Clinical Oral Investigations
https://doi.org/10.1007/s00784-023-05335-1

RESEARCH

Automatic caries detection in bitewing radiographs: part I—deep learning

Lukáš Kunt[1] · Jan Kybic[1] · Valéria Nagyová[2] · Antonín Tichý[2]

Received: 30 June 2023 / Accepted: 11 October 2023

Abstract

Objective The aim of this work was to assemble a large annotated dataset of bitewing radiographs and to use convolutional neural networks to automate the detection of dental caries in bitewing radiographs with human-level performance.

Materials and methods A dataset of 3989 bitewing radiographs was created, and 7257 carious lesions were annotated using minimal bounding boxes. The dataset was then divided into 3 parts for the training (70%), validation (15%), and testing (15%) of multiple object detection convolutional neural networks (CNN). The tested CNN architectures included YOLOv5, Faster R-CNN, RetinaNet, and EfficientDet. To further improve the detection performance, model ensembling was used, and nested predictions were removed during post-processing. The models were compared in terms of the F_1 score and average precision (AP) with various thresholds of the intersection over union (IoU).

Results The twelve tested architectures had F_1 scores of 0.72–0.76. Their performance was improved by ensembling which increased the F_1 score to 0.79–0.80. The best-performing ensemble detected caries with the precision of 0.83, recall of 0.77, $F_1 = 0.80$, and AP of 0.86 at IoU=0.5. Small carious lesions were predicted with slightly lower accuracy (AP 0.82) than medium or large lesions (AP 0.88).

Conclusions The trained ensemble of object detection CNNs detected caries with satisfactory accuracy and performed at least as well as experienced dentists (see companion paper, Part II). The performance on small lesions was likely limited by inconsistencies in the training dataset.

Clinical significance Caries can be automatically detected using convolutional neural networks. However, detecting incipient carious lesions remains challenging.

Keywords Dental caries detection · Convolutional neural networks · Ensembling · Bitewing · X-ray images

Introduction

Machine learning and especially neural networks have improved remarkably over the last decade, surpassing human-level performance in many tasks, such as the ImageNet image classification task [6, 10] or breast cancer detection [31]. In dentistry, neural networks are also increasingly used [13]. Among other applications, they have been used to detect dental caries in bitewings, periapical radiographs, orthopantomograms, and photographs [24, 29]. Such a method can work as a second reader, providing an independent image assessment and giving dentists an opportunity to cross-check their decisions. The automatic method could also reduce the probability of caries being overlooked, and it could be used to determine the caries' position for dental records or teaching purposes.

Related work

This work is focused on automatic caries detection in bitewings [1, 5, 7, 9, 23, 25, 27, 28, 33] which has been previously approached in various ways. The first group of methods extracts individual teeth in radiographs and trains a *classifier* deciding whether there is a carious lesion in the tooth or not. Kuang et al. [14] proposed a neural network, which was able to outperform an ordinary dentist by more than 5% while being 6% worse than an experienced one. The best results were obtained by a kernel SVM classifier. Moran et al. [25] used histogram equalization, Otsu's thresholding, and morphological operations to extract individual teeth. The

✉ Jan Kybic
kybic@fel.cvut.cz

Extended author information available on the last page of the article

Published online: 16 November 2023

teeth were assigned to three categories: sound teeth, incipient lesions, and advanced lesions. Using 112 radiographs with 480 annotated teeth and the ResNet and Inception models for classification task, the best achieved tooth-level accuracy was 73.3% [25]. Mao et al. [23] used a similar approach while using tooth images split in mesial and distal halves. On 3716 such images, AlexNet reached a 90.3% accuracy. On 3000 tooth images extracted from periapical radiographs without dental restorations, Lee et al. [16] reached an accuracy of 89% for molars and 82% for both premolars and molars combined, using GoogLeNet and Inception-v3 neural networks.

The task of detecting caries can also be formulated as a *segmentation problem,* producing a binary mask of the caries. Cantu et al. [5] created a large dataset of 3686 bitewing images. Three dentists independently drew a polygonal-shaped mask over caries in each image. The per-pixel performance of a U-Net model with EfficientNet B5 as a backbone outperformed the mean performance of seven dentists in all metrics. Lian et al. [18] achieved an IoU (intersection over union) of 0.785 on panoramic images, while the best-performing dentist achieved an IoU of 0.717. Lee et al. [17] annotated not only caries but also the enamel, dentin, pulp, and gutta-percha restorations and used two independent U-Net models. While the model achieved a relatively modest F_1 score of 0.641, it was shown to help the dentists to improve their sensitivity by 7–10%. A competition in segmenting the bitewing radiographs was organized in 2015 [37]. However, the results on caries detection were poor, with pixelwise $F_1 \approx 12\%$.

