BMC Oral Health

RESEARCH

Open Access

Application of Mask R-CNN for automatic recognition of teeth and caries in cone-beam computerized tomography

Yujie Ma[1†], Maged Ali Al-Aroomi[1†], Yutian Zheng[2], Wenjie Ren[1], Peixuan Liu[1], Qing Wu[3], Ye Liang[1*] and Canhua Jiang[1*]

Abstract

Objectives Deep convolutional neural networks (CNNs) are advancing rapidly in medical research, demonstrating promising results in diagnosis and prediction within radiology and pathology. This study evaluates the efficacy of deep learning algorithms for detecting and diagnosing dental caries using cone-beam computed tomography (CBCT) with the Mask R-CNN architecture while comparing various hyperparameters to enhance detection.

Materials and methods A total of 2,128 CBCT images were divided into training and validation and test datasets in a 7:1:1 ratio. For the verification of tooth recognition, the data from the validation set were randomly selected for analysis. Three groups of Mask R-CNN networks were compared: A scratch-trained baseline using randomly initialized weights (R group); A transfer learning approach with models pre-trained on COCO for object detection (C group); A variant pre-trained on ImageNetfor for object detection (I group). All configurations maintained identical hyperparameter settings to ensure fair comparison. The deep learning model used ResNet-50 as the backbone network and was trained to 300epoch respectively. We assessed training loss, detection and training times, diagnostic accuracy, specificity, positive and negative predictive values, and coverage precision to compare performance across the groups.

Results Transfer learning significantly reduced training times compared to non-transfer learning approach ($p < 0.05$). The average detection time for group R was 0.269 ± 0.176 s, whereas groups I (0.323 ± 0.196 s) and C (0.346 ± 0.195 s) exhibited significantly longer detection times ($p < 0.05$). C-group, trained for 200 epochs, achieved a mean average precision (mAP) of 81.095, outperforming all other groups. The mAP for caries recognition in group R, trained for 300 epochs, was 53.328, with detection times under 0.5 s. Overall, C-group demonstrated significantly higher average precision across all epochs (100, 200, and 300) ($p < 0.05$).

Conclusion Neural networks pre-trained with COCO transfer learning exhibit superior annotation accuracy compared to those pre-trained with ImageNet. This suggests that COCO's diverse and richly annotated images offer more relevant features for detecting dental structures and carious lesions. Furthermore, employing ResNet-50

[†]Yujie Ma and Maged Ali Al-Aroomi contributed equally to this work and are co-first authors.

*Correspondence:
Ye Liang
liangye@csu.edu.cn
Canhua Jiang
canhuaj@csu.edu.cn
Full list of author information is available at the end of the article

as the backbone architecture enhances the detection of teeth and carious regions, achieving significant improvements with just 200 training epochs, potentially increasing the efficiency of clinical image interpretation.

Keywords CBCT, Image recognition, Caries, Mask R-CNN, Deep learning

Introduction

Dental caries represent the initial stage of tooth lesions, which, if left untreated, can progressively damage the pulp and tooth structure, leading to pulpitis, periapical inflammation, and, ultimately, tooth loss [1]. In some cases, it may even cause osteomyelitis of the jaw [2, 3]. The World Health Organization has recently ranked caries, alongside tumors and cardiovascular diseases, as one of the top three diseases requiring prevention and treatment [4]. Early and accurate detection of caries lesions is essential in guiding clinical planning by determining the appropriate treatment. Therefore, timely and precise diagnosis is critical to successful treatment outcomes [5].

Traditionally, clinicians have diagnosed caries through percussion, dental probing, and visual examination [6]. However, caries may be too small in some instances that they cannot be visualized or diagnosed without radiographic imaging assistance. In this regard, panoramic, periapical, and bitewing radiography are regularly practiced [7, 8]. Despite their widespread use, detecting caries through conventional radiographs remains challenging. The main limitation of traditional radiography lies in its 2D representation of caries lesions, which are inherently 3D structures. This 2D limitation can result in the loss of critical information, making it harder to assess the full extent of the caries. The images are also susceptible to image noise, superimposition of anatomical structures, and distortion and have low image resolution [9].

In contrast, cone-beam computed tomography (CBCT) offers a detailed 3D view of the teeth, providing more precise information about the severity of caries and facilitating improved treatment planning. Studies have shown that CBCT significantly enhances the detection of tooth root canal spaces and periapical areas, making it more effective than conventional imaging methods for evaluating dental infections and pathologies [10, 11]. This improvement allows for better identification of periapical bone lesions, thereby enhancing diagnostic accuracy, treatment planning, and overall prognostic outcomes. As access to CBCT increases, dental practices are gradually transitioning from 2 to 3D imaging [11, 12]. However, CBCT is not specifically designed for diagnosing dental caries. Its sensitivity in detecting caries is lower than that 2D imaging, although its negative predictive value is comparable [13]. Recent advances in deep learning for CBCT image analysis have significantly improved its ability to detect early-stage dental caries, boosting its accuracy, sensitivity, specificity, and classification capabilities [14]. Nonetheless, interpreting CBCT images still requires considerable time and expertise, relying heavily on the clinician's experience and attention to detail. Despite its advantages, CBCT has certain limitations [15, 16].

Recently, there has been considerable innovation in artificial intelligence for healthcare, which can also aid digital dentistry and telemedicine [17]. Convolutional neural networks (CNNs) perform exceptionally in object detection and classification. By learning directly from raw input data, CNNs classify images without requiring manual feature extraction. Region-based convolutional neural networks (R-CNNs) have been developed explicitly for object detection, enabling automatic identification and annotation of target objects (regions of interest) [18, 19]. The R-CNN was later improved to Faster R-CNN, which offers greater efficiency. Based on Faster R-CNN, the Mask R-CNN method was developed to detect targets in images and provide high-quality segmentation results [20, 21]. Recently, Mask R-CNN has been applied in various fields, such as target extraction [22], pathological cell identification [23], and crop identification [24]. It has also been utilized in more specialized areas, including oral pathology, for conditions like herpes labialis, aphthous ulcers [25], and periapical pathosis, helping to identify the causative tooth [26].

Mask R-CNN has been increasingly popular in CBCT image analysis for tasks like dental anomaly detection due to its ability to perform precise pixel-level segmentation. However, applying Mask R-CNN to CBCT images presents several challenges, such as low resolution, image noise, and the complexity of dental structures, which can affect model performance [27]. Additionally, the variability in tooth anatomy also requires advanced preprocessing to enhance segmentation accuracy. To address these challenges, transfer learning is commonly used, where a pre-trained model is fine-tuned for specific tasks like analyzing CBCT images. Transfer learning can enhance model performance, especially when labeled data is limited, by leveraging knowledge from large, diverse datasets [28]. However, the effectiveness of transfer learning depends on the similarity between the pre-training and target datasets. In dental imaging, differences between the source data (e.g., general medical imaging) and CBCT images may affect model accuracy [29].

To the best of the authors' knowledge, there are currently no studies investigating the impact of various hyperparameters on the accuracy of deep learning algorithms for detecting carious lesions in CBCT images. This study aims to explore the application of the Mask R-CNN algorithm for identifying teeth and carious lesions in CBCT images, thereby facilitating the diagnostic process for clinicians and reducing the rates of underdiagnosis. We established training sets using CBCT images to evaluate the effectiveness of the Mask R-CNN algorithm in detecting teeth and caries. Additionally, we compared various hyperparameters to identify more efficient methods that could enhance caries detection in CBCT imaging.

Materials and methods

Image acquisition, preprocessing, and training set construction

The institutional ethics committee of Xiangya Hospital of Central South University approved the present research (approval number 2022020716). The CBCT images were taken for orthodontics, implant placement, or evaluation of impacted teeth or paranasal sinuses between January 2018 and December 2021. Inclusion criteria for the dataset included the presence of one or more carious teeth detected in the CBCT images and the existence of permanent dentition. Exclusion criteria included the presence of artifacts caused by restoration, filling materials, orthodontic appliances, titanium plates, etc., that hinder accurate caries diagnosis or poor quality of the CBCT images, non-odontogenic lesions, the occurrence of caries after root canal treatment, and situations where more than half of the crown was missing. Additionally, poor quality of the CBCT images, characterized by severe distortion, blurriness, or noise, also led to exclusion. The analysis involved examining CBCT images acquired using the I-CAT Imaging System (KaVo Dental GmbH, Biberach, Germany). The imaging settings were configured as follows: 120 kV, 5 mA, a field of view measuring 16 cm × 13 cm, an exposure time of 17.8 s, a voxel size of 0.25 mm, and a slice thickness of 0.25 mm, resulting in a resolution of 640 × 640 pixels. The basic process of deep learning algorithm training is shown in Fig. 1.

