

RESEARCH

Open Access



AI-Dentify: deep learning for proximal caries detection on bitewing x-ray - HUNT4 Oral Health Study

Javier Pérez de Frutos^{1*}, Ragnhild Holden Helland¹, Shreya Desai², Line Cathrine Nymo^{4,5}, Thomas Langø¹, Theodor Remman³ and Abhijit Sen^{4,5}

Abstract

Background Dental caries diagnosis requires the manual inspection of diagnostic bitewing images of the patient, followed by a visual inspection and probing of the identified dental pieces with potential lesions. Yet the use of artificial intelligence, and in particular deep-learning, has the potential to aid in the diagnosis by providing a quick and informative analysis of the bitewing images.

Methods A dataset of 13,887 bitewings from the HUNT4 Oral Health Study were annotated individually by six different experts, and used to train three different object detection deep-learning architectures: RetinaNet (ResNet50), YOLOv5 (M size), and EfficientDet (D0 and D1 sizes). A consensus dataset of 197 images, annotated jointly by the same six dental clinicians, was used for evaluation. A five-fold cross validation scheme was used to evaluate the performance of the AI models.

Results The trained models show an increase in average precision and F1-score, and decrease of false negative rate, with respect to the dental clinicians. When compared against the dental clinicians, the YOLOv5 model shows the largest improvement, reporting 0.647 mean average precision, 0.548 mean F1-score, and 0.149 mean false negative rate. Whereas the best annotators on each of these metrics reported 0.299, 0.495, and 0.164 respectively.

Conclusion Deep-learning models have shown the potential to assist dental professionals in the diagnosis of caries. Yet, the task remains challenging due to the artifacts natural to the bitewing images.

Keywords Caries detection, Bitewing, Digital dentistry, Deep learning, Object detection

Introduction

As reported in the WHO Global Oral Health Status Report in 2022 [1], globally 3.5 billion people are afflicted by some form of oral disease, and 2 billion suffer from caries in permanent teeth. Furthermore, untreated dental caries in permanent teeth is the most common dental health condition. Diagnosis of such lesions requires both the inspection of clinical images e.g., X-ray (bi-dimensional images) or cone beam computed tomography (tri-dimensional images), as well as the visual examination and probing of the affected tooth or teeth. This procedure is time consuming, and requires a high level experience

*Correspondence:

Javier Pérez de Frutos

javier.perezdefrutos@sintef.no

¹ Department of Health Research, SINTEF Digital, Professor Brochs gate 2, Trondheim 7030, Norway

² Boneprox A.B., Gothenburg, Sweden

³ Boneprox A.S., Tønsberg, Norway

⁴ Department of public Health and Nursing, Norwegian University of Science and Technology, Trondheim, Norway

⁵ Kompetansesenteret Tannhelse Midt (TkMidt), Trondheim, Norway



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

when analysing the clinical images. The two main image modalities used to assist and support the examination of caries are bitewing (BW) and panoramic radiography (OPG) [2, 3]. Caries, particularly proximal caries, a type of carious lesion located on the surfaces between adjacent teeth, are difficult to detect manually or visually (i.e. using radiographic X-ray images) due to artifacts. Also, poor angulation can hinder the correct identification of the lesions or even occlude lesser grade caries.

Since 2008, the research on the application of artificial intelligence (AI) and, more specifically, deep learning (DL) convolutional neural networks (CNN) models for the analysis of dental has noticeably increased [4–13]. However, research on this field is still limited compared to other clinical areas. Data availability and reliable annotations [8, 13] are the main bottlenecks in the development of machine learning (ML) methods in dentistry. A large portion of the published work uses a dataset of fewer than 300 images, only few studies have access to large datasets [8] with more than 1,000 images like [11, 14, 15]. Of these publications, the work presented in [4–6, 11, 15] focus on object detection, which is the scope of the present study. Object detection or object recognition refers to the task of localising and classifying objects in a picture [16]. The localisation is usually marked using axis-aligned bounding boxes, surrounding the outermost boundary of the item of interest.