The third approach applies *object detection* techniques, yielding a rectangular bounding box for each lesion, without attempting to identify their precise boundaries. This significantly simplifies the annotation effort and also avoids the task of identifying individual teeth [38]. Srivastava et al. [33] trained a fully convolutional neural network with over 100 layers on a dataset containing more than 3000 bitewing radiographs. It produced a pixel mask, which was then post-processed by fitting a minimal bounding rectangle, obtaining $F_1 = 0.7$ at IoU=0.8. In a related work using U-Net and trained on 6000 bitewing X-ray images, $F_1 = 0.61$ was reported [15]. Bayrakdar et al. [1] performed both semantic segmentation and object detection on a dataset of 621 bitewing images, reporting object detection precision (positive predictive value) of 0.78, recall (sensitivity) of 0.77, and F_1 score of 0.78.The model outperformed 2 dentists with 2–3 years of experience while being outperformed by 3 dentists with over 10 years of experience.

Bayraktar et al. [2] used YOLOv3 and a dataset of 1000 bitewing images evaluated in terms of classifying individual 11,521 approximal surfaces (i.e., not images or detections, making the values not directly comparable) of which 1847 were decayed and reported recall of 0.72, precision of 0.86, and specificity of 0.98, corresponding to $F_1 = 0.83$. Using the same architecture and a dataset of 994 images, Panyarak et al. [27] reported recall of 0.67 and precision of 0.75 at IoU = 0.5, with good results for extensive caries but failing to predict enamel caries reliably. In the most recent object detection study by Chen et al. [7], the Faster R-CNN model was trained on 818 labeled bitewing radiographs and achieved the F_1 score of 0.74, outperforming postgraduate students with less than 3 years of experience.

Estai et al. [8] used a two-step approach on 2468 bitewing images: regions of interests (ROI) were detected using Faster R-CNN and then classified as carious or sound by the Inception-ResNet-v2 neural network with $F_1 = 0.87$ (on manually selected ROIs, i.e., not reflecting the ROI detection performance).

Finally, some authors have attempted to *classify* the lesions according to their stage. Panyarak et al. [27, 28] classified the lesions into 4 or 7 classes according to the International Caries Classification and Management System (ICCMS), instead of just two (carious and sound). However, the task seems to be difficult; the classification error reached 0.36 in the 4-class case and 0.42 in the 7-class case. Better results were obtained by Chen et al. [7] who reported a sensitivity of 0.65 for enamel caries (E1/E2), 0.69 for lesions involving the outer third of dentin (D1), and 0.85 for deeper dentin lesions (D2/D3).

Proposed approach

In this work, the problem of dental caries localization was formulated as an object detection and was solved using state-of-the-art object detection CNNs (Sect. 2). The objectives of this work were to assemble a large dataset of bitewing radiographs with annotated carious lesions and to develop an automatic algorithm for caries detection. The method was experimentally evaluated in Sect. 3. For a more extensive performance comparison with 7 additional human annotators, please see the companion manuscript (Part II) [36].

Methods

Data

Recent bitewing X-ray images of adult patients from routinely performed diagnostics scans were retrieved from a hospital information system and anonymized. See Section 5.2 for ethical considerations. The radiographs were acquired using four different intraoral X-ray units, three of which used direct radiography and one employed indirect radiography. Sensor physical dimensions ranged from 31 × 41 mm to 27 × 54 mm. To simplify processing, all images were rescaled to 896 × 1024 pixels, with the wide-sensor

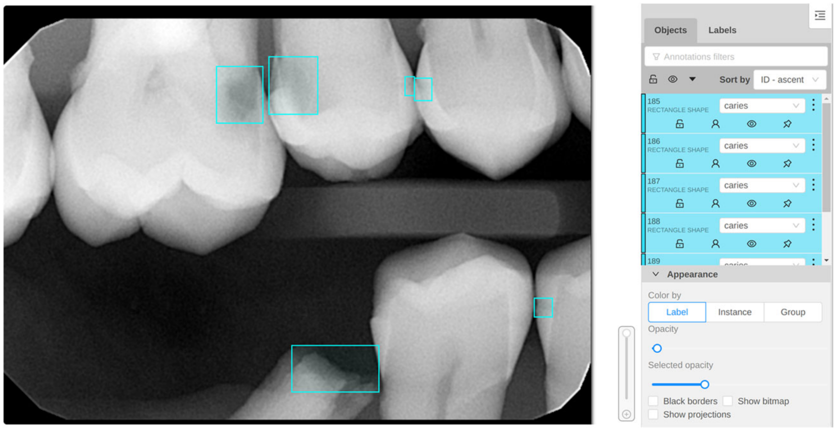

Fig. 1 An example bitewing X-ray image with annotated carious lesions in the CVAT web application

images padded with black horizontal margins to preserve the aspect ratio. Consequently, the image scaling is not identical for all images, but the differences are small and can be handled by the CNN object detectors thanks to the augmentation (Sect. 2.3).

Annotations were performed by a specialist in cariology and operative dentistry with 5 years of experience (A.T.), henceforth denoted as expert E_0. The annotation process was conducted in the Computer Vision Annotation Tool (CVAT).[1] Each lesion was marked by a minimum bounding box, i.e., the smallest possible axis-parallel rectangle containing the entire lesion (Fig. 1).