The selected DICOM files are imported into the E-3D system, and after reconstructing multi-layer images, the patient's positioning direction is set. To enhance tissue contrast in CBCT images, a default window level range of −5 Hu to 1991 Hu with a window width of 1996 Hu is applied. Minor adjustments are made for some patients with bone density variations, with fluctuations not exceeding 500. In total, 2,128 CBCT images were labeled by 2 oral and maxillofacial radiologists with 10 years of experience to create the reference standard. The CBCT's

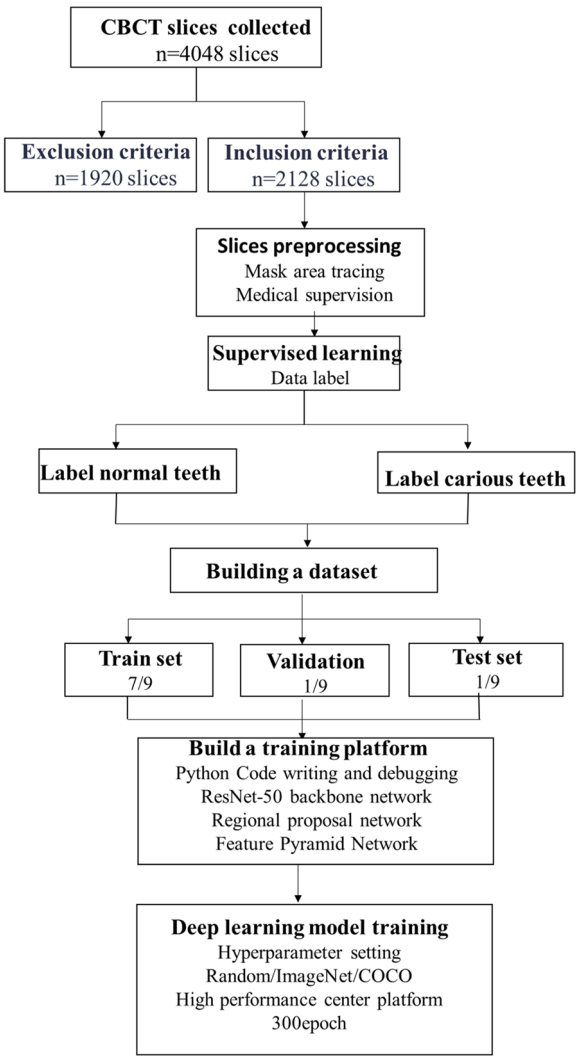


Fig. 1 A diagram illustrating the experimental design and basic process of deep learning algorithm training experiment

MPR interface displays images in sagittal, coronal, and axial planes. The images were categorized as dental and non-dental caries cases. Based on WHO diagnostic criteria for caries [30], simple caries appears as small, arc-shaped concave defects in the enamel or cementum; moderate caries extends into superficial dentin with similar concave defects; and severe caries penetrate deeper dentin layers, presenting more pronounced defects. Carious lesions thus appear as low-density, concave images in CBCT. The intra-examiner reliability test results showed a Kappa coefficient of 0.86 (range: 0.81–0.89), indicating high consistency by the same examiner, while the inter-examiner Kappa coefficient was 0.79 (range: 0.72–0.83), demonstrating good consistency between different examiners. After segmenting target tooth #28, isolated pixels

are removed, region-growing is applied to attach neighboring pixels, forming a growth region, while non-neighboring floating noise is excluded. A source mask (green #28, Fig. 2, A-C and target mask (yellow #28, (Fig. 2, D-F) are set, and selecting the target tooth initiates calculations to generate a new mask (yellow) with reduced noise. Remaining irrelevant areas, such as adjacent tooth structures and jawbone sections, are removed using the Multiple Slice Edit tool, with initial anchor points set on the axial plane and further adjustments made in the sagittal and coronal planes. For the carious tooth, a carious region mask is created based on the refined. The Multiple Slice Edit tool fills the defect layer by layer across the views, producing carious teeth (Fig. 2, G-I). Boolean subtraction of mask target health tooth part from carious tooth yields carious region (Fig. 2, J-L).

The dataset ultimately included 2,128 images, comprising 237 dental caries images and 498 non-dental caries images, which were then divided into training, validation, and test sets in a 7:1:1 ratio. We trained the deep learning model using the cross-entropy loss function [31] and tracked its performance during training and validation with TensorBoard [32]. For the verification of tooth recognition, the data from the validation set were randomly selected for analysis. Finally, this data was provided as input to deep CNNs, which were then trained (Fig. 3).

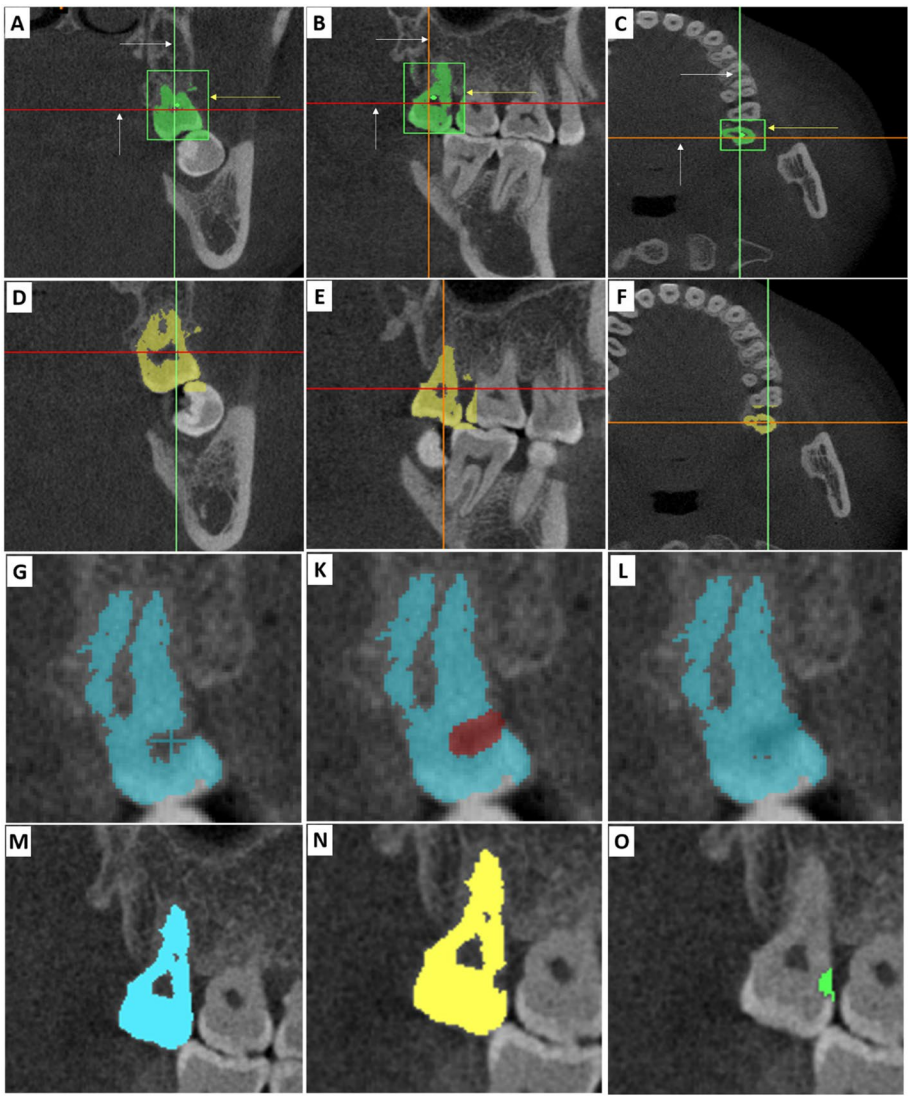


Fig. 2 The white arrow shows the positioning line, and the yellow arrow marks the cropping box **a**–**c**. The region-growing method removes floating pixels and creates the target mask (**d**–**f**). The multi-layer brush fills the target tooth across planes (**G**–**I**). Boolean operations isolate the carious region (M: tooth and caries; N: tooth only; O: caries only)