In Devito et al. [4], a multi-layer perceptron with 51 artificial neurons (25 in the input layer, 25 in the hidden layer, and one in the output layer) is used to detect proximal caries on BW images, using a dataset of 160 images annotated by 25 experts. Whereas in Srivastava et al. [11], a caries detector built using a tailor designed fully connected neural network was trained with 3,000 annotated BW images. In Singh et al. [6], hand-crafted features for X-ray images are built using Radon and discrete cosine transformations, and further classified using an ensemble of ML techniques such as random forest. Park et al. [15] proposed an ensemble of U-Net and Fast R-CNN for caries detection in colour image, trained with 2,348 RGB intraoral photographic images. Even though, the work done by Cantu et al. [14] focuses on image segmentation, it is worth mentioning because of the dataset used: 3,686 BW images, with caries segmentation annotations, to train a U-Net model for segmentation.

Study goals

In this study we compare three state-of-the-art deep learning architectures for object detection on the task of proximal caries detection, namely RetinaNet, YOLOv5, and EfficientDet. By using an extensive and annotated dataset, we hypothesised that AI object detection models can perform in equal or better terms than dental

clinicians. Hence, in this study we trained the aforementioned architectures in detection and classification of enamel caries, dentine caries, and secondary lesions, in BW images. Then, the models were compared to human annotators in order to test our hypothesis. In addition, a novel processing pipeline for merging multi-observer object detection annotations, based on Gaussian Mixture Models, is proposed.

Methods

Dataset

The bitewing images used in this study were collected as part of the HUNT4 Oral Health Study on the prevalence of periodontitis in a Norwegian population, a sub-study of the fourth phase of the HUNT study [17]. The HUNT4 Oral Health Study is a collaborative study between several Norwegian institutes including: the HUNT research centre, the Kompetansesenteret Tannhelse Midt (TkMidt), the Norwegian University of Technology (NTNU), the University of Oslo (UiO), the Tannhelsetjenestens Kompetansesenter Øst (TkØ), and the Norwegian National Centre for Ageing and Health.

The data collected consisted of clinical and radiographic oral examination, which took place between 2017 and 2019. A total of 7,347 participants were invited to participate in the study, out of a population of 137,233 people (2017) [18]. Only 4,933 participants were included in the Oral Health survey study, out of which, 4,913 completed both clinical and radiographic examination [18, 19]. A total of 19,210 BW and 4,871 OPG images were collected from the participants. The demographics of the dataset showed a distribution of 2,759 (56%) female and 2,174 (44%) male participants, with ages ranging from 19 to 94 years (51.8 ± 16.6 years on average) [18]. For this study, only the BW images were considered.

The following subsections will further describe the steps of the workflow followed in the present study, which is depicted in Fig. 1.

Data annotation

The data was annotated by six dental clinicians with extensive experience in the diagnosis of proximal caries, using the open-source annotation tool AnnotationWeb [20]. The caries were classified in five different categories shown in Table 1. Further details of the annotation procedure can be found in Section 1 of the Additional Materials 1.

To clean the annotations so as to get a ground truth to train the AI models, a novel object detection multi-observers annotations combination strategy was envisioned for this project. First, the annotated bounding boxes were grouped based on the intersection over union (IoU) score, a metric which describes how well the