The dataset was created in a step-wise (bootstrap) fashion. The *initial dataset* consisting of 2599 images with 4575 annotated carious lesions was used for the first training of the YOLOv5 object detection model (Sect. 2.2). The model was then applied on additional 1400 images, its predictions were reviewed by E_0, and bounding boxes were adjusted if needed. Around 20 predictions per 100 images had to be either added or removed to get the same annotation quality as in the initial dataset. The review was approximately twice as fast as annotating the images from scratch.

The YOLOv5 model was retrained on this extended dataset, and its predictions were compared with the ground truth, i.e., annotations by E_0. There were 1543 images with at least a single false-positive or false-negative detection. The annotations for these images were reviewed and possibly corrected by the annotator, taking into account the automatic predictions. Finally, corrupted and low-quality images were removed, yielding a *final dataset* D_0 with 3989 X-ray images and 7257 annotations.

The histograms of the number of lesions per image and the bounding box dimensions in D_0 are shown in Figs. 2 and 3, with numerical values in Table 1. It can be observed that there were approximately 2 lesions per image. Since the annotation bounding boxes were reasonably tight, their sizes were assumed to be a good indicator of the approximate lesion size. Given the scale 19–25 pixels/mm, the size of most lesions was 1–3 mm.

Object detection architectures

Several existing general-purpose deep learning CNN architectures for object detection from images were tested: Faster R-CNN [30], RetinaNet [19], YOLOv5 [12], and EfficientDet [34]. For Faster R-CNN and RetinaNet, alternative backbones (feature extractors) were also tested, namely ResNet-50, ResNet-101 [11], and SwinTransformer (tiny) [20]. For the YOLOv5 architecture, the originally proposed backbones were used, denoted small (S), medium (M), and large (L), all of them based on the CSPDarknet53 architecture [3]. For the EfficientDet, a family of backbones D0, D1…D5 (from the smallest to the largest) was tested. In the following, the networks are denoted *architecture-backbone* (see, e.g., Table 2), with SwinTransformer and ResNet abbreviated as SwinT and R, respectively. YOLOv3 (used in [2]) was also tested, but the results were inferior to YOLOv5.

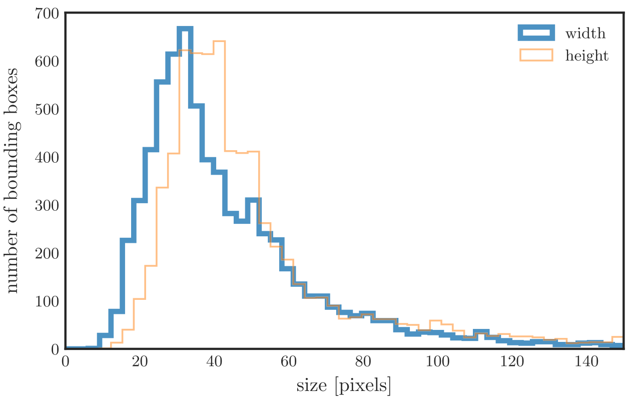

Fig. 2 Histogram of bounding box dimensions in the final dataset D_0. Sizes greater than 150 pixels are omitted

[1] https://github.com/opencv/cvat

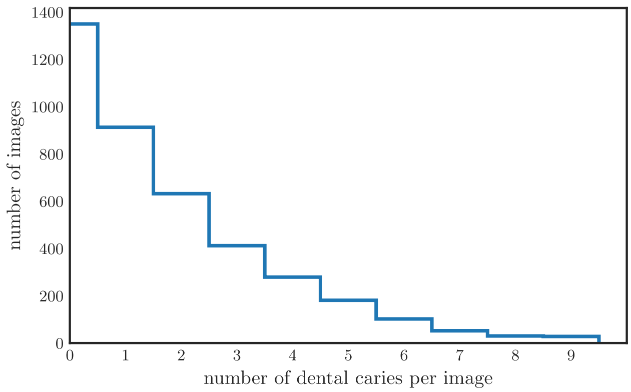

Fig. 3 Histogram of the number of annotated carious lesions per image in the final dataset D_0

Preprocessing and augmentation

The final dataset D_0 (Sect. 2.1) was randomly split into training (70%), validation (15%), and test parts (15%). The intensity was normalized for all images to have the same mean and variance.

During training, the following augmentation operations were applied to artificially increase the dataset size to improve the generalization: horizontal and vertical flip, translation by up to 10% of the image size, rotation by up to 10°, Gaussian blur with $\sigma = 7 \sim 31$ px, and gamma correction with $\gamma = 0.6 \sim 1.4$. The last three operations were applied with a probability of 0.3, and the others with a probability of 0.5.

Optimization and GPU acceleration

AdamW optimizer [22] provided the most consistent results on our data. The initial learning rate was determined by a coarse grid search and then modified by the cosine annealing learning rate scheduler [21].