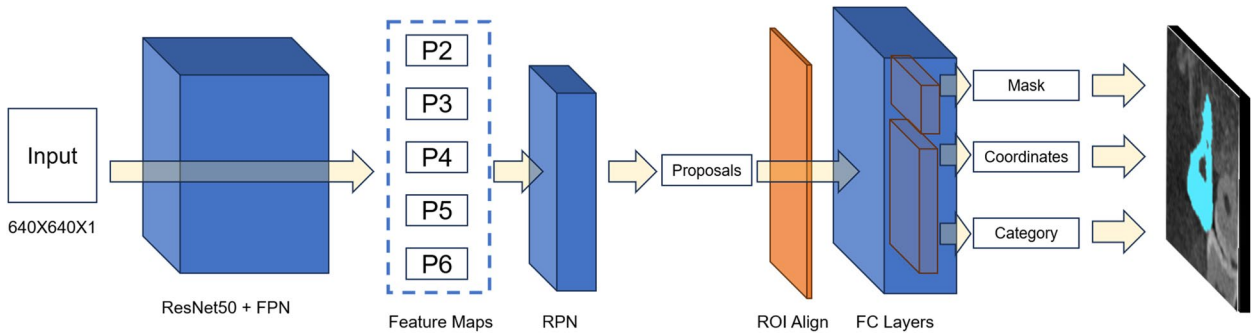


Fig. 3 Overall architecture of the Mask R-CNN model– Input image is processed to generate feature maps (P2-P6), proposals, and aligned regions, producing object masks, coordinates, and categories

Mask R-CNN training parameter setting and study grouping

This study employed the Mask R-CNN model by He et al. [20], using a ResNet-50 backbone with a Feature Pyramid Network (FPN) to detect teeth and caries. Feature extraction was performed by the Region Proposal Network (RPN), with RoIAlign converting proposed regions into fixed-size feature maps. Each Region of Interest (ROI) was segmented, with masks generated by a fully convolutional network (FCN) [33]. Testing used three types of initial neural network values: A scratch-trained baseline using randomly initialized weights (R group),A transfer learning approach with models pre-trained on COCO (https://github.com/matterport/Mask_RCNN/releases/download/v2.0/mask_rcnn_coco.h5).

for object detection (C group) [34], and A variant pre-trained on ImageNet.

(https://github.com/fchollet/deep-learning-models/releases/download/v0.2/resnet50_weights_tf_dim_ordering_tf_kernels_notop.h5) for object detection (I group) [35]. The model's fully connected layer was adapted to classify three categories—background, teeth, and caries. Lower layers captured general features (e.g., shape), while higher layers extracted semantic features (e.g., category), with the final layer refining tooth and caries segmentation and localization.

The Mask R-CNN model was implemented using TensorFlow and Keras within a high-performance computational environment designed to handle resource-intensive deep learning tasks. The system was equipped with GPU acceleration provided by an NVIDIA Tesla V100, featuring 32 GB of GDDR6 dedicated memory, which delivers robust computational power. The CPU configuration included a 4-socket Intel Xeon Gold 5218 processor array, offering a total of 110 compute cores, ensuring efficient multi-threaded processing. The virtualization environment was built using the ESXi operating system, with CentOS 7 serving as the underlying OS. The system was configured with 230 GB of RAM, supporting effective resource management and smooth execution of memory-demanding tasks. The software stack was built on Python 3.6, ensuring compatibility with the necessary libraries and frameworks for model training and evaluation. For DICOM processing, the pydicom 2.2.2 library was utilized to standardize the parsing of DICOM data, enabling seamless handling of medical imaging datasets and facilitating the integration of medical images into the deep learning pipeline.

Training was conducted on a Tesla V100 GPU platform with 2.3 GHz CPUs, 230 GB of memory, and CentOS7. System performance was independently assessed, followed by a clinical study after confirming the metrics met the required standards. Groups C and I were pre-training for 100 epochs, fine-tuning at a 0.01 learning rate for 300 epochs to ensure stable convergence. To investigate the impact of initial network values on recognition accuracy, tests were conducted on C and I groups with transfer learning and R group without [36]. The model was trained for 300 epochs, with sampling at epochs 100, 200, and 300, to analyze the relationship between training duration and recognition performance (Table 1). For each group, loss, detection, and training times, mean average precision (mAP), and coverage precision were calculated to compare performance across groups.

Table 1 Experimental grouping of the deep learning model

Group	Database	Type of Learning	Epoch
Group R	Random training data	Nontransfer learning	100
	Random training data	Nontransfer learning	200
	Random training data	Nontransfer learning	300
Group I	ImageNet database	Transfer learning	100
	ImageNet database	Transfer learning	200
	ImageNet database	Transfer learning	300
	COCO database	Transfer learning	100
Group C	COCO database	Transfer learning	200
	COCO database	Transfer learning	300

Comparison method for loss function variation and time consumption

The loss function measures the discrepancy between predicted and actual values in the model, and it is always non-negative. A smaller loss indicates that the model's predictions are closer to the training results. Mask R-CNN is crucial for separating the mask tasks and class prediction. As the model learns to perform these tasks independently, multitasking losses are defined as follows:

$$L = L_{cls} + L_{bbox} + L_{mask}$$

Here L_{cls} is the classification loss, L_{bbox} is the bounding box loss, and L_{mask} is the mask loss. The mask loss is calculated as the average binary cross-entropy loss for each ROI linked to the true bounding box, focusing only on the final pixel mask branch. A non-competitive relationship network is employed in the loss function to prevent interference between pixels.

The training set is divided into 132 batches, with all batches trained in each of the 300 epochs. The training time is measured from the start of each epoch to its completion, including loss evaluation. The detection time is recorded from when the image and neural network begin processing to when the detection result is completed. The mean and standard deviation of training and detection time are calculated separately to compare the learning speeds of the three model groups.

Calculation of recognition precision

This study evaluated the deep learning model using metrics such as precision, recall, positive predictive value, negative predictive value, average precision (AP), and precision-recall (PR) curves. The intersection-over-union (IoU) threshold, denoted as α, defined the metrics: a true positive (TP) occurs when IoU $\geq \alpha$, a false positive (FP) when IoU $< \alpha$, and a false negative (FN) when ground truth is missed (IoU $< \alpha$). When calculating the IOU, the clinician's evaluation of the caries-affected area is thoroughly incorporated. For instance, consider the depicted tooth: the purple region delineated by the clinician represents the caries extent determined from clinical data, referred to as the Ground Truth. The green region, by contrast, signifies the caries extent as predicted by the deep learning model (Supplemental Fig. 1). Based on the notations mentioned above, precision (P) and recall (R) are calculated as follows:

$$recall = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

The AP metric balances both since a single P or R value does not fully assess model performance. AP@α represents the area under the precision-recall curve (AUC-PR) at IoU threshold α, with a larger area indicating higher mean average precision (mAP):

$$AP@\alpha = \int_0^1 P\,(r)\,dr$$

Due to the non-monotonic nature of the PR curve, AP is computed using the widely used 11-point interpolation method [37].

Coverage precision comparison method

Coverage precision is calculated using the same method as mAP. It is defined for each input image, with each image corresponding to an mAP value that reflects accuracy in recognizing specific regions. Coverage refers to the overlap between the area depicted in the training set and the area recognized by the model, assessed under specified IoU thresholds. In this evaluation, defined positives are teeth and caries manually annotated on CBCT images, while negatives indicate no diagnosis. The TP, FP, and FN are defined as previously described.