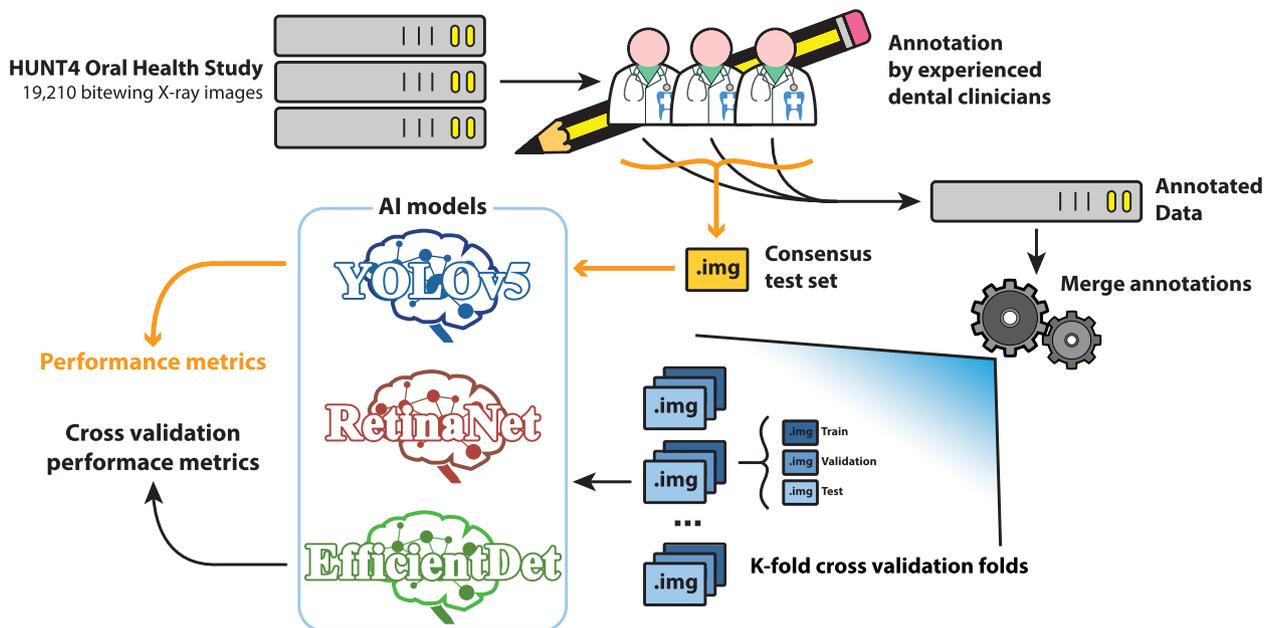


Fig. 1 Data workflow. The HUNT4 Oral Health Study bitewings were stored on a dedicated server, and made available to the expert dentists and dental hygienists for annotation, resulting in the annotated data and the consensus test set. The resulting annotations were merged to build the datasets used in this study. The training dataset was further split following a K-fold (in this study $K = 5$) cross-validation (CV), and pre-processed. The AI models were trained and evaluated on both the CV test set and the consensus test set

Table 1 Definition of the classes used to annotate the dataset

Label name	Description
Grade 1	Radiolucent in outer half of the enamel [21, 22]
Grade 2	Radiolucent in the inner half of the enamel, but not in the dentine [21, 22]
Grade 3	Radiolucent in the outer third of the dentine [21, 22]
Grade 4	Radiolucent in 2/3 of the dentine [21, 22]
Grade 5	Radiolucent in the inner third of the dentine [21, 22]
Secondary lesion	Caries related to sealants or restorations
Unknown grade	Caries whose grade cannot be clearly identified

boxes overlap. Then, a Gaussian distribution was fitted to each bounding box in the group, along the vertical and horizontal axes. A mixture density function (MDF) of a Gaussian Mixture Model in which all distributions have the same weight, was obtained by combining the probability density functions of the fitted Gaussian distributions. The common bounding box was then obtained from the MDF given a probability threshold (p), as detailed in Algorithm 2, in the Additional Materials 1. Alternatively, the non-maximum suppression (NMS) algorithm can be used to find the best fitting bounding box. However, since all the annotations had the same level of confidence, unlike the predictions done by an AI model, NMS will be biased towards the first bounding box selected as

a reference. Lastly, the label of the common bonding box was determined based on the most voted class among the bounding boxes in the group. In case of tie, the most severe class was chosen e.g., dentine caries over enamel caries.

A total of 13,887 images were annotated by one to six of the dental clinicians (see Fig. 2), having a total of 13,585 images annotated by more than one dental clinician.

The distribution of labels in Fig. 3 shows a higher volume of secondary lesions than all the other grades. After discussion with the dental clinicians, it was agreed to merge the grade one and two under the label of “enamel caries”, and grades three to five under the group of “dentine caries”. Secondary caries and unknown grade groups were kept as separate label groups.

In addition, 197 images were annotated by consensus agreement among all the expert annotators, so as to build a test set for evaluation purposes. To create this dataset, hereafter consensus test set, all annotators (dental clinicians) were brought together in the same room and agreement was achieved by consensus on the annotation of the images. The images in the consensus test set had previously been annotated by all annotators individually, with a considerable time gap between the individual annotations and the creation of the consensus agreement annotations, so that the annotations could be considered independent of each other.