Nvidia GTX 1080-Ti GPU with 12 GB of memory was used. Depending on the network, it allowed batch sizes (BS) between 16 for the smallest YOLOv5-S model and 1 for EfficientDet-D4 and D5. To compensate for the small batch size, the gradients were accumulated for 1–16 batches, such that the number of accumulated gradient evaluations remains the same (16). For EfficientDet-D4 and D5, with batch size 1, batch-normalization had to be replaced by group-normalization.

Table 1 Statistics of bounding box annotation dimensions in the final dataset D_0

	Width (px)	Height (px)
Image size	1068	795–847
Minimum box size	8	9
Maximum box size	384	315
Mean box size	47.55	53.15
Box size st. deviation	37.99	35.33

Pruning

For each image i, each method a provided a set of bounding boxes

$$\mathsf{B}_{ia} = \{b_{ia}^1, b_{ia}^2, \ldots\} \tag{1}$$

The automatic methods were set to produce 300 boxes per image, much more than the maximum expected number of carious lesions in the image. For each box, a confidence score $0 \leq c(b_{ia}^j) \leq 1$ was also predicted. Boxes with a confidence lower than a threshold, $c < \delta_a$, were discarded. The method-dependent threshold δ_a was determined on the validation set to maximize the F_1 score.

Model ensembling and box fusion

Model ensembling was used to improve the detection performance. Three different ensembles of four models each were created: Ensemble 1 combined four independently trained YOLOv5-M models, while Ensemble 2 combined YOLOv5 models with small (S), medium (M), large (L), and extra large (XL) backbones. Ensemble 3 combined different types of models: RetinaNet-SwinT, Faster R-CNN-ResNet50, YOLOv5-M, and RetinaNet-R101.

For each image i, the sets of bounding boxes B_{ia} predicted by n_a individual detection methods $a = 1, \ldots, n_a$ were first concatenated

$$\mathsf{B}_i = \bigcup_{a=1}^{n_a} \mathsf{B}_{ia} = \{b_{i1}^1, b_{i1}^2, \ldots \\ \ldots b_{i2}^1, b_{i2}^2, \ldots, b_{in_a}^1, b_{in_a}^2, \ldots\} \tag{2}$$

with updated confidence scores

$$c'(b_{ia}^j) = w_a c(b_{ia}^j) \tag{3}$$

where $0 \leq w_a \leq 1$ were fixed, method-dependent weights (see below).

There were usually a lot of nearly identical and overlapping boxes in B_i. These duplicates or near-duplicates could be eliminated for example by non-maximal suppression, soft non-maximal suppression [4], or non-maximum weighted suppression [39]. Weighted boxes fusion (WBF) [32] was used here based on preliminary experiments. It greedily clustered boxes with IoU $> \tau$, where τ is a hyperparameter. For each cluster C, it produced a fused bounding box b_C by taking a weighed average of the box coordinates, using the updated confidence scores c' (3) as weights

$$b_C = \frac{1}{Z_C} \sum_{b \in C} c'(b)\, b \tag{4}$$

$$\text{with} \quad Z_C = \sum_{b \in C} c'(b) \tag{5}$$

and assigning it the confidence score as a weighted mean

$$c(b_C) = \frac{\sum\limits_{b_{ia}^j \in C} w_a c\left(b_{ia}^j\right)}{\sum\limits_{b_{ia}^j \in C} w_a + \sum\limits_{a \in \mathcal{M}} w_a} \tag{6}$$

$$\text{where} \quad \mathcal{M} = \left\{a; \nexists b_{ia}^j \in C, a \in \left\{1, \ldots, n_a\right\}\right\}$$

is a set of methods that do not appear in the cluster C. This way, the confidence $c(b_C)$ is decreased if fewer methods agree.

The hyperparameters τ and w_a were found by numerical optimization of the F_1 score of the fused boxes on the validation set (without the preprocessing described in Sect. 2.7). Exact values of weights w_a were not critical. On the other hand, the results were fairly sensitive to the IoU threshold τ; the best values for τ were $0.45 \sim 0.65$.

Post-processing

The boxes produced by the automatic detection were sometimes nested, as in Fig. 4. The larger box usually matched the ground truth, as it enclosed the entire carious lesion. The smaller box marked only the enamel penetration zone within the larger lesion. This was probably caused by a large number of caries limited to the enamel in the training dataset, while dentin involvement, which implies damage to the enamel, was less frequent. The enclosed detections were therefore pruned as follows:

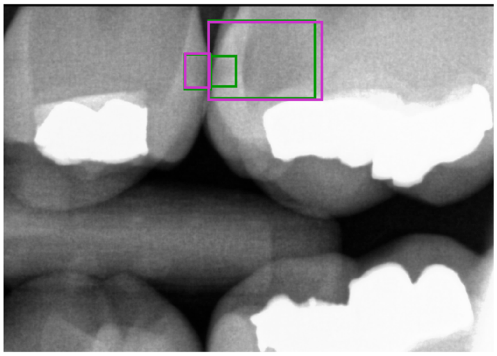

Fig. 4 The automatic method *(green)* produced two nested predictions in the left maxillary first molar (tooth depicted on the top right). The large box enclosed the entire carious lesion and matched the ground truth annotation *(magenta)*. The small box enclosed only the enamel lesion, neglecting the extension to the dentin

For all predicted fused bounding boxes b, sorted in a decreasing order by the confidence score $c(b)$, the remaining predicted boxes b' with smaller confidence $c(b') < c(b)$ and area $S(b') < S(b)$ were considered. If b' was mostly enclosed in b, i.e., if

$$S\left(b \cap b'\right) > \beta S(b'), \tag{7}$$

then the prediction b' was removed. The hyperparameter $\beta = 0.8$ was found by optimizing the F_1 score on the validation set. The post-processing was used in all subsequent experiments except for the comparison of the architecture-backbone average precision (AP).