For coverage comparison, following literature conventions, the IoU is set at 50, 75, and at 5% intervals from 50 to 95, referred to as AP@50, AP@75, and AP@50:5:95, respectively. AP@50 indicates that the overlap between the machine's recognized area and the original depiction exceeds 50%, while AP@75 indicates more than 75%. AP@50:5:95 represents the mean mAP calculated at 50%, 55%, 60%, and up to 95%. Since the original depiction includes both teeth and caries, this coverage comparison assesses the model's overall ability to identify both, reflecting a more comprehensive accuracy in depiction consistency.

Verification of Clinical Model Detection Efficiency

Three highly experience radiologists diagnosed 56 randomly selected CBCT images containing teeth and caries under three conditions: without AI assistance, with Faster R-CNN (detection box), and with Mask R-CNN (detection box and boundary box). The detection time for each image was recorded in seconds per image.

Statistical analysis

Data were analyzed using IBM SPSS Statistics for Windows, version 21.0 (IBM Corp). Measurement data were presented as mean ± standard deviation. ANOVA was employed for multiple C-groupomparisons of training

and testing times, and Dunnett's T3 was used for interC-groupomparisons when significant differences were found. A t-test was applied to assess group differences in AP after verifying the homogeneity of variance. Statistical significance was determined at $\alpha = 0.05$, with significance indicated by $P < 0.05$. Since the detection efficiency data for the three clinical models were non-normally distributed, a non-parametric Kruskal–Wallis test was conducted. When the asymptotic significance (P) was less than 0.001, significant differences were observed among the three groups. Post-hoc pairwise comparisons were performed using Mann–Whitney U test.

Results

The loss functions and algorithm time consumption

Figure 4 illustrates the changes in the loss function during the training of the Mask R-CNN using the ResNet-50 deep learning model. In Fig. 4A, the training loss shows a steady decline. Figure 4B displays the total loss variation on the validation set, which fluctuates significantly but still trends downward. Both figures present the loss function variations for the R, I, and C groups. However, the loss values for the R-group are not shown, as they remained infinite throughout the training process.

The training process for R group used non-transfer learning, with an average training time of 1001.478 ± 130.920 s per epoch. In comparison, I and C groups, which utilized transfer learning, had a training time of 895.613 ± 65.740 s and 897.867 ± 229.608 s, respectively. An ANOVA test revealed a significant difference in training times between transfer learning and non-transfer learning, with transfer learning being faster ($p < 0.05$). During the detection process, the average detection time per image for R-group was 0.269 ± 0.176 s. When compared to I group (0.323 ± 0.196 s) and C group (0.346 ± 0.195 s), showing that the detection times for groups I and C were significantly longer than for R-group ($p < 0.05$). Table 2 presents the results for detection and training times.

Typical recognition image results

During the model training, the area surrounded by the red curve contains unmarked tooth shapes and root canals, classifying it as a true positive (Fig. 5A). The red curve also encloses the green-marked region, accurately identifying the teeth and carious areas, making it another true positive (Fig. 5B). In contrast, the yellow highlighted area does not contain maxillary and mandibular teeth, but the red curve outlines the teeth and root canals, classifying this as a false positive (Fig. 5C). The yellow-highlighted area corresponds to the hard bone of the maxilla and mandible, while the red curve outlines the teeth's cross-sectional shape, which is also a false positive (Fig. 5D). Figure 5E shows a premolar in the yellow-highlighted area, where the tooth's cross-section was not recognized, classifying it as a false negative. Finally, Fig. 5F displays a molar in the yellow-highlighted area, where neither the tooth nor the root canal shapes were recognized, also classified as a false negative.

Recognition of precision outcomes

Figure 6A shows the PR plots for I-group at 100, 200, and 300 epochs during tooth recognition, while Fig. 6B presents the PR curve plots for C- group at the same epochs. In Fig. 6C, the mAP calculated using the 11-point interpolation method reveals that C-group at 200 epochs achieved an mAP of 81.095, surpassing all other groups. In contrast, Fig. 6D illustrates the PR plots for I group

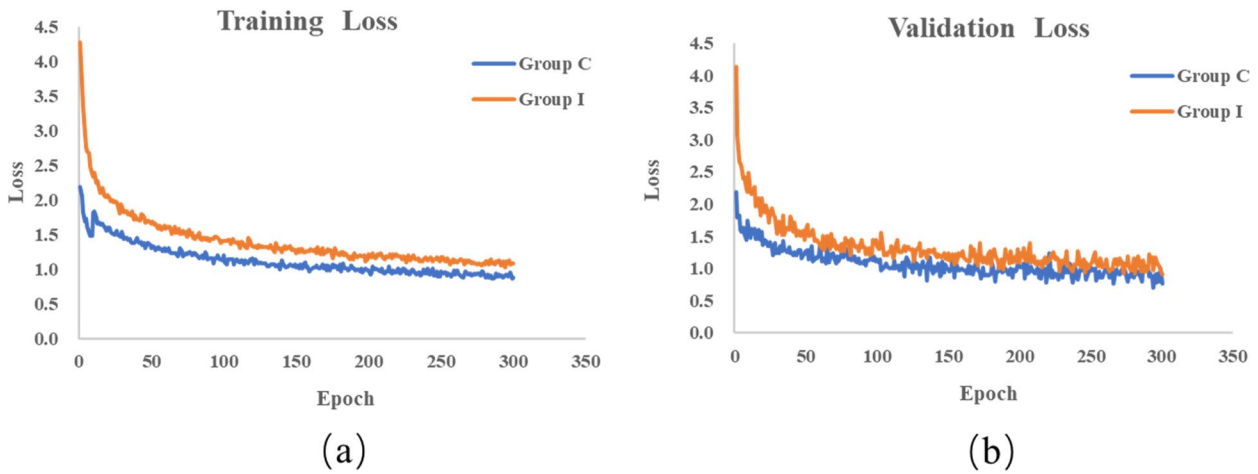


Fig. 4 Training process of the Mask R-CNN deep learning network: **a** Changes in total loss values for the training set using initial values from ImageNet and COCO. **b** Changes in total loss values for the validation set using initial values from ImageNet and COCO

Table 2 Training time per epoch and detection time per image

Group	Method	Training time (s/per epoch)	Detection time (s/per image)
Group R	Random	1001.478 ± 130.920	0.269 ± 0.176
Group I	ImageNet	895.613 ± 65.740*	0.323 ± 0.196*
Group C	COCO	897.867 ± 229.608*	0.346 ± 0.195*

NOTE: * represents a statistically significant difference from group R by ANOVA variance test, $p < 0.05$

during caries recognition at 100, 200, and 300 epochs, while Fig. 6 E displays the PR curve plots for C group in the same context. Figure 6 F shows the mAP calculated through the 11-point interpolation method, with C group at 300 epochs attaining an mAP of 53.328, which was higher than all other groups.

Coverage precision outcomes

As shown in Fig. 7, at the coverage of AP@50, C-group at 100 epochs achieved an AP of 68.969, significantly higher than I-group's 57.522 ($p < 0.05$). By epochs 200 and 300, the precision value for I group improved, and no significant difference was observed between groups I and C at 200 epochs. At AP@75 coverage, the AP values for both groups at 100 epochs were lower and not significantly different. At 200 epochs, C-group's AP increased by 10 points from 100 epochs, while I-group increased by 6 points. C-group reached an AP of 55.722 at 200 epochs, compared to I group's 41.377 ($p < 0.05$), indicating that C group performed significantly better. At 300 epochs, both groups saw a 3-point increase in AP compared to 200 epochs. Statistical analysis revealed that C group's AP@75 value was 58.396, significantly higher than I-group's 44.941 at 300 epochs ($p < 0.05$). Overall, at AP@50:5:95, C-group consistently outperformed I-group at all epochs (100, 200, and 300), with significant differences ($p < 0.05$).

