350 radiographs were included in the final dataset.

Reference standard and annotation protocol

The radiographic reference standard was established through observer calibration, independent assessment, and adjudication. Before formal assessment, the

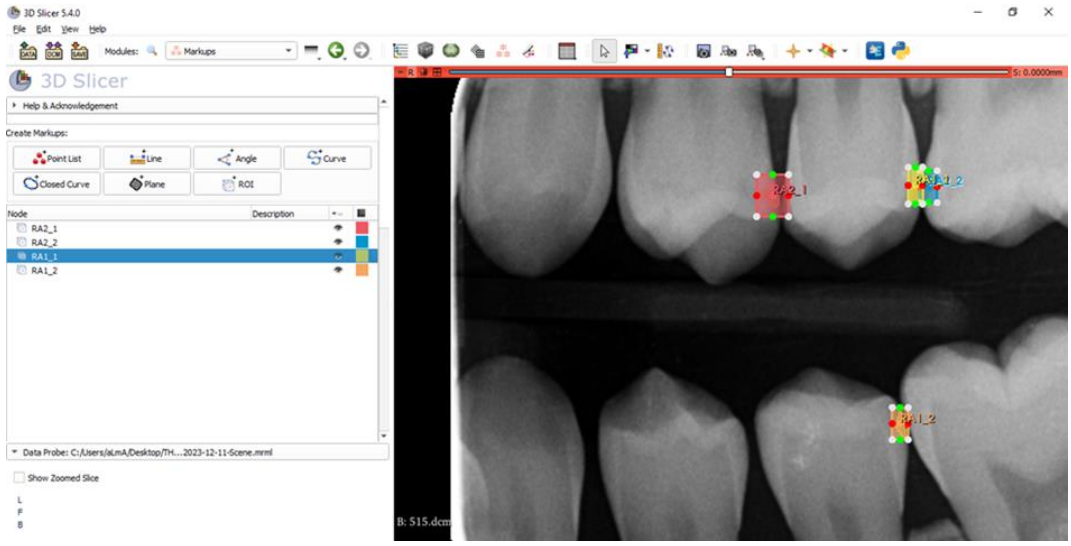

Figure 1. Representative region-of-interest annotation and radiographic caries grading by expert observers. The selected area was manually delineated and assigned an ICCMS-based radiographic severity category.

observers completed a calibration session using 50 bitewing radiographs that were not included in the final dataset. This calibration step was performed to standardize the application of radiographic ICCMS criteria and reduce inter-observer variability.

Two specialists in restorative dentistry independently reviewed all included bitewing radiographs and assigned radiographic caries severity labels. The two observers were blinded to each other's labels. Any disagreement between the two primary observers was reviewed by an oral and maxillofacial radiologist, who made the final decision on the radiographic label. Accordingly, the final reference labels were based on expert consensus interpretation of the bitewing radiographs. Because no histopathological or operative confirmation was available, these labels should be interpreted as an expert radiographic reference standard rather than definitive ground truth.

After final reference labels had been established, regions of interest (ROIs) were manually annotated using 3D Slicer software (version 1.1.4.5; Brigham and Women's Hospital Inc., Boston, MA, USA) as presented in Figure 1. Each carious lesion was delineated and assigned the corresponding final reference label. When multiple lesions were present on the same radiograph, each lesion was annotated and recorded as a separate region of interest.

When manual contours were used, they were converted into rectangular bounding boxes so that they could be used as reference annotations for the Faster R-CNN object-detection model. The finalized ROI annotations and corresponding labels were then exported and used for model training, validation, and testing. Representative examples of caries grading are shown in Figure 2.

For model development and evaluation, carious ICCMS-based radiographic categories were classified in two ways. The six-class task included RA1, RA2, RA3, RB4, RC5, and RC6. For the three-class task, these categories were collapsed into broader severity groups:

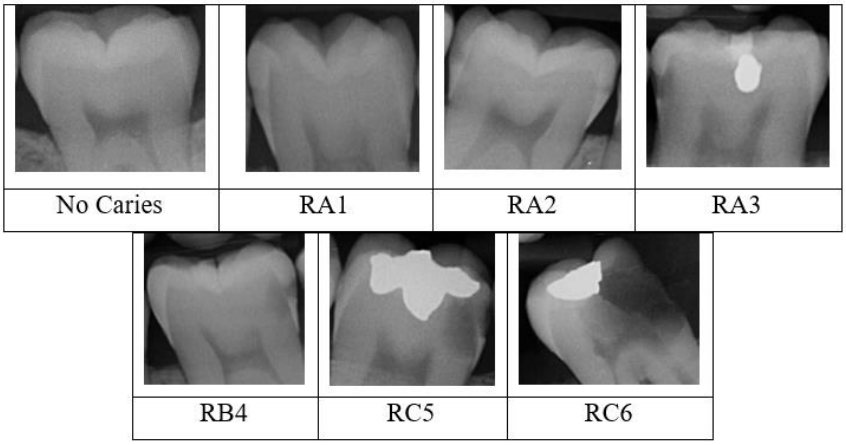

Figure 2. Distribution of annotated radiographic caries categories according to the ICCMS-based classification system. The figure shows the number of regions of interest assigned to each diagnostic category.

RA, representing initial-stage caries; RB, representing moderate-stage caries; and RC, representing extensive-stage caries. In total, 4126 annotated carious ROIs were included in the analysis. The definitions of the three-class and six-class categories, together with the total ROI distribution, are presented in Table 1.

Data splitting and validation strategy

The dataset was divided into training, validation, and independent test sets at the radiograph level. This means that all regions of interest extracted from the same radiograph were assigned to the same subset. This approach was used to prevent information from the same radiograph appearing in both training and testing data, which could otherwise lead to overly optimistic performance estimates. The final split included 278 radiographs in the training set, 36 radiographs in the validation set, and 36 radiographs in the test set.

The training set was used for model fitting, the validation set was used for monitoring model performance and early stopping, and the independent test set was reserved exclusively for final performance evaluation.

Image preprocessing and data augmentation

All annotated radiographs were exported from the 3D Slicer software in DICOM format and processed using the Python programming language (version 3.12.1; Python Software Foundation, Beaverton, OR, USA). Images were resized to 512 × 512 pixels to standardize the input dimensions for the detection model, and the corresponding bounding-box coordinates were adjusted accordingly. Pixel intensity values were normalized before model input. A 3 × 3 Gaussian smoothing filter was applied to reduce image noise.

To improve model generalization, training images were augmented using random horizontal flipping, limited random rotation, brightness and contrast adjustment, and zoom variation. Augmentation was applied only to the training subset and was not applied to the validation or independent test subsets.

ROI detection using Faster R-CNN

The complete workflow of the proposed method is shown in Figure 3.

Faster R-CNN was used to identify candidate regions that could contain carious lesions. The model was trained using the expert-prepared bounding-box annotations. Each preprocessed bitewing radiograph was first processed by the convolutional backbone to extract image features. These feature maps were then passed to the region proposal network (RPN), which generated multiple candidate bounding boxes with objectness scores indicating the likelihood that each box contained a relevant region. The proposed boxes were subsequently refined by the detection head, which adjusted the box coordinates and assigned a confidence score to each detected region.

Because several candidate boxes could overlap around the same lesion, non-maximum suppression (NMS) was applied using an intersection-over-union threshold of 0.50. The remaining bounding boxes were considered detected ROIs and were forwarded to the classification stage.

Table 1. ICCMS-based radiographic caries categories and ROI distribution used for the three-class and six-class classification tasks

Three-class classification	Definition	Three-class ROIs, n (%)	Six-class classification	Radiographic definition	Six-class ROIs, n (%)
RA	Initial stage	3707 (89.8)	RA1	Radiolucency in the outer half of enamel	1570 (38.1)
			RA2	Radiolucency in the inner half of enamel, with or without extension to the enamel–dentin junction	1635 (39.6)
			RA3	Radiolucency limited to the outer one-third of dentin	502 (12.2)
RB	Moderate stage	194 (4.7)	RB4	Radiolucency reaching the middle one-third of dentin	194 (4.7)
RC	Extensive stage	225 (5.5)	RC5	Radiolucency reaching the inner one-third of dentin; clinically non-cavitated	68 (1.6)
			RC6	Radiolucency extending into or approaching the pulp; clinically cavitated	157 (3.8)

Percentages were calculated using the total number of annotated carious ROIs included in the study (n = 4126).
ICCMS: International Caries Classification and Management System; ROI: region of interest.