Evaluation of the models

The performance of the automatic methods was evaluated on the test part of the final dataset D_0 (598 radiographs). The MS COCO competition approach was adopted, and the lesion detection performance was quantified using *average precision* AP@τ [26], where the average is taken over all recall levels between *0* and *1* and τ is the IoU threshold for the detection to be considered correct. AP (without τ) is the average of AP@τ for the subscripts S, M, and L (e.g., AP_S), denote results for small, medium, and large lesions, evaluated for ground truth bounding boxes with area smaller than 32^2 pixels, between 32^2 and 96^2 pixels, and larger than 96^2 pixels, respectively.

Results

Model comparison

Table 2 compares the different neural network detection architectures and backbones (models) by evaluating their AP on the test part of the D_0 dataset. The best-performing architectures were RetinaNet [19] with the SwinTransformer [34] backbone and YOLOv5 [12] with the large backbone. The table also shows that AP indeed decreased with increasing IoU threshold (τ). The largest drop was observed between τ=0.5 and τ=0.75. This indicated that the overlap of predictions and annotations was mostly more than 50% but less than 75%. Finally, it was revealed that the performance of the models was similar for medium and large lesions and lower for small lesions.

Table 2 Comparison of different neural network architectures and backbones (named *architecture-backbone*) on the test part of dataset D_0 in terms of average precision (AP@τ) for different IoU thresholds τ, as an average over all τ (denoted AP) and for the subsets of small, medium, and large lesions (denoted by subscripts S, M, L)

Models	AP	$AP@.3$	$AP@.5$	$AP@.75$	$AP@.5_S$	$AP@.5_M$	$AP@.5_L$
EffDet-D0	0.360	0.832	0.762	0.292	0.682	0.798	0.779
EffDet-D1	0.387	0.848	0.792	0.336	0.696	0.841	0.763
EffDet-D4	0.349	0.821	0.742	0.285	0.660	0.784	0.672
RetinaNet-R50	0.397	0.864	0.814	0.346	0.768	0.84	0.803
RetinaNet-R101	0.384	0.855	0.789	0.323	0.728	0.817	0.815
RetinaNet-SwinT	0.393	**0.872**	**0.827**	0.311	**0.786**	**0.847**	**0.852**
Faster R-CNN-R50	0.391	0.856	0.801	0.326	0.758	0.822	0.794
Faster R-CNN-R101	0.375	0.861	0.791	0.294	0.743	0.818	0.768
YOLOv5-L	**0.408**	0.853	0.815	**0.375**	0.783	0.840	0.769
YOLOv5-S	0.374	0.834	0.773	0.301	0.713	0.804	0.745
YOLOv5-M	0.402	0.861	0.808	0.364	0.765	0.837	0.755

Best values are shown in bold

Table 3 Precision, recall, and F_1-score for the tested neural networks (architectures and backbones) on test part of the dataset D_0

Models	Precision	Recall	F_1
EffDet-D0	0.71	0.72	0.72
EffDet-D1	0.77	0.76	**0.76**
EffDet-D4	0.72	0.73	0.73
RetinaNet-R50	0.77	0.75	0.76
RetinaNet-R101	0.76	0.73	0.74
RetinaNet-SwinT	0.77	0.76	**0.76**
Faster R-CNN-R50	0.73	0.78	0.76
Faster R-CNN-R101	0.75	0.75	0.75
YOLOv5-S	0.76	0.71	0.74
YOLOv5-M	**0.77**	0.74	0.76
YOLOv5-L	0.74	**0.79**	**0.76**

Best values are shown in bold

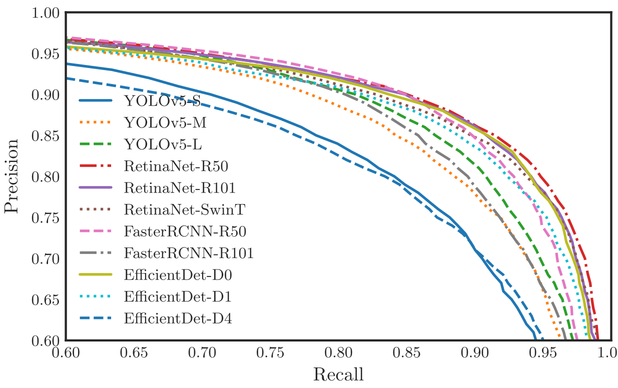

Fig. 5 A zoom of the precision-recall curves for the tested neural networks on the test part of dataset D_0

The resulting F_1 score, precision, and recall are shown in Table 3. EfficientDet-D1 [34] worked best along with RetinaNet-SwinT and YOLOv5-L. The precision-recall curves are shown in Fig. 5.