Clinical model detection efficiency verification

Figure 8 illustrates the distribution of detection times (seconds per image) for three methods: without AI

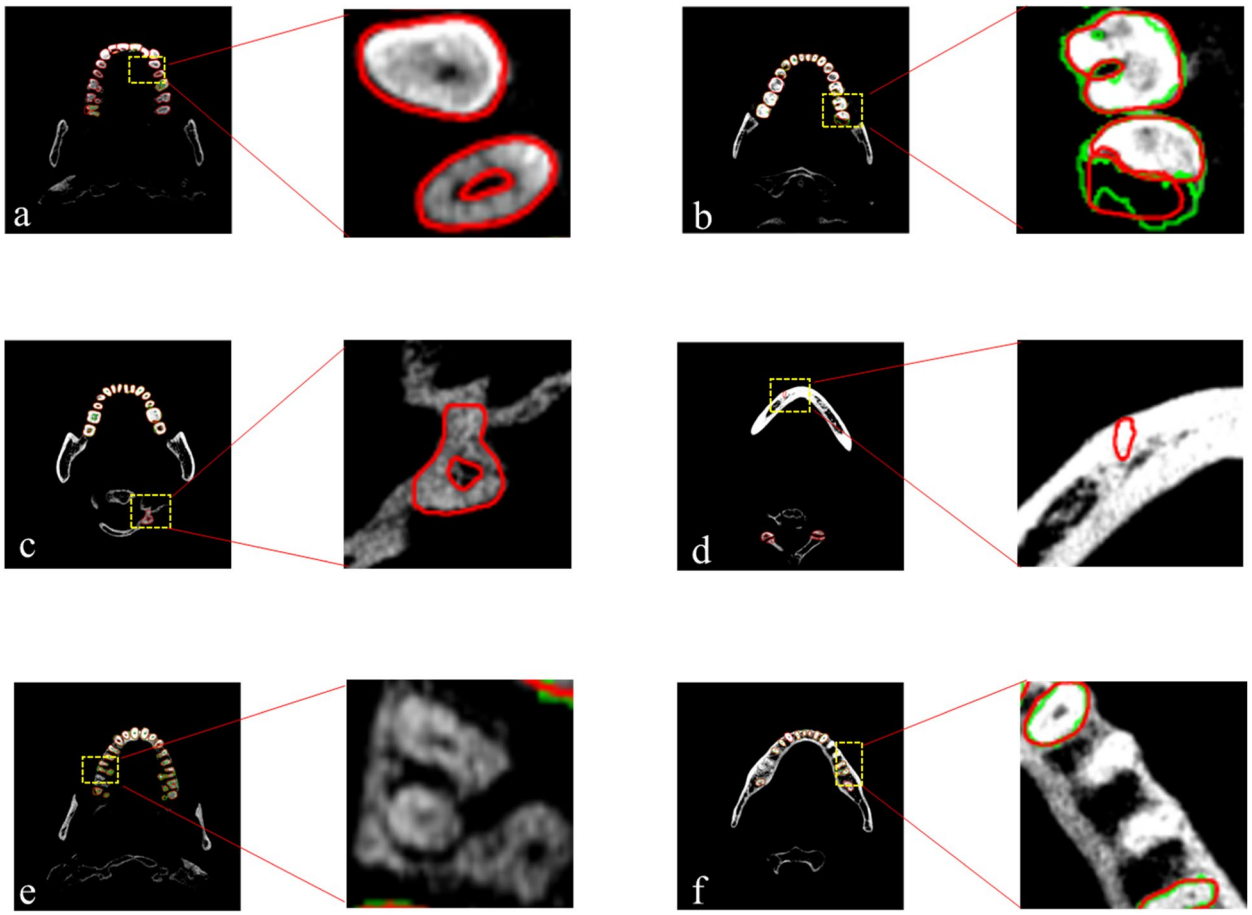


Fig. 5 Typical images recognized by the Mask R-CNN deep learning model: True positive results, correctly identifying the root canal morphology of teeth, intact teeth, and decayed teeth at an IOU > 0.05 (**a**-**b**). False positive results, incorrectly classifying the dense and loose bone of the hyoid and mandibular bones as teeth and dental root canals (**c**-**d**). True negative results, where the root and crown morphologies were not correctly identified (**e**–**f**)

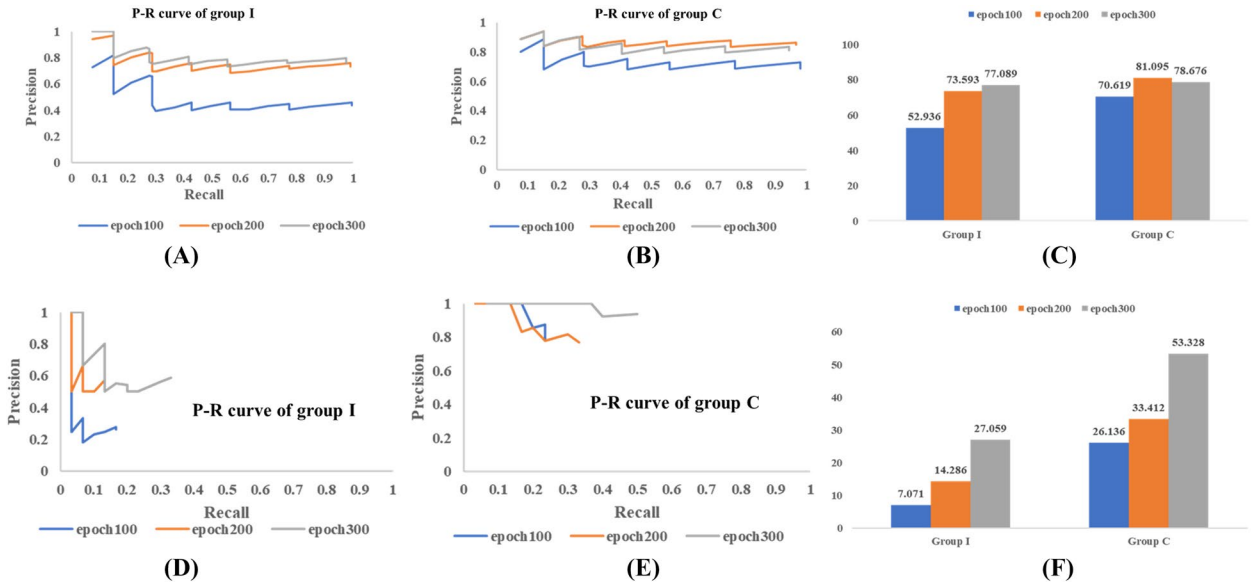


Fig. 6 Recognition precision comparison for established teeth detected as positive: (**a**) Precision-recall curve with the ImageNet database, (**b**) variation in mean average precision values, (**c**) Precision-recall curve with the COCO database. Recognition precision comparison for established caries detected as positive: (**d**) Precision-recall curve with the ImageNet database, (**e**) variation in mean average precision values, (**f**) Precision-recall curve with the COCO database

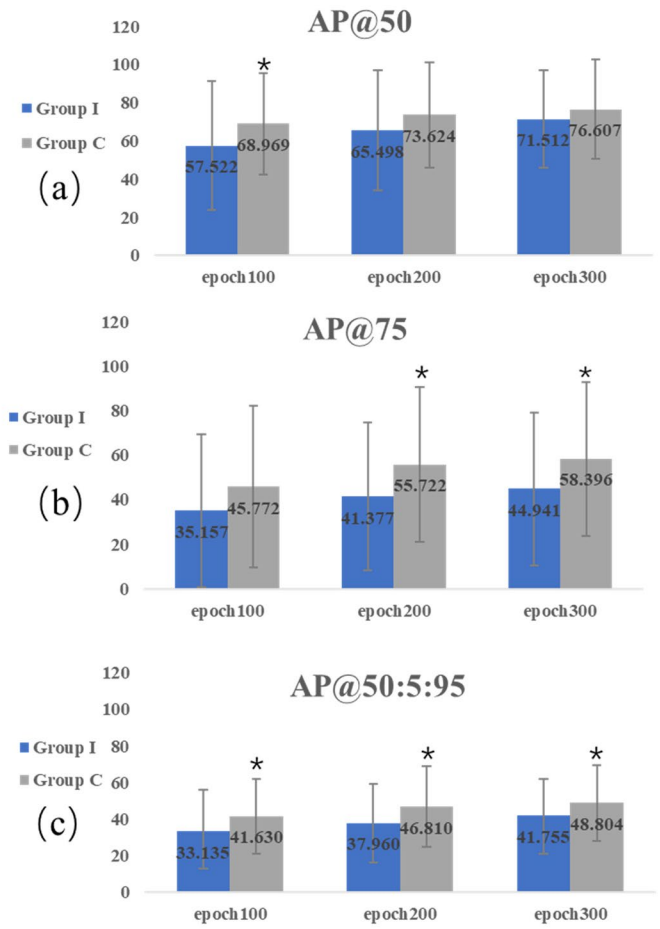
assistance, with Faster R-CNN, and with Mask R-CNN. The detection time in the no-AI group spans a wide range (16.390–23.450 s), with a median of 23.535 s (IQR = 19.520–21.270 s). AI assistance led to significant improvements in detection time, with the Faster R-CNN group showing a median of 9.235 s (IQR = 8.760–9.623 s, mean = 9.133 s). The Mask R-CNN group further reduced the median to 4.350 s (IQR = 4.193–4.725 s, mean = 4.415 s), exhibiting the lowest variability (standard deviation = 0.356 s). The Kruskal–Wallis test revealed highly significant differences between the three groups (H (2) = 148.458, $p < 0.001$), and post-hoc Mann–Whitney U tests confirmed significant pairwise differences ($p < 0.001$).