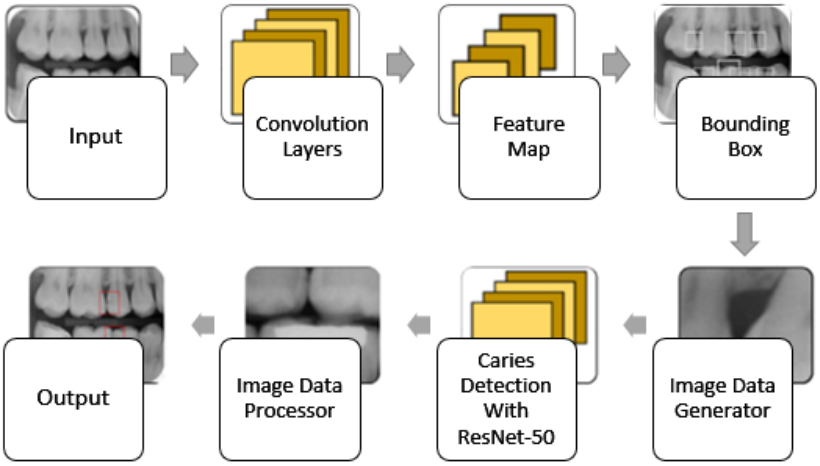

Figure 3. Workflow of the proposed two-stage deep learning framework for automated caries assessment on bitewing radiographs. The pipeline included image preprocessing, ROI localization using Faster R-CNN, and severity classification using ResNet-50 architectures.

Severity classification using ResNet-50

Detected ROIs were cropped from the preprocessed radiographs and resized to the input dimensions required by the ResNet-50 classifier. Two ResNet-50 variants were evaluated: ResNet-50 V1 and ResNet-50 V2. ResNet-50 V1 uses the original post-activation residual block design, whereas ResNet-50 V2 uses a pre-activation design in which batch normalization and activation precede convolutional operations. The pre-activation structure may improve gradient propagation and optimization stability in deeper networks.

Both ResNet-50 variants were initialized with ImageNet pre-trained weights. The original ImageNet classification head was removed and replaced with a new task-specific fully connected layer followed by a softmax output layer. The number of output neurons was adjusted according to the classification task: three output neurons for the three-class analysis and six output neurons for the six-class analysis. Therefore, separate models were trained for each combination of architecture and classification task, resulting in four classification models: ResNet-50 V1 for the three-class task, ResNet-50 V2 for the three-class task, ResNet-50 V1 for the six-class task, and ResNet-50 V2 for the six-class task.

Training and validation procedure

The deep learning pipeline was implemented in Python using the TensorFlow/Keras deep learning framework. Image preprocessing was performed using the Python Imaging Library. Computational analyses were performed on a workstation equipped with an NVIDIA GeForce RTX 3080 graphics processing unit with 10 GB of video memory and 32 GB of random-access memory.

The ResNet-50 classifiers were trained using categorical cross-entropy loss and the Adam optimizer with an initial learning rate of 0.001. The batch size was set to 16, and each model was trained for a maximum of 100 epochs. During training, model performance was monitored on the validation set. Early stopping was applied based on validation loss with a patience of 10 epochs to reduce overfitting.

Performance evaluation

Final model performance was evaluated using the independent test set, which remained unseen during training and validation. Performance was assessed using accuracy, precision (positive predictive value; PPV), sensitivity (recall), specificity, F1-score, and mean average precision. For multi-class classification, class-wise metrics were calculated using a one-versus-rest approach.

For each class, true positives (TP) were defined as ROIs correctly classified as the target class. False positives (FP) were ROIs incorrectly classified as the target class. False negatives (FN) were ROIs belonging to the target class but classified as another class or missed by the detection stage. True negatives (TN) were ROIs correctly classified as not belonging to the target class.

The following formulas were used:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision (Positive Predictive Value) = TP / (TP + FP)

Sensitivity (Recall) = TP / (TP + FN)

Specificity = TN / (TN + FP)

F1-score = 2 × Precision × Recall / (Precision + Recall)

Table 2. Overall performance metrics of ResNet-50 V1 and ResNet-50 V2 in the three-class and six-class ICCMS-based caries classification tasks. Accuracy, precision, recall, specificity, F1-score, and mean average precision are reported for each model.

Model Variant	Classification	Accuracy	Precision	Recall	Specificity	F1-Score	mAP
ResNet-50 V2	6-Classes	0.84	0.60	0.55	0.93	0.57	0.44
	3-Classes	0.88	0.68	0.68	0.96	0.68	0.52
ResNet-50 V1	6-Classes	0.61	0.42	0.31	0.92	0.35	0.25
	3-Classes	0.92	0.70	0.67	0.96	0.68	0.52

mAP:Mean Average Precision

Precision–recall (PR) curves were generated for the six-class classification task to evaluate how precision and recall changed across different decision thresholds. For each diagnostic class, average precision (AP) was calculated from the corresponding PR curve and used as a summary measure of class-specific detection/classification performance. AP reflects the model's ability to maintain high precision while also identifying a high proportion of true lesions across threshold values.