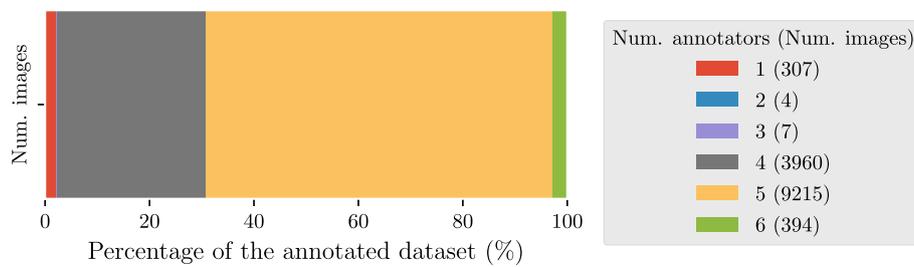


Fig. 2 Distribution of annotated images in the annotated dataset. In the legend, the number of annotated images for each interval is shown within brackets

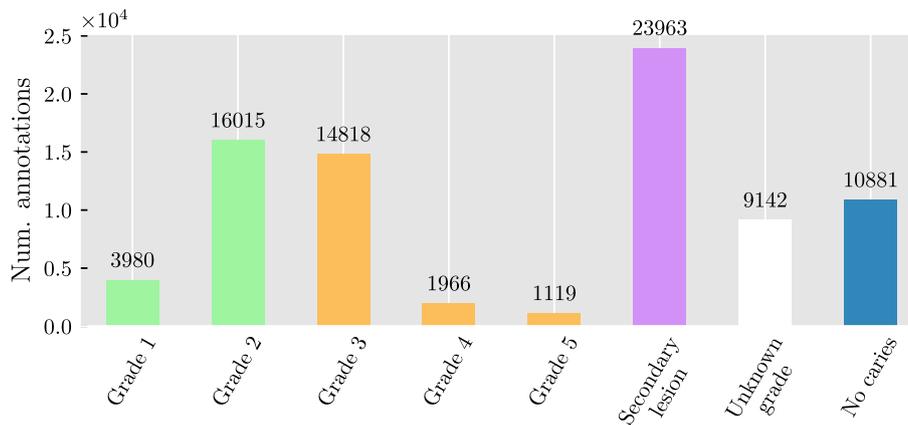


Fig. 3 Distribution of annotations in the dataset annotated by the six dental clinicians. Enamel proximal caries (Grades 1 and 2, total 19,995 annotations) are pictured in light green, dentine lesions (Grade 3 to 5, total 17,903 annotations) are in orange, secondary lesions are depicted in pink, and caries of uncertain grade have been highlighted in white. Image free of lesions (No caries) are shown in dark blue, here the number of annotations matches the number of images

Object detection models

Three state of the art object detection architectures were evaluated for caries detection: RetinaNet (Keras implementation) [23] (ResNet50 backbone), YOLOv5 [24] (size M), and EfficientDet [25] (pretrained D0 and D1). All the models used transfer learning, which is a common strategy when adapting object detection models to a particular dataset, by loading the weights of models trained on a larger dataset set e.g., ImageNet or COCO datasets. RetinaNet was initialised with the weights of ResNet50 trained on ImageNet dataset, YOLOv5 loaded the weights pretrained on COCO dataset (provided in the original repository <https://github.com/ultralytics/yolov5>), and EfficientDet pretrained weights were obtained from <https://github.com/rwightman/efficientdet-pytorch>. For better comparison of the architectures, Table 2 shows the number of parameters for each architecture. Due to time restrictions, not all the versions of YOLOv5 and EfficientDet are included in the current results. The pre-processing and post-processing were kept the same for all models and experiments. Preliminary experiments were conducted with the contrast limited adaptive histogram

Table 2 Number of parameters of each architecture

Architecture	Number of parameters (millions)
YOLOv5 M	21.2M
RetinaNet (ResNet50)	36.4M
EfficientDet D0	3.9M
EfficientDet D1 M	6.6M

equalization (CLAHE) method, inspired by Georgieva et al. [26], but these experiments were eliminated before the final round of cross-validation because they did not lead to any improvement in the scores. Only intensity standardisation to the range [0, 1], and horizontal and vertical flipping were used to augment the training dataset, both being applied with a probability of 0.5.

The training was done on a dedicated server running Ubuntu 20.04. The machine featured a NVidia Quadro RTX 5000 GPU with 16 GB VRAM, a Intel Core i7-9700 CPU, 32 GB RAM, 1 TB SSD, and 8 TB HDD. The training

parameters for each model are summarised in Table 3. In the case of YOLOv5 and RetinaNet, the learning rate was monitored using a learning rate scheduler. For YOLOv5, the OneCycleLR scheduler from PyTorch was used with a maximum learning rate of 10^{-1} . Whereas for RetinaNet, a step learning rate scheduler was used with the patience set to 10 epochs. For both RetinaNet and EfficientDet, an early stopper was used to prevent overfitting. A patience of 20 and minimum loss increment of 5×10^{-3} was configured for RetinaNet, and a patience of 50 and minimum increment of 10^{-1} for EfficientDet.