Discussion

Transfer learning effectively enhances machine learning efficiency, allowing neural networks to acquire relevant skills quickly. This study compares transfer learning with non-transfer learning for identifying dental and carious features in CBCT images. The results indicate significant differences between the two approaches. As training iterations increased, both neural networks utilizing ImageNet and COCO through transfer learning demonstrated a general decline in the loss function, suggesting improved learning. After 100 epochs, the loss reduction began to plateau, indicating model convergence. In contrast, the non-transfer learning group (R group), which began with random values, consistently faced infinite loss values, highlighting the difficulties in recognizing patterns without prior knowledge. Although the group R neural networks were initialized similarly to groups I and C, substantial training would be required to yield results, necessitating significant computational resources [38].

Comparing the loss values between the transfer learning groups (I and C), we found that C group consistently exhibited lower loss values in both training and validation sets. This suggests higher learning efficiency, enabling faster dental and carious area identification. The training set showed a steady decline in loss, while the validation set fluctuated, consistent with standard neural network training principles. Both sets demonstrated generalizability and robust performance [39–41]. The diverse object recognition capabilities of the network parameters significantly reduced the time needed to identify patterns in CBCT images, emphasizing the importance of prioritizing transfer learning in developing neural networks for image recognition.

Significant differences were observed between R, I, and C groups during training, with R group requiring considerably more training time. R group struggled to recognize typical features of teeth or caries, indicating it was still searching for characteristic patterns and lacked a clear direction for stochastic gradient descent, which likely contributed to its extended training duration [42, 43]. While R group had shorter detection times, this was likely due to its inability to identify typical features,



NOTE: * represents statistically significant difference from group I for the same epoch by paired t test

Fig. 7 Precision of tooth and caries coverage: (**a**) Average precision AP values at coverage of AP@50, (**b**) AP values at coverage of AP@75, (c) AP values at coverage of AP@50:5:95

leading to reduced model complexity. Conversely, there were no significant differences in training or detection times between I and C groups, suggesting that the neural network source had minimal impact on these metrics. Therefore, the detection efficiency of I and C groups is fundamentally similar.

Our study carefully integrated the dentist's expertise, ensuring that all caries were identified and confirmed using clinical characteristics. Furthermore, the calculation of the IOU fully accounted for the actual caries regions annotated by the dentist. Additionally, the average detection time for an experienced oral radiologist in the control group (C) was 3.007 ± 0.709 s, significantly longer than the machine learning times, underscoring the clinical potential of the Mask R-CNN model for identifying teeth and carious lesions. To assess algorithm recognition accuracy, we used clinical judgments as a reference standard to determine whether the teeth in the images were accurately identified, reducing bias from unmarked teeth in the training set. The C-group model at 200 epochs achieved an mAP of 80.974, outperforming the other groups. However, the mAP for the C group at 300 epochs dropped by 1.865 compared to 200 epochs, likely due to overfitting, which negatively impacted generalization and increased error. Thus, training for 200 epochs was determined to be the most effective under the study's conditions.

The Faster R-CNN [44] and Mask R-CNN [20], developed by Kaiming in 2015 and 2017, are advanced deep-learning models that combine residual networks, CNNs, and task-specific processors. These models excel in image recognition and instance segmentation. Mask R-CNN builds on Faster R-CNN by adding pixel-level mask prediction through RoIAlign, which improves alignment for more accurate results. It also introduces a dual-task approach with a mask loss function, enhancing both recognition accuracy and processing speed, making it a powerful tool in computer vision [21]. When compared

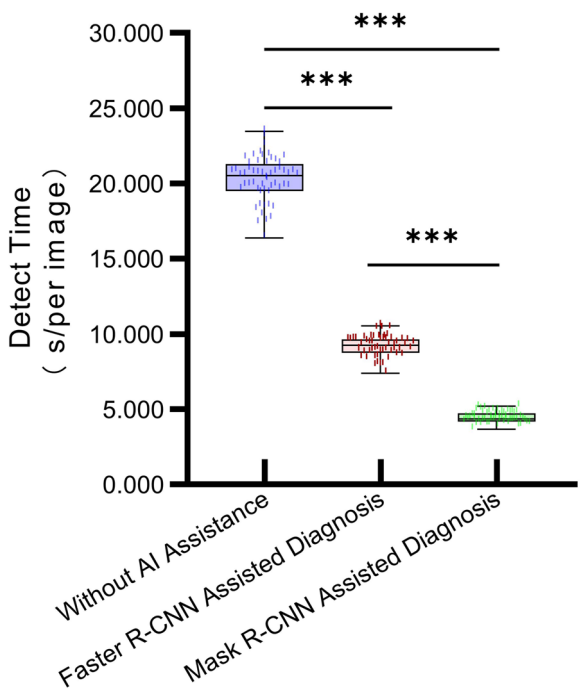


Fig. 8 Comparison of CBCT tooth and caries image detection times with and without AI assistance, and with Faster R-CNN and Mask R-CNN assistance (s/per image)

to other models, the trade-offs generally involve speed and accuracy. Mask R-CNN provides high segmentation accuracy but is slower than You Only Look Once (YOLO) [45], which is better suited for real-time detection, though it may sacrifice some precision. For tasks requiring pixel-level segmentation, such as lesion boundary detection, Mask R-CNN are preferred over Faster R-CNN and YOLO. Faster R-CNN and YOLO, on the other hand, are more suitable for tasks that require rapid detection of multiple objects. While Mask R-CNN is more complex, it is also capable of handling these tasks. Ultimately, the choice between these models depends on whether speed or accuracy is prioritized, and a comparative evaluation can help determine the most suitable option for a specific use case.

Several studies have compared the accuracy, sensitivity, and specificity of using CNN networks and deep learning for detecting dental caries in dental radiographs. Previous research has reported CNN-based accuracy for caries detection ranging from 82 to 97%. The variations in these results can be attributed to several factors, including differences in the datasets used. Datasets can differ in size, diversity, quality, and distribution, all of which can influence the outcomes [46–48]. The selection of layers, types of layers, number of neurons, learning rate, and other model parameters can also significantly impact the results. Regarding caries recognition in our study, C group at 200 epochs achieved an mAP of 18.182, outperforming other groups and exceeding the maximum mAP of I-group by 3.030. Nonetheless, both groups had mAP values below 20, which is insufficient for clinical identification of caries, likely due to the dataset's low proportion of caries samples [49]. The parameters from C group at 300 epochs proved superior for caries detection. Identifying carious lesions is challenging due to their open nature, making them hard to distinguish from the background; this complexity supports the notion that recognizing such variable areas is a strength of deep learning [50]. The Mask R-CNN algorithm's pixel-level recognition capability enhances its performance in these scenarios [20]. Despite the C group at 200 epochs achieving an mAP of 81.095 for tooth recognition, caries detection still needs improvement. The variability of caries regions, their small image representation, and insufficient interlayer connections may limit current recognition capabilities [50]. Future optimization efforts could focus on addressing these limitations.