Mean average precision (mAP) was then calculated by averaging the AP values across all diagnostic classes:

$mAP = 1/n \sum AP_k$

Where AP_k represents the average precision for class k, and n represents the total number of diagnostic classes. In this study, mAP was used because it provides a more informative summary of multi-class performance than accuracy alone, particularly when the dataset contains uneven class distributions.

Because the dataset was imbalanced, model performance was interpreted using both overall and per-class metrics. Particular attention was given to precision, recall, and F1-score for less frequent classes, especially RC5 and RC6. In imbalanced datasets, overall accuracy may remain high even when the model performs poorly in minority classes. Recall is important because it reflects the proportion of true lesions in a class that were correctly identified, while precision reflects how many model predictions for that class were actually correct. F1-score was therefore useful as a balanced measure when both missed lesions and false-positive predictions were clinically relevant.

Results

Overall performance in three-class and six-class classifications

In the three-class classification task, ResNet-50 V1 achieved the highest overall accuracy, with an accuracy of 0.92, precision of 0.70, recall of 0.67, specificity of 0.96, F1-score of 0.68, and mAP of 0.52. ResNet-50 V2 achieved an accuracy of 0.88, precision of 0.68, recall of 0.68, specificity of 0.96, F1-score of 0.68, and mAP of 0.52 (Table 2).

In the six-class classification task, ResNet-50 V2 showed better overall performance than ResNet-50 V1. ResNet-50 V2 achieved an accuracy of 0.84, precision of 0.60, recall of 0.55, specificity of 0.93, F1-score of 0.57, and mAP of 0.44. By comparison, ResNet-50 V1 achieved an accuracy of 0.61, precision of 0.42, recall of 0.31, specificity of 0.92, F1-score of 0.35, and mAP of 0.25.

Overall, both architectures performed better in the three-class task than in the six-class task.

Per-class performance in six-class classification

Per-class performance metrics for the six-class classification task are presented in Table 3. Because the dataset had an uneven class distribution, precision, recall, and F1-score were particularly important for interpreting class-specific performance. Precision reflected the reliability of the model's positive predictions for each category, recall reflected the ability to identify true lesions within each category, and F1-score provided a balanced summary of both measures.

ResNet-50 V2 achieved higher precision, recall, and F1-score than ResNet-50 V1 across all six diagnostic categories. For the RA1 category, ResNet-50 V2 achieved a precision of 0.82, recall of 0.85, and F1-score of 0.83, compared with 0.65, 0.58, and 0.61 for ResNet-50 V1, respectively. For the RA2 caries category, ResNet-50 V2 achieved a precision of 0.68, recall of 0.72, and F1-score of 0.70, compared with 0.51, 0.44, and 0.47 for ResNet-50 V1.

Performance was lower for the more advanced categories, especially RC5 and RC6. In the RC5 category, ResNet-50 V2 achieved a precision of 0.45, recall of 0.38, and F1-score of 0.41, whereas ResNet-50 V1 achieved 0.28, 0.15, and 0.20, respectively. In the RC6 category, ResNet-50 V2 achieved a precision of 0.58, recall of 0.40, and F1-score of 0.47, whereas ResNet-50 V1 achieved 0.35, 0.20, and 0.25, respectively. Specificity remained relatively high across diagnostic categories for both architectures, ranging from 0.82 to 0.97.

Table 3. Per-class perrmafonce metrics of ResNet-50 V1 and ResNet-50 V2 in the six-class ICCMS-based classification task. Accuracy, precision, recall, specificity, and F1-score are reported separately for each diagnostic category.