Validation protocol

After removing the images rejected by the annotators (980), images with unknown grade annotations (4,565), and those in the consensus test dataset (197), the remaining 8,342 images were split into five folds to perform a cross-validation (CV) study. Random sampling without replacement was used to build the folds. The cross-validation training and evaluation was performed with a three-way-split, i.e. for each iteration, three folds were used for training, one fold was used for validation during training, to avoid overfitting; and the final fold was kept aside as a test set for the final performance evaluation.

To test our hypothesis on the performance of AI models, both the trained models in each fold and the annotators were evaluated against the consensus test set. However, due to time constraints, one of the annotators did not complete the individual annotation task, missing one image, and thus the resulting metrics of this annotators are not strictly comparable to those of the models

and other annotators. This annotator is marked with an asterisk (*) in Table 7.

Performance evaluation

As aforementioned, the models described in “Object detection models” section and the annotators were evaluated on the consensus test set. The metrics used in the evaluation were the standard metrics for evaluating object detection model performance: average precision (AP) for each of the classes, the mean average precision (mAP) across classes, the F1-score (F1) for each of the classes, the mean F1-score (mF1), as a surrogate for the recall and precision, the false negative rate (FNR) for each class, and the average across classes (mFNR). These three metrics are in the range [0, 1].

Bootstrap confidence intervals (95%) were computed for the test results of both the models and the annotators, to compare the performance of these. For this comparison, the models trained on the 5th fold were used. The intervals were computed using the bias-corrected and accelerated bootstrap algorithm [27], with 1,000 iterations for confidence interval. Significance in score differences between annotators and models were determined based on overlap of the confidence intervals.

Results

The evaluation results on the consensus test set for the five-fold cross-validation can be found in Tables 4, 5 and 6 (the results per fold can be found in Table 1 in Section 4 of the Additional Material 1). The metrics were computed using the PASCAL VOC metrics implemented in [28], with an IoU thresholds of 0.3. The threshold was deemed an adequate trade-off between

Table 3 Training parameters for each model

Architecture	Batch size	Learning rate	Max. epochs	Optimiser	Framework
YOLOv5 M	8	10^{-2}	180	SGD	PyTorch
RetinaNet	4	10^{-5}	200	Adam	Keras
EfficientDet D0	8	10^{-4}	10^6	AdamW	PyTorch
EfficientDet D1	8	10^{-4}	10^6	AdamW	PyTorch

Table 4 Average precision (AP) results (mean and standard deviation) of the five-fold cross-validation evaluated on the consensus test set. The best metrics are highlighted in bold

Model	Enamel caries	Dentine caries	Secondary lesion	mAP
YOLOv5 M	0.597 ± 0.034	0.622 ± 0.029	0.681 ± 0.022	0.633 ± 0.025
RetinaNet	0.371 ± 0.023	0.333 ± 0.024	0.412 ± 0.034	0.372 ± 0.017
EfficientDet D0	0.310 ± 0.048	0.278 ± 0.017	0.267 ± 0.017	0.285 ± 0.012
EfficientDet D1	0.340 ± 0.029	0.387 ± 0.032	0.377 ± 0.028	0.368 ± 0.027

precision and recall for the current application, where detection of potential caries is preferred. The YOLOv5 model reached the highest AP scores for all classes, as well as the highest F1-scores for two out of three classes, and the lowest FNR for all classes.

The confidence intervals for the means of the distributions of the performance metrics, calculated for each model and each annotator, can be found in Table 7. As described in “Performance evaluation” section, these intervals were used to assess statistical significance between the different architectures, as well as between the models and the human expert rater performance. A graphical representation of these is show in Fig. 4, for ease of interpretation.