We extended the training to 800 epochs after normalizing the caries proportion to evaluate whether increasing training epochs could improve model recognition. C group at 300 epochs maintained the highest mAP, with values fluctuating between 44.156 and 50.932 from epochs 400 to 800, indicating a stable convergence state. This suggests that simply increasing training epochs did not lead to further improvements. One possible reason for this lack of improvement could be the imbalance in the ratio between carious and intact teeth, as the dataset used in this study had a much lower ratio of carious to healthy teeth, which may have affected the training outcomes.

Our coverage analysis used annotated masks from the training set as the reference standard. AP@50 indicates coverage where the overlapping area marked by the neural network exceeds 50%, while AP@75 indicates high coverage with overlaps over 75%. AP@50:5:95 represents overall coverage. Our comparative analysis of nine sample pairs revealed differences in coverage at training epochs of 100, 200, and 300. For the low coverage indicator AP@50, C-group scored significantly higher at 100 epochs. However, no significant differences were found between the groups at 200 and 300 epochs. This suggests that C group's neural network quickly adapted to recognizing caries patterns, resulting in better early performance. As training progressed, I group also improved its recognition, but enhancements became limited as both groups approached their recognition thresholds, explaining the lack of statistical differences at higher epochs. Thus, C and I groups achieved good results under the 50% coverage criterion.

Regarding high coverage, the C group model at 200 epochs achieved an accuracy of 55.722, significantly higher than I group, and at 300 epochs reached 58.396, surpassing I group. This indicates that networks using COCO transfer learning align more closely with the annotated masks in the training set, with overall accuracy for C group significantly better than I group, suggesting stronger early performance in transfer learning. We did not analyze coverage above AP@95 due to considerable image noise [51, 52], as even experienced doctors exhibit slight discrepancies in annotations, which do not affect clinical diagnoses.

Several commonly used transfer learning techniques include feature extraction, fine-tuning, and multi-task learning. Feature extraction involves using a pre-trained model on a large dataset as a fixed feature extractor, with the extracted features then being fed into a new classifier. On the other hand, fine-tuning adjusts specific layers of the pre-trained model to adapt it to a new task better. Additionally, multi-task learning allows the model to simultaneously learn multiple related tasks, which can improve its generalization ability and overall performance. Research has shown that transfer learning can outperform traditional machine learning methods, particularly in cases involving small datasets [53]. Accurately segmenting certain dental caries remains a challenge, particularly due to the variability in caries forms (e.g., occlusal, interproximal) and the presence of noise or artifacts in dental radiographs. Pre-trained models like COCO and ImageNet may not capture the specific features needed to detect dental caries. To improve segmentation accuracy, fine-tuning these models with domain-specific datasets, such as annotated dental radiographs, can help the model learn the distinctive features of caries. Data augmentation, transfer learning, and post-processing techniques can also enhance the model's generalizability and performance. Also, hybrid models combining CNNs with traditional image processing techniques could refine segmentation results further. These strategies can help overcome current limitations and improve the robustness of deep learning models in clinical caries detection.

A comparison of detection efficiency between no AI assistance and the two AI algorithms (Faster R-CNN and Mask R-CNN) revealed that AI significantly reduced image processing time, with Mask R-CNN achieving the highest efficiency (an 81% reduction in median time). Additionally, the data distribution for Mask R-CNN was more concentrated (IQR: 4.193–4.725 s), indicating superior stability compared to both the Faster R-CNN group (median = 9.235 s) and the no-AI assistance group (median = 23.535 s). This demonstrates that, while retaining the boundary box detection from Faster R-CNN, the experiment integrated dual-object detection using segmentation masks. This design allows clinicians to simultaneously obtain both spatial localization (boundary box) and morphological details (pixel-level mask) of the target more easily. Although Mask R-CNN shows the best performance, the presence of extreme values suggests that its adaptability in complex scenarios still needs improvement.

This study has its limitations. While CBCT is not commonly used for caries detection, it is increasingly utilized for diagnosing and planning treatment for various dental conditions. As a result, some patients may have incidental findings, such as caries, during CBCT scans. The dataset used in this study had a much lower ratio of carious to healthy teeth, which may have impacted the assessment of CBCT's diagnostic accuracy, as the higher number of healthy teeth could reduce sensitivity. Additionally, this ratio may not fully reflect clinical reality, as caries prevalence varies across different populations and regions. Future research should aim to include a higher proportion of decayed teeth to address these limitations and explore additional methodologies. Moreover, comparative studies between CBCT and 2D imaging are necessary to better understand these imaging technologies'relative advantages and limitations in dental diagnosis.

Conclusion

The research demonstrates that the neural network pre-trained with transfer learning from the COCO dataset achieves better annotation accuracy than the network pre-trained with ImageNet. This suggests that the COCO dataset, with its diverse and richly annotated images, provides more relevant features for detecting dental structures and carious lesions. Additionally, by utilizing ResNet-50 as the backbone architecture, the model can effectively recognize teeth and carious regions after only 200 epochs of training. This approach holds promise for streamlining diagnostic workflows and reducing clinicians'time on image analysis.

Abbreviations

CBCT	Cone-beam computed tomography
CNNs	Convolutional neural networks
R-CNNs	Region-based convolutional neural networks
FPN	Feature Pyramid Network
RPN	Region Proposal Network
ROI	Region of Interest
FCN	Fully convolutional network
mAP	Mean Average Precision
AP	Average precision
PR	Precision-recall
IoU	Intersection-over-union
TP	True positive
FP	False positive
FN	False negative
P	Precision
R	Recall

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12903-025-06293-8.

Supplementary Material 1. Supplemental Fig. 1. Demonstrates an example of Intersection over Union (IOU) calculation.

Acknowledgements
We are grateful to the High-Performance Computing Center of Central South University for assistance with the computations. We thank Professor ShengHui Liao (Digital Medicine and 3D Printing Research Center, Central South University) for providing E-3D software and performing the targeted optimization of the software. We thank Binzhu Wang and Shouhui Huang for their assistance with CBCT data collection.

Authors' contributions
YM: study design, data acquisition, quality control, data analysis, statistical analysis, manuscript preparation, and editing; MAA: manuscript drafting, study design, and critical revision; YZ: quality control and manuscript preparation; WR: statistical analysis and manuscript review; PL, QW: data acquisition and manuscript review; YL, CJ: study concept, critical revision, and overall supervision. All authors reviewed and approved the final version.

Funding
This work was financially supported by grants from the National Natural Science Foundation of China (82271040, 82103473); Outstanding Youth Science Foundation of Hunan Province (2023JJ20095); Key R&D Program of Hunan Province (2022SK2048); Healthcare High-level Talents in Hunan Province (2023018, 2023069); Science and Technology Project of Ningbo City (2023Z202, 2021S084); Research Program of Advanced Interdisciplinary Studies of Central South University (2023QYJC018).

Data availability
All data generated or analyzed during this study are included in this published article.

Declarations

Ethics approval and consent to participate
The institutional ethics committee of Xiangya Hospital of Central South University approved the present research (approval number 2022020716) and also met the ethical principles of the World Medical Association Declaration of Helsinki. All participants provided informed consent to participate in the study.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Oral and Maxillofacial Surgery, Center of Stomatology, Xiangya Hospital, Central South University, Changsha, Hunan Province 410008, China. [2]The College of Mechanical and Electrical Engineering, Central South University, Changsha, Hunan Province, China. [3]High Performance Computing Center, Central South University, Changsha, Hunan Province, China.