Caries Stage (ICCMS)	Model	Accuracy	Precision	Recall	Specificity	F1-Score
RA1	V1	0.81	0.65	0.58	0.88	0.61
	V2	0.89	0.82	0.85	0.91	0.83
RA2	V1	0.74	0.51	0.44	0.82	0.47
	V2	0.82	0.68	0.72	0.86	0.70
RA3	V1	0.68	0.38	0.29	0.89	0.33
	V2	0.79	0.55	0.51	0.92	0.53
RB4	V1	0.65	0.32	0.22	0.91	0.26
	V2	0.76	0.52	0.48	0.94	0.50
RC5	V1	0.62	0.28	0.15	0.94	0.20
	V2	0.73	0.45	0.38	0.95	0.41
RC6	V1	0.64	0.35	0.20	0.96	0.25
	V2	0.75	0.58	0.40	0.97	0.47

Precision–recall analysis

Precision–recall curves for the six-class classification task are presented in Figure 4. These curves show the trade-off between precision and recall across decision thresholds for each model. ResNet-50 V2 showed higher overall mAP than ResNet-50 V1 in the six-class task, consistent with the quantitative results reported in Table 1.

The precision–recall curves also supported the per-class findings in Table 2. Lower precision–recall performance was observed particularly for the advanced categories, RC5 and RC6, where both models showed reduced recall and F1-scores. This reduction was more evident for ResNet-50 V1, which showed lower mAP and weaker class-specific performance than ResNet-50 V2 in the six-class task.

Discussion

This study developed and internally evaluated a two-stage deep learning framework for automated localization and radiographic severity classification of dental caries on bitewing radiographs. The first stage used Faster R-CNN to localize candidate regions of interest, and the second stage used ResNet-50 V1 and ResNet-50 V2 to classify the detected regions according to ICCMS-based radiographic severity categories. Model performance was evaluated in two classification settings: a detailed six-class task, which separated initial enamel lesions and different levels of dentinal involvement, and a broader three-class task, in which these detailed categories were collapsed into wider severity groups. The main finding was that the proposed framework showed preliminary feasibility for automated caries assessment, particularly in the

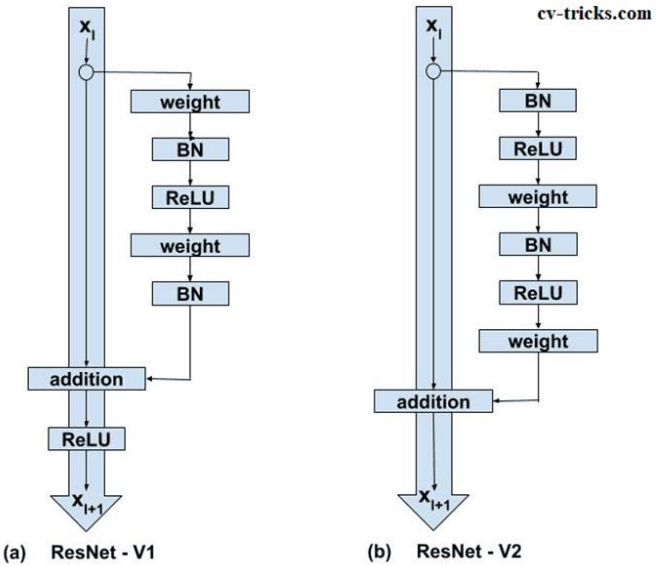

Figure 4. Schematic architecture of the ResNet-50 V1 and ResNet-50 V2 models used for caries severity classification. ResNet-50 V1 uses a post-activation residual block design, whereas ResNet-50 V2 uses a pre-activation residual block design.

broader three-class setting and for initial-stage lesions. However, performance decreased when the task required finer six-class discrimination, especially for advanced categories with fewer annotated examples.

In the three-class classification task, ResNet-50 V1 achieved the highest overall accuracy, with an accuracy of 0.92, precision of 0.70, recall of 0.67, specificity of 0.96, F1-score of 0.68, and mAP of 0.52. ResNet-50 V2 showed slightly lower accuracy in this task, at 0.88, but comparable recall, specificity, F1-score, and mAP values.

These findings suggest that when adjacent radiographic stages are collapsed into broader severity groups, both architectures can capture clinically relevant differences consistently. In contrast, the six-class task was more challenging. In this setting, ResNet-50 V2 outperformed ResNet-50 V1 across the overall metrics, achieving an accuracy of 0.84, precision of 0.60, recall of 0.55, specificity of 0.93, F1-score of 0.57, and mAP of 0.44, compared with 0.61, 0.42, 0.31, 0.92, 0.35, and 0.25, respectively, for ResNet-50 V1. This difference may be related to the pre-activation residual design of ResNet-50 V2, in which batch normalization and activation are applied before convolutional layers. This architecture may facilitate gradient propagation and improve optimization stability, particularly when the classification task requires separation of subtle radiographic differences between adjacent caries stages (18).