Model ensembling

The results of the ensemble models are shown in Tables 4 and 5 in terms of the F_1 score and AP, respectively. It can be seen that the ensemble models outperformed the individual models in both F_1 and AP (compared with Table 2). The best-performing ensemble model was the most heterogeneous Ensemble 3. See a companion paper [36] for its comparison with 7 additional human annotators.

Discussion

In this work, multiple deep neural networks were trained to detect caries. As it is common in object detection, their performance was evaluated mainly using average precision AP and the F_1 score. These measures take both precision (positive predictive value) and recall (sensitivity) into account, as AP is defined as the area under the precision-recall curve, and the F_1 score is calculated as the harmonic mean of precision and recall at the working point. Naturally, the results depend on the IoU threshold (τ), i.e., to what extent a model

Table 4 Precision, recall, and the F_1 score of the ensemble models on the test part of the dataset D_0

Model	Precision	Recall	F_1
Ensemble 1	0.81	0.77	0.79
Ensemble 2	0.79	**0.79**	0.79
Ensemble 3	**0.83**	0.77	**0.80**

Best values are shown in bold

Table 5 Evaluation of the ensemble models on the test part of the dataset D_0 in terms of the average precision (AP@τ) for different IoU thresholds τ, as an average over all τ (denoted AP) and for the subsets of large, medium, and small lesions (denoted by subscripts S, M, L)

Model	AP	$AP@.03$	$AP@.5$	$AP@.75$	$AP@.5_S$	$AP@.5_M$	$AP@.5_L$
Ensemble 1	0.434	0.879	0.841	0.390	0.814	0.862	0.804
Ensemble 2	0.442	0.878	0.850	**0.418**	**0.821**	0.866	0.817
Ensemble 3	**0.444**	**0.898**	**0.861**	0.405	**0.821**	**0.879**	**0.876**

Best values shown in bold

prediction must overlap with the corresponding annotation to be considered correct. As for this application, detecting the caries is more important than their exact location, and the threshold $\tau = 0.5$ seems appropriate.

All the networks performed satisfactorily (AP@0.5 0.742–0.827, F_1 score 0.716–0.764), with RetinaNet with the SwinTransformer backbone and YOLOv5 with the large backbone performing the best. A marked improvement in precision was achieved by combining results of several networks, taking advantage of their diversity (see Tables 4 and 5). The best-performing model (Ensemble 3) comprised of RetinaNet-SwinTransformer, YOLOv5-M, Faster R-CNN-ResNet50, and RetinaNet-ResNet101 models and yielded the AP@0.5 of 0.86 and F_1 score of 0.80.

Several automatic caries detection methods have been described in the literature [1, 9, 15, 27, 33]. Direct comparison is unfortunately not possible as neither the test data nor the implementations are available for these methods. In contrast, we provide the source code (see the Data availability statement) and the test dataset [35, 36]. There are also fundamental differences in the composition of the datasets. The included radiographs were acquired using different X-ray machines/sensors and exported in different formats. Furthermore, various exclusion criteria were used, such as the presence of primary teeth [5, 7], the presence of restorations on proximal surfaces, or the absence of caries in the radiograph [7].

Second, some works were concerned only with proximal caries [2, 8, 9, 25] or consisted of manually selected ROIs [8]. In this work, the evaluation was performed at the lesion level to avoid the need to identify individual teeth. However, one consequence of this choice is that accuracy or specificity cannot be calculated since no “true negatives” were labeled in the dataset. Moreover, the performance measure calculated at the level of lesions is numerically different from measures calculated at the level of individual teeth [9] or tooth surfaces [7]. For example, the reported tooth-level specificity of 0.85 and sensitivity of 0.69 [9] corresponds to a precision about 0.62 on the lesion level, assuming that lesions exist on the average in 25% of teeth (as in the study by Chen et al. [7]). Based on such approximative calculations, Ensemble 3 seems to outperform the alternative methods mentioned above.

This study also showed that predicting small carious lesions is less accurate compared to medium and large lesions. This was expected, since the interpretation of incipient lesions is clinically difficult and the annotations may therefore be inconsistent. The present results partially agree with the findings of Chen et al. [7] who also reported the lowest sensitivity in enamel lesions (E1/E2). However, the sensitivity in shallow dentin lesions was similar, and only deep dentin lesions were predicted with a significantly higher sensitivity. Even more complex classification was attempted by Panyarak et al. but the models underperformed [27], and this issue therefore remains to be solved.

The presented results are based on a dataset containing 3989 bitewing radiographs with 7257 carious lesions labeled using bounding boxes. While this dataset seems to be one of the largest datasets of this type, there are limitations. First, the image scaling is not exactly the same, and second, all labels were created by a single annotator. This approach was selected because large discrepancies were found when comparing annotations by multiple experts [36].