Overall, the scores of all of the object detection models were similar to or better than that of the human expert annotators. In terms of AP, the YOLOv5 model achieved significantly higher scores than all of the annotators, as well as the RetinaNet and the EfficientDet D0. The RetinaNet and EfficientDet models also achieved mAP-scores that were similar to or significantly better than the annotators. Regarding F1, The YOLOv5 model achieved significantly higher scores than the RetinaNet model and 4 out of 6 annotators, but the difference with the EfficientDet models were not significant. The EfficientDet models achieved mF1-scores similar to or better than the annotators, whereas the mF1-score of the RetinaNet model was significantly lower than most of the annotators. In

Table 5 F1-score results (mean and standard deviation) of the five-fold cross-validation evaluated on the consensus test set. The best metrics are highlighted in bold

Model	Enamel caries	Dentine caries	Secondary lesion	mF1
YOLOv5 M	0.513 ± 0.011	0.588 ± 0.019	0.563 ± 0.029	0.555 ± 0.011
RetinaNet	0.234 ± 0.032	0.312 ± 0.015	0.228 ± 0.027	0.258 ± 0.013
EfficientDet D0	0.465 ± 0.055	0.459 ± 0.021	0.444 ± 0.006	0.456 ± 0.017
EfficientDet D1	0.533 ± 0.021	0.561 ± 0.020	0.507 ± 0.028	0.534 ± 0.019

Table 6 False negative rate (FNR) results (mean and standard deviation) of the five-fold cross-validation evaluated on the consensus test set. The best metrics are highlighted in bold

Model	Enamel caries	Dentine caries	Secondary lesion	mFNR
YOLOv5 M	0.153 ± 0.021	0.215 ± 0.032	0.160 ± 0.035	0.176 ± 0.025
RetinaNet	0.185 ± 0.060	0.364 ± 0.044	0.200 ± 0.045	0.250 ± 0.031
EfficientDet D0	0.606 ± 0.060	0.636 ± 0.020	0.533 ± 0.020	0.592 ± 0.025
EfficientDet D1	0.479 ± 0.018	0.524 ± 0.020	0.459 ± 0.021	0.487 ± 0.011

Table 7 Mean average precision (mAP), mean F1-score (mF1), and mean false negative rate (mFNR) evaluation of the models (trained on the fifth fold) and individual annotators on the consensus test set, with an IoU-threshold of 0.3. All metrics are reported as score over the whole test set, and a 95% confidence interval. The best results among the models and the annotators have been highlighted in bold letter. * annotators who did not complete the individual annotation task (see "Validation protocol")

Model / Annotator	mAP	mF1	mFNR
YOLOv5, 5 th fold	0.647 [0.566, 0.707]	0.548 [0.506, 0.598]	0.149 [0.110, 0.203]
RetinaNet, 5 th fold	0.407 [0.355, 0.458]	0.177 [0.154, 0.202]	0.210 [0.167, 0.262]
EfficientDet D0, 5 th fold	0.360 [0.290, 0.431]	0.522 [0.461, 0.588]	0.484 [0.422, 0.552]
EfficientDet D1, 5 th fold	0.503 [0.421, 0.569]	0.503 [0.421, 0.569]	0.359 [0.306, 0.431]
Annotator 1*	0.284 [0.231, 0.347]	0.495 [0.447, 0.552]	0.480 [0.413, 0.552]
Annotator 2	0.250 [0.247, 0.285]	0.385 [0.346, 0.420]	0.309 [0.251, 0.374]
Annotator 3	0.242 [0.199, 0.320]	0.403 [0.343, 0.470]	0.631 [0.564, 0.686]
Annotator 4	0.299 [0.270, 0.353]	0.450 [0.411, 0.492]	0.237 [0.180, 0.292]
Annotator 5	0.288 [0.244, 0.356]	0.479 [0.423, 0.528]	0.444 [0.376, 0.515]
Annotator 6	0.261 [0.248, 0.301]	0.376 [0.346, 0.410]	0.164 [0.124, 0.217]