The lower performance in the six-class task can be explained by the greater difficulty of separating adjacent radiographic severity categories. Although the broader three-class grouping may be more practical for clinical decision-making, the six-class task required the model to distinguish subtle differences between enamel involvement, early dentinal extension, and deeper dentinal radiolucencies. These distinctions may be difficult even in expert radiographic interpretation. In addition, the dataset was imbalanced, with fewer examples in advanced categories such as RC5 and RC6.

In this setting, overall accuracy may provide an incomplete impression of model performance because it can be influenced mainly by the more frequent classes. Therefore, per-class precision, recall, F1-score, and mAP are more informative. Precision shows how reliable the model's positive predictions are for each category, recall shows how many true lesions in each category were detected, F1-score balances precision and recall, and mAP summarizes precision–recall performance across decision thresholds. Together, these metrics provide a clearer assessment of model performance across all severity categories, especially for underrepresented classes.

The six-class per-category analysis showed that ResNet-50 V2 performed better than ResNet-50 V1 across all diagnostic categories, although performance varied by caries stage. ResNet-50 V2 showed its strongest performance in the early categories, with F1-scores of 0.83 for RA1 and 0.70 for RA2. Performance was lower for the advanced categories RC5 and RC6, where ResNet-50 V2 achieved F1-scores of 0.41 and 0.47, respectively. Recall was also reduced in these categories, indicating that some advanced lesions were not correctly classified by the model. Although specificity remained high across categories, the lower recall and F1-scores for RC5 and RC6 most likely reflect the smaller number of annotated examples available for these classes rather than the inherent clinical difficulty of detecting advanced caries.

The reduced recall observed in some diagnostic categories has important clinical implications because false-negative and false-positive classifications can affect patient management. False-negative results may delay necessary intervention, whereas false-positive detections may contribute to unnecessary restorative treatment. For this reason, the model should be considered a decision-support tool rather than a replacement for clinical judgment. This is particularly important for advanced lesions, where missed detection may delay treatment, and for early lesions, where overdiagnosis may lead to unnecessary operative intervention.

The present results are consistent with previous studies showing that deep learning can assist in caries detection and radiographic assessment on bitewing images. Estai et al. (19) evaluated a deep learning system for automatic detection of proximal caries on bitewing radiographs and reported favorable diagnostic performance. Similarly, Cantu et al. (20) showed that deep learning models could detect caries lesions with different radiographic extensions. In the present study, the broader three-class task also showed acceptable performance, particularly with ResNet-50 V1, which achieved an accuracy of 0.92, precision of 0.70, recall of 0.67, F1-score of 0.68, and mAP of 0.52. These findings support the general observation that deep learning models can identify radiographic patterns related to dental caries, especially when severity categories are grouped into broader diagnostic levels.

Some previous studies have reported higher diagnostic values than those observed in the present six-class task. Bayraktar and Ayan (21) reported an accuracy

of 0.94, sensitivity of 0.72, specificity of 0.98, and AUC of 0.87 for interproximal caries detection using a deep CNN. Bayrakdar et al. (10) also reported high performance for caries detection and segmentation on bitewing radiographs. Kunt et al. (22), using a substantially larger dataset of 3,989 bitewing radiographs, reported an accuracy of approximately 0.83 for automatic caries detection. In comparison, the present ResNet-50 V2 model achieved an accuracy of 0.84 in the six-class task, but its recall, F1-score, and mAP were 0.55, 0.57, and 0.44, respectively. Therefore, although the overall accuracy was comparable with some previous studies, the lower recall, F1-score, and mAP indicate that detailed severity classification remained more challenging.

The differences between the present results and previous reports are likely related to differences in task definition, dataset size, class distribution, annotation strategy, and evaluation metrics. Many earlier studies focused on binary caries detection, lesion presence versus absence, or segmentation of carious regions (10, 19, 21, 22). In contrast, the present study classified carious radiographic findings into ICCMS-based severity categories, including initial enamel lesions and different levels of dentinal involvement. This required the model to distinguish not only whether caries was present, but also the estimated radiographic depth of the lesion. Because adjacent ICCMS categories may show subtle radiographic differences and because advanced categories were less represented in the dataset, lower recall, F1-score, and mAP in the six-class task are expected compared with binary or detection-focused models.