Received: 6 November 2024 Accepted: 28 May 2025
Published online: 06 June 2025

References

1. Mathur VP, Dhillon JK. Dental caries: a disease which needs attention. The Indian Journal of Pediatrics. 2018;85:202–6.
2. Liu D, Zhang J, Li T, et al. Chronic osteomyelitis with proliferative periostitis of the mandibular body: report of a case and review of the literature. The Annals of The Royal College of Surgeons of England. 2019;101:328–32.
3. Mardini S, Gohel A. Imaging of odontogenic infections. Radiol Clin North Am. 2018;56:31–44.
4. Benzian H, Watt R, Makino Y, et al. WHO calls to end the global crisis of oral health. The Lancet. 2022;400:1909–10.
5. Baelum V, Heidmann J, Nyvad B. Dental caries paradigms in diagnosis and diagnostic research. Eur J Oral Sci. 2006;114:263–77.
6. Kong C, Davis M, Fong K, Abbott PV. General dentists' use of diagnostic equipment and methods. Aust Endod J. 2017;43:66–72.
7. Chan M, Dadul T, Langlais R, et al. Accuracy of extraoral bite-wing radiography in detecting proximal caries and crestal bone loss. J Am Dent Assoc. 2018;149:51–8.
8. Vandenberghe B, Jacobs R, Bosmans H. Modern dental imaging: a review of the current technology and clinical applications in dental practice. Eur Radiol. 2010;20:2637–55.
9. Kamburoğlu K, Kurt H, Kolsuz E, et al. Occlusal caries depth measurements obtained by five different imaging modalities. J Digit Imaging. 2011;24:804–13.
10. Liang Y, Jiang L, Gao X, et al. Detection and measurement of artificial periapical lesions by cone-beam computed tomography. Int Endod J. 2014;47:332–8.
11. Patel S, Dawood A, Mannocci F, et al. Detection of periapical bone defects in human jaws using cone beam computed tomography and intraoral radiography. Int Endod J. 2009;42:507–15.
12. Takeshita WM, Iwaki LCV, Da Silva MC, Tonin RH. Evaluation of diagnostic accuracy of conventional and digital periapical radiography, panoramic radiography, and cone-beam computed tomography in the assessment of alveolar bone loss. Contemp Clin Dent. 2014;5:318–23.
13. Safi Y, Mahmoudi NS, Aghdasi MM, et al. Diagnostic accuracy of cone beam computed tomography, conventional and digital radiographs in detecting interproximal caries. J Med Life. 2015;8:77.
14. Esmaeilyfard R, Bonyadifard H, Paknahad M. Dental Caries Detection and Classification in CBCT Images Using Deep Learning. Int Dent J. 2024;74:328–34.
15. Hung K, Yeung AWK, Tanaka R, Bornstein MM. Current applications, opportunities, and limitations of AI for 3D imaging in dental research and practice. Int J Environ Res Public Health. 2020;17:4424.
16. Dawasaz AA, Alshahrani I, Yassin SM, et al. Detection of dental caries' and dermatoglyphics' association with relative enamel thickness using CBCT images in Saudi subpopulation: a novel approach. Biomed Res Int. 2021;2021:5550916.
17. Gherlone E, Polizzi E, Tete G, Cappare P. Dentistry and Covid-19 pandemic: Operative indications post-lockdown. New Microbiol. 2021;44:1–11.
18. Huang J, Rathod V, Sun C, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 7310–7311.
19. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. p 580–587.
20. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. 2017. p. 2961–2969.
21. Zu L, Zhao Y, Liu J, et al. Detection and segmentation of mature green tomatoes based on mask R-CNN with automatic image acquisition approach. Sensors. 2021;21:7842.
22. Tiede D, Schwendemann G, Alobaidi A, et al. Mask R-CNN-based building extraction from VHR satellite data in operational humanitarian action: An example related to Covid-19 response in Khartoum, Sudan. Trans GIS. 2021;25:1213–27.
23. Chen J, Zhang B. Segmentation of overlapping cervical cells with mask region convolutional neural network. Comput Math Methods Med. 2021;2021:3890988.
24. Yao N, Ni F, Wu M, et al. Deep learning-based segmentation of peach diseases using convolutional neural network. Front Plant Sci. 2022;13:876357.
25. Anantharaman R, Velazquez M, Lee Y. Utilizing mask R-CNN for detection and segmentation of oral diseases. In: 2018 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE, 2018. p. 2197–2204.
26. Adnan N, Umer F, Malik S, Hussain OA. Multi-model deep learning approach for segmentation of teeth and periapical lesions

on pantomographs. Oral Surg Oral Med Oral Pathol Oral Radiol. 2024;138:196–204.
27. Chen X, Ma N, Xu T, Xu C. Deep learning-based tooth segmentation methods in medical imaging: A review. Proc Inst Mech Eng H. 2024;238:115–31.
28. Tajbakhsh N, Shin JY, Gurudu SR, et al. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Trans Med Imaging. 2016;35:1299–312.
29. Zhou SK, Greenspan H, Davatzikos C, et al. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proc IEEE. 2021;109:820–38.
30. Uribe SE, Innes N, Maldupa I. The global prevalence of early childhood caries: a systematic review with meta-analysis using the WHO diagnostic criteria. Int J Paediatr Dent. 2021;31:817–30.
31. Adler AI, Painsky A. Feature importance in gradient boosting trees with cross-validation feature selection. Entropy. 2022;24:687.
32. Vogelsang DC, Erickson BJ. Magician's corner: 6. TensorFlow and TensorBoard Radiol Artif Intell. 2020;2: e200012.
33. Wang S, Sun G, Zheng B, Du Y. A crop image segmentation and extraction algorithm based on mask RCNN. Entropy. 2021;23:1160.
34. Choi W-H, Choi Y-S. Effective pre-training method and its compositional intelligence for image captioning. Sensors. 2022;22:3433.
35. Sun ED, Dekel R. ImageNet-trained deep neural networks exhibit illusion-like response to the Scintillating grid. J Vis. 2021;21:15.
36. Man CK, Quddus M, Theofilatos A. Transfer learning for spatio-temporal transferability of real-time crash prediction models. Accid Anal Prev. 2022;165: 106511.
37. Padilla R, Netto SL, Da Silva EAB. A survey on performance metrics for object-detection algorithms. In: 2020 international conference on systems, signals and image processing (IWSSIP). IEEE, 2020. p. 237–242.
38. Zhao Z, Lu W, Zeng Z, et al. Self-supervised assisted active learning for skin lesion segmentation. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2022. p. 5043–5046.
39. Liang S, Gu Y. Towards robust and accurate detection of abnormalities in musculoskeletal radiographs with a multi-network model. Sensors. 2020;20:3153.
40. Hassan M, Murtza I, Hira A, et al. Robust spatial fuzzy GMM based MRI segmentation and carotid artery plaque detection in ultrasound images. Comput Methods Programs Biomed. 2019;175:179–92.
41. Masci AL, Menesale EB, Chen W-C, et al. Integration of fluorescence detection and image-based automated counting increases speed, sensitivity, and robustness of plaque assays. Mol Ther Methods Clin Dev. 2019;14:270–4.
42. Hosseini SAH, Yaman B, Moeller S, et al. Dense recurrent neural networks for accelerated MRI: History-cognizant unrolling of optimization algorithms. IEEE J Sel Top Signal Process. 2020;14:1280–91.
43. Igarashi Y, Kondo S, Kida S, et al. Mandibular premolar identification system based on a deep learning model. J Oral Biosci. 2022;64:321–8.
44. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2016;39:1137–49.
45. Jiang P, Ergu D, Liu F, et al. A Review of Yolo algorithm developments. Procedia Comput Sci. 2022;199:1066–73.
46. Mohammad-Rahimi H, Motamedian SR, Rohban MH, et al. Deep learning for caries detection: A systematic review. J Dent. 2022;122: 104115.
47. Moharrami M, Farmer J, Singhal S, et al. Detecting dental caries on oral photographs using artificial intelligence: A systematic review. Oral Dis. 2024;30:1765–83.
48. Tabatabaian F, Vora SR, Mirabbasi S. Applications, functions, and accuracy of artificial intelligence in restorative dentistry: A literature review. J Esthet Restor Dent. 2023;35:842–59.
49. Balki I, Amirabadi A, Levman J, et al. Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. Can Assoc Radiol J. 2019;70:344–53.
50. Geng C, Huang S, Chen S. Recent advances in open set recognition: A survey. IEEE Trans Pattern Anal Mach Intell. 2020;43:3614–31.
51. Bueno MR, Azevedo BC, Estrela C. A critical review of the differential diagnosis of root fracture line in CBCT scans. Braz Dent J. 2021;32:114–28.
52. Jacobson MW, Lehmann M, Huber P, et al. Abbreviated on-treatment CBCT using roughness penalized mono-energization of kV-MV data and a multi-layer MV imager. Phys Med Biol. 2021;66: 135001.
53. Cai C, Wang S, Xu Y, et al. Transfer learning for drug discovery. J Med Chem. 2020;63:8683–94.

Publisher's Note