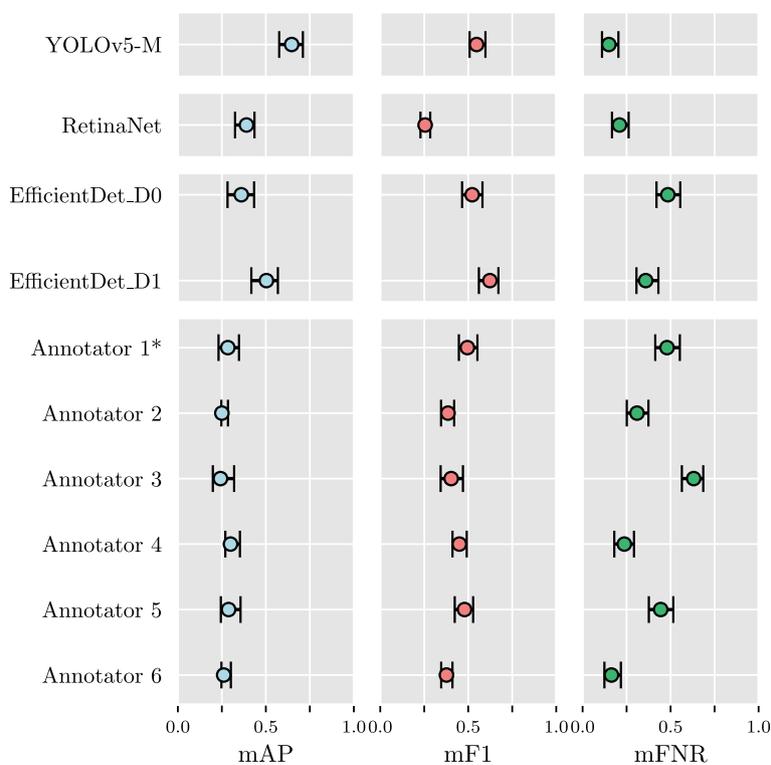


Fig. 4 Bootstrap 95% confidence intervals for the metrics mAP, mF1 and mFNR, for the models and the annotators

terms of FNR, the YOLOv5 model was significantly better (lower scores) than 4 annotators, and similarly, the mFNR of the RetinaNet was significantly better than 3 of the annotators. The EfficientDet models achieved mFNR scores that were similar to or significantly higher than the annotators, meaning that the performance was similar to worse than that of the annotators. The results per class can be found in Table 2 in Section 4 of the Additional Materials 1.

Discussion

In the presented study, three different object detection DL architectures were trained and evaluated on the task of detection of proximal caries in BW X-ray images. The caries were annotated by dental clinicians and classified into three groups: enamel, dentine, and secondary lesions. The predictive performance of the models was assessed in terms of the object detection metrics AP, F1-score, and FNR, and compared against the performance of human expert annotators on a consensus test set. The main finding is that all model performances were on par with or better than the human annotators, with the best model achieving significantly higher scores than the human annotators for all metrics.

The dataset presented in this study features 13,882 BW images, with carious lesions annotated by six dental

clinicians. To the best of our knowledge, this is the largest dataset presented so far for the task of training object detection models for caries detection, exceeding the size of the dataset described in [11] with 3,000 images, and in [14] with 3,686 BW images. A novel strategy for combining the annotations from multiple annotators on the same image was presented, creating robust ground truth annotations for training by combining the expert knowledge of all the annotators. In addition, a test set consisting of 197 images was jointly annotated by all the annotators by consensus agreement. The consensus test set was used to compare the model performances against the performance of the individual annotators, allowing for an assessment of the models usefulness by comparison against a baseline of human expert knowledge.

As detailed in “Validation protocol” section, the performance of each of the architectures was assessed using five-fold cross validation. The folds were built through random sampling of the images without replacement. This approach could lead to data leakage as the split was performed at image level instead of patient level. Nonetheless, due to the size of the dataset, the augmentation during the pre-processing of the images, and the fact that those scans corresponding to a single patient show different regions of the denture, the risk of leakage is minimised. In addition, all of the models were evaluated

on the consensus test set, presented in Tables 4, 5 and 6. The selected metrics, AP, F1-score, and FNR, were deemed appropriate for this experiment, as they summarise the goodness of the models to correctly identify the caries (AP), the trade off between precision and recall (F1-score), and the rate at which the object detectors disregard the caries which are in the BW images (FNR). By using the PASCAL VOC implementation of the metrics, the AP precision is regressed using a larger amount of points, compared to the 11-point interpolation of the AP curve, used in the COCO implementation of AP [28]. This resulted in a better estimate of this metric, and was therefore considered adequate for this study. Lastly, to assess the statistical difference in performance of the

models and the expert annotators, confidence intervals were estimated using the BCa algorithm [27].

The YOLOv5 model achieved the best performance in terms of the metrics used in the study. Both the EfficientDet D1 and YOLOv5 achieved significantly better performance than the RetinaNet in terms of mAP and mF1-score, even though the number of parameters for these models are lower than that of the of RetinaNet. Indeed, EfficientDet D1 is one fifth the size of RetinaNet, and yet it performed better in terms of mAP and F1. On the other hand, both the YOLOv5 and the RetinaNet achieved significantly lower FNR-scores than the EfficientDet models. In sum, all of the presented architectures exhibited different strengths and weaknesses,