The severity-classification aspect of the present study is most closely related to the work of Panyarak et al. (13), who evaluated deep learning for caries classification based on the ICCMS radiographic scoring system. Similar to that study, the present work emphasized clinically meaningful severity grading rather than simple caries detection. Both studies highlight that ICCMS-based radiographic grading is methodologically demanding because it requires discrimination between lesion depths that may be subtle on bitewing radiographs, particularly when image quality, projection geometry, or proximal overlap affects lesion visibility.

The better performance of ResNet-50 V2 in the six-class task may be related to differences in model architecture. Compared with ResNet-50 V1, ResNet-50 V2 is designed to train more smoothly in deeper networks, which may help the model distinguish finer radiographic differences between caries categories . However, because no formal statistical comparison was performed between the two architectures, this explanation should be considered only a possible reason for the observed difference.

The clinical value of the proposed framework lies in its potential to support more standardized radiographic interpretation. A system that localizes suspected lesions and suggests a radiographic severity category may help clinicians review bitewing radiographs more systematically, particularly in busy clinical settings or when early lesions are difficult to detect. The use of ICCMS-based categories also improves clinical interpretability compared with models that simply classify images as carious or non-carious. Nevertheless, the current results do not establish clinical effectiveness. Before such a system can be recommended for routine use, it must be externally validated across different institutions, radiographic devices, patient populations, and acquisition protocols.

A major consideration in interpreting the present results is the nature of the reference standard. The labels were based on adjudicated expert interpretation of bitewing radiographs rather than histological, operative, or longitudinal confirmation. This approach is practical and common in retrospective radiographic artificial intelligence studies, but it does not represent definitive biological ground truth. Bitewing radiographs have inherent limitations in estimating true lesion depth and activity, particularly in the presence of overlapping proximal surfaces, variations in projection geometry, or subtle enamel lesions (8). Therefore, the reported metrics should be interpreted as agreement with expert radiographic assessment rather than true diagnostic accuracy against a definitive disease standard.

The present study has several limitations that should be considered when interpreting the findings. First, the dataset was obtained from a single center, and the independent test set was relatively small, which may limit the generalizability of the results. Second, the class distribution was imbalanced, particularly for advanced lesions, and this likely contributed to the lower recall and F1-scores observed in minority categories such as RC5 and RC6. Third, the study did not include external validation, prospective clinical testing, or a reader-performance comparison with clinicians. Therefore, future studies should use larger, more balanced, multi-center datasets with standardized annotation protocols and external validation. In addition, future reader-performance studies are needed to determine whether artificial intelligence-assisted interpretation improves

clinician accuracy and decision-making compared with unaided assessment.

Conclusions

The Faster R-CNN and ResNet-50 framework showed preliminary feasibility for automated caries localization and ICCMS-based severity classification on bitewing radiographs. ResNet-50 V2 performed better in the six-class task, while broader three-class grouping produced more stable results. However, lower performance in advanced classes indicates that larger, balanced, multi-center datasets and external validation are needed before clinical implementation.

Acknowledgements

We sincerely thank Mona Abbasi Pirouz for the valuable support and cooperation in this research and assistance with the software section.

Conflict of interest

The authors declare no conflict of interest.

Author contributions

F.M. and M.J.M. contributed to conceptualization, supervision, project administration, original draft preparation, and review and editing. A.M.M. contributed to original draft preparation, review and editing, data curation, and software. A.A.T., H.Z., and F.B.A. contributed to data curation, formal analysis, review and editing, and software. R.I. contributed to investigation, original draft preparation, and review and editing. Z.K. contributed to conceptualization, original draft preparation, and review and editing. S.S. and S.M. contributed to methodology, investigation, and review and editing.

Ethical considerations

The study was approved by the ethics committee of Mashhad University of Medical Sciences (IR.MUMS.DENTISTRY.REC.1403.050). All radiographs were de-identified before image processing and model development.

Funding

This study was supported by Mashhad University of Medical Sciences under grant number 4012410. The article was derived from a student thesis.