Conclusions

Various CNN object detection architectures and backbones were trained to detect caries in bitewing radiographs. While individual networks yielded decent results, they were further improved by postprocessing and fusing results of multiple methods. All models performed best in large or medium-sized lesions which are easier to detect than incipient caries. Please see a companion paper [36] for an extensive comparison of the method described here with 7 additional human annotators.

Author Contributions L.K.: implementation, experiments, writing and editing. J.K.: image analysis, machine learning and statistical methodology, supervision of the implementation and experiments, writing and editing. V.N.: data validation and annotation, writing and editing. A.T.: conceptualization, data curation and annotation, writing and editing.

Funding This work was supported by the General University Hospital in Prague (project GIP-21-SL-01-232) and by the OP VVV funded project CZ.02.1.01/0.0/0.0/16_019/0000765 “Research Center for Informatics”. The study sponsors had no involvement in the study design, analysis, interpretation of the data, writing, or choosing the publication venue.

Data availability statement The cource code for the method described here is available at https://github.com/kuntiik/MT/ and the test dataset is available in Mendelay Data [35].

Declarations

Ethics approval This research was approved by the Ethics Committee of the General University Hospital in Prague, protocol number 82/21. The patients signed a written informed consent, agreeing with the use of their data in anonymized form for research purposes.

Conflict of interest The authors declare no competing interests.

References

1. Bayrakdar IS, Orhan K, Akarsu S, et al (2021) Deep-learning approach for caries detection and segmentation on dental bitewing radiographs. Oral Radiol 38(4). https://doi.org/10.1007/s11282-021-00577-9
2. Bayraktar Y, Ayan E (2021) Diagnosis of interproximal caries lesions with deep convolutional neural network in digital bitewing radiographs. Clin Oral Investig 26(1). https://doi.org/10.1007/s00784-021-04040-1
3. Bochkovskiy A, Wang C, Liao HM (2020) YOLOv4: optimal speed and accuracy of object detection. CoRR abs/2004.10934. https://doi.org/10.48550/arXiv.2004.10934
4. Bodla N, Singh B, Chellappa R, et al (2017) Soft-NMS – improving object detection with one line of code. In: International conference on computer vision (ICCV), pp 5561–5569. https://doi.org/10.48550/ARXIV.1704.04503
5. Cantu AG, Gehrung S, Krois J et al (2020) Detecting caries lesions of different radiographic extension on bitewings using deep learning. J Dent 100:103425. https://doi.org/10.1016/j.jdent.2020.103425
6. Chen L, Li S, Bai Q et al (2021) Review of image classification algorithms based on convolutional neural networks. Remote Sens 13(22):4712. https://doi.org/10.3390/rs13224712
7. Chen X, Guo J, Ye J et al (2023) Detection of proximal caries lesions on bitewing radiographs using deep learning method. Caries Res 56(5–6):455–463. https://doi.org/10.1159/000527418
8. Estai M, Tennant M, Gebauer D et al (2023) 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 134(2):262–270. https://doi.org/10.1016/j.oooo.2022.03.008
9. García-Cañas A, Bonfanti-Gris M, Paraíso-Medina S et al (2022) Diagnosis of interproximal caries lesions in bitewing radiographs using a deep convolutional neural network-based software. Caries Res 56(5–6):503–511. https://doi.org/10.1159/000527491
10. He K, Zhang X, Ren S, et al (2015) Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: IEEE International conference on computer vision (ICCV), pp 1026–1034. https://doi.org/10.1109/ICCV.2015.123
11. He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 770–778. https://doi.org/10.1109/CVPR.2016.90
12. Jocher G, Chaurasia A, Stoken A, et al (2022) YOLOv5 SOTA realtime instance segmentation. https://doi.org/10.5281/zenodo.7347926
13. Khanagar SB, Al-ehaideb A, Maganur PC et al (2021) Developments, application, and performance of artificial intelligence in dentistry–a systematic review. J Dent Sci 16(1):508–522. https://doi.org/10.1016/j.jds.2020.06.019
14. Kuang W, Ye W, (2008) A kernel-modified SVM based computer-aided diagnosis system in initial caries. In, (2008) Second international symposium on intelligent information technology application. IEEE. https://doi.org/10.1109/iita.2008.206
15. Kumar P, Srivastava MM (2018) Example mining for incremental learning in medical imaging. In: IEEE Symposium Series on Computational Intelligence (SSCI). https://doi.org/10.1109/SSCI.2018.8628895
16. Lee JH, Kim DH, Jeong SN et al (2018) Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. J Dent 77:106–111. https://doi.org/10.1016/j.jdent.2018.07.015
17. Lee S, Oh S, Jo J, et al (2021) Deep learning for early dental caries detection in bitewing radiographs. Sci Reports 11(1). https://doi.org/10.1038/s41598-021-96368-7
18. Lian L, Zhu T, Zhu F et al (2021) Deep learning for caries detection and classification. Diagnostics 11(9):1672. https://doi.org/10.3390/diagnostics11091672
19. Lin TY, Goyal P, Girshick R, et al (2017) Focal loss for dense object detection. In: International conference on computer vision (ICCV), pp 2999–300. https://doi.org/10.1109/ICCV.2017.324
20. Liu Z, Lin Y, Cao Y, et al (2021) Swin transformer: hierarchical vision transformer using shifted windows. In: IEEE International conference on computer vision (ICCV), pp 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986
21. Loshchilov I, Hutter F (2017) SGDR: stochastic gradient descent with warm restarts. In: International conference on learning representations (ICLR). https://doi.org/10.48550/arXiv.1608.03983
22. Loshchilov I, Hutter F (2019) Decoupled weight decay regularization. In: International conference on learning representations (ICLR). https://doi.org/10.48550/ARXIV.1711.05101
23. Mao YC, Chen TY, Chou HS et al (2021) Caries and restoration detection using bitewing film based on transfer learning with CNNs. Sensors 21. https://doi.org/10.3390/s21134613
24. Mohammad-Rahimi H, Motamedian SR, Rohban MH et al (2022) Deep learning for caries detection: a systematic review. J Dent 122. https://doi.org/10.1016/j.jdent.2022.104115
25. Moran M, Faria M, Giraldi G et al (2021) Classification of approximal caries in bitewing radiographs using convolutional neural networks. Sensors 21(15):5192. https://doi.org/10.3390/s21155192
26. Padilla R, Passos WL, Dias TLB et al (2021) A comparative analysis of object detection metrics with a companion open-source toolkit. Electronics 10(3):279. https://doi.org/10.3390/electronics10030279
27. Panyarak W, Suttapak W, Wantanajittikul K et al (2023) Assessment of YOLOv3 for caries detection in bitewing radiographs based on the ICCMS radiographic scoring system. Clin Oral Investig 27:1731–1742. https://doi.org/10.1007/s00784-022-04801-6
28. Panyarak W, Wantanajittikul K, Suttapak W et al (2023) 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 135(2):272–281. https://doi.org/10.1016/j.oooo.2022.06.012
29. Prados-Privado M, Villalón JG, Martínez-Martínez CH et al (2020) Dental caries diagnosis and detection using neural networks: a systematic review. J Clin Med 9(11):3579. https://doi.org/10.3390/jcm9113579
30. Ren S, He K, Girshick RB, et al (2015) Faster R-CNN: towards real-time object detection with region proposal networks. In: Neural information processing systems (NIPS). https://doi.org/10.48550/arXiv.1506.01497
31. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al (2019) Standalone artificial intelligence for breast cancer detection in mammog-

raphy: comparison with 101 radiologists. JNCI: J National Cancer Inst 111(9):916–922. https://doi.org/10.1093/jnci/djy222

32. Solovyev R, Wang W, Gabruseva T (2019) Weighted boxes fusion: ensembling boxes from different object detection models. Image Vis Comput. https://doi.org/10.48550/ARXIV.1910.13302
33. Srivastava MM, Kumar P, Pradhan L, et al (2017) Detection of tooth caries in bitewing radiographs using deep learning. In: NIPS workshop on machine learning for health. https://doi.org/10.48550/arXiv.1711.07312
34. Tan M, Pang R, Le QV (2020) EfficientDet: scalable and efficient object detection. In: Computer vision and pattern recognition conference (CVPR). https://doi.org/10.48550/arXiv.1711.07312
35. Tichý A, Kunt L, Kybic J (2023a) Dental caries in bitewing radiographs. Mendeley Data. https://doi.org/10.17632/4fbdxs7s7w.1
36. Tichý A, Kunt L, Nagyová V, et al (2023b) Automatic caries detection in bitewing radiographs. part II: Experimental comparison. Clin Oral Investig. https://doi.org/10.1007/s00784-023-05335-1
37. Wang CW, Huang CT, Lee JH et al (2016) A benchmark for comparison of dental radiography analysis algorithms. Med Image Anal 31:63–76. https://doi.org/10.1016/j.media.2016.02.004
38. Yasa Y, Çelik O, Bayrakdar IS et al (2020) An artificial intelligence proposal to automatic teeth detection and numbering in dental bitewing radiographs. Acta Odontol Scand 79(4):275–281. https://doi.org/10.1080/00016357.2020.1840624
39. Zhou H, Li Z, Ning C, et al (2017) CAD: scale invariant framework for real-time object detection. In: 2017 EEE International conference on computer vision workshops (ICCVW). https://doi.org/10.1109/iccvw.2017.95

Authors and Affiliations

Lukáš Kunt[1] · Jan Kybic[1] · Valéria Nagyová[2] · Antonín Tichý[2]

Lukáš Kunt
kunt.lukas@gmail.com

Valéria Nagyová
valeria.nagyova@lf1.cuni.cz

Antonín Tichý
antonin.tichy@lf1.cuni.cz

[1] Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic

[2] Institute of Dental Medicine, First Faculty of Medicine of the Charles University and General University Hospital, Prague, Czech Republic