References

1. Zhang Q, Wu X, Wang L, Huang J. Self-equilibrium segmentation of near-infrared images of dental microcracks. Infrared Phys Technol 2024;138:105246.
2. Abdelaziz M. Detection, Diagnosis, and Monitoring of Early Caries: The Future of Individualized Dental Care. Diagnostics (Basel) 2023;13(24):3649.
3. Ahrari F, Akbari M, Mohammadi M, Fallahrastegar A, Najafi MN. The validity of laser fluorescence (LF) and near-infrared reflection (NIRR) in detecting early proximal cavities. Clin Oral Investig 2021;25(8):4817–4824.
4. Fernández CE, González-Cabezas C, Fontana M. Minimum intervention dentistry in the US: an update from a cariology perspective. Br Dent J 2020;229(7):483–486.
5. Wang F, Liu J-y, Wang X-c, Wang Y. Experimental investigation on the caries characteristic of dental tissues by photothermal radiometry scanning imaging. Infrared Phys Technol 2018;89:64–71.
6. Gomez Rossi J, Rojas-Perilla N, Krois J, Schwendicke F. Cost-effectiveness of Artificial Intelligence as a Decision-Support System Applied to the Detection and Grading of Melanoma, Dental Caries, and Diabetic Retinopathy. JAMA Netw Open 2022;5(3):e220269.
7. Ghaffari M, Zhu Y, Shrestha A. A review of advancements of artificial intelligence in dentistry. Dent Rev 2024;4(2):100081.
8. Kühnisch J, Aps JK, Splieth C, Lussi A, Jablonski-Momeni A, Mendes FM, et al. ORCA-EFCD consensus report on clinical recommendation for caries diagnosis. Paper I: caries lesion detection and depth assessment. Clin Oral Investig 2024;28(4):227.
9. Horvath KSH, Gjerdet NR, Shi XQ. Advancements in Caries Diagnostics Using Bitewing Radiography: A Systematic Review of Deep Learning Approaches. Caries Res 2026;60(2):127–151.
10. Bayrakdar IS, Orhan K, Akarsu S, Çelik Ö, Atasoy S, Pekince A, et al. Deep-learning approach for caries detection and segmentation on dental bitewing radiographs. Oral Radiol 2022;38(4):468–479.
11. Chen X, Guo J, Ye J, Zhang M, Liang Y. Detection of Proximal Caries Lesions on Bitewing Radiographs Using Deep Learning Method. Caries Res 2022;56(5-6):455–463.
12. Pérez de Frutos J, Holden Helland R, Desai S, Nymoen LC, Langø T, Remman T, et al. AI-Dentify: deep learning for proximal caries detection on bitewing x-ray - HUNT4 Oral Health Study. BMC Oral Health 2024;24(1):344.
13. Panyarak W, Wantanajittikul K, Suttapak W, Charuakkra A, Prapayasatok S. Feasibility of deep learning for dental caries classification in bitewing radiographs based on the ICCMS™ radiographic scoring system. Oral Surg Oral Med Oral Pathol Oral Radiol 2023;135(2):272–281.
14. Ammar N, Kühnisch J. Diagnostic performance of artificial intelligence-aided caries detection on bitewing radiographs: a systematic review and meta-analysis. Jpn Dent Sci Rev 2024;60:128–136.

15. Lee S, Oh SI, Jo J, Kang S, Shin Y, Park JW. Deep learning for early dental caries detection in bitewing radiographs. Sci Rep 2021;11(1):16807.
16. Alsolamy M, Nadeem F, Azhari AA, Ahmed WM. Automated Detection, Localization, and Severity Assessment of Proximal Dental Caries from Bitewing Radiographs Using Deep Learning. Diagnostics (Basel) 2025;15(7).
17. Ismail AI, Pitts NB, Tellez M, Banerjee A, Deery C, Douglas G, et al. The International Caries Classification and Management System (ICCMS™) An Example of a Caries Management Pathway. BMC Oral Health 2015;15 Suppl 1(Suppl 1):S9.
18. He F, Liu T, Tao D. Why ResNet Works? Residuals Generalize. IEEE Trans Neural Netw Learn Syst 2020;31(12):5349–5362.
19. Estai M, Tennant M, Gebauer D, Brostek A, Vignarajan J, Mehdizadeh M, et al. Evaluation of a deep learning system for automatic detection of proximal surface dental caries on bitewing radiographs. Oral Surg Oral Med Oral Pathol Oral Radiol 2022;134(2):262–270.
20. Cantu AG, Gehrung S, Krois J, Chaurasia A, Rossi JG, Gaudin R, et al. Detecting caries lesions of different radiographic extension on bitewings using deep learning. J Dent 2020;100:103425.
21. Bayraktar Y, Ayan E. Diagnosis of interproximal caries lesions with deep convolutional neural network in digital bitewing radiographs. Clin Oral Investig 2022;26(1):623–632.
22. Kunt L, Kybic J, Nagyová V, Tichý A. Automatic caries detection in bitewing radiographs: part I-deep learning. Clin Oral Investig 2023;27(12):7463–7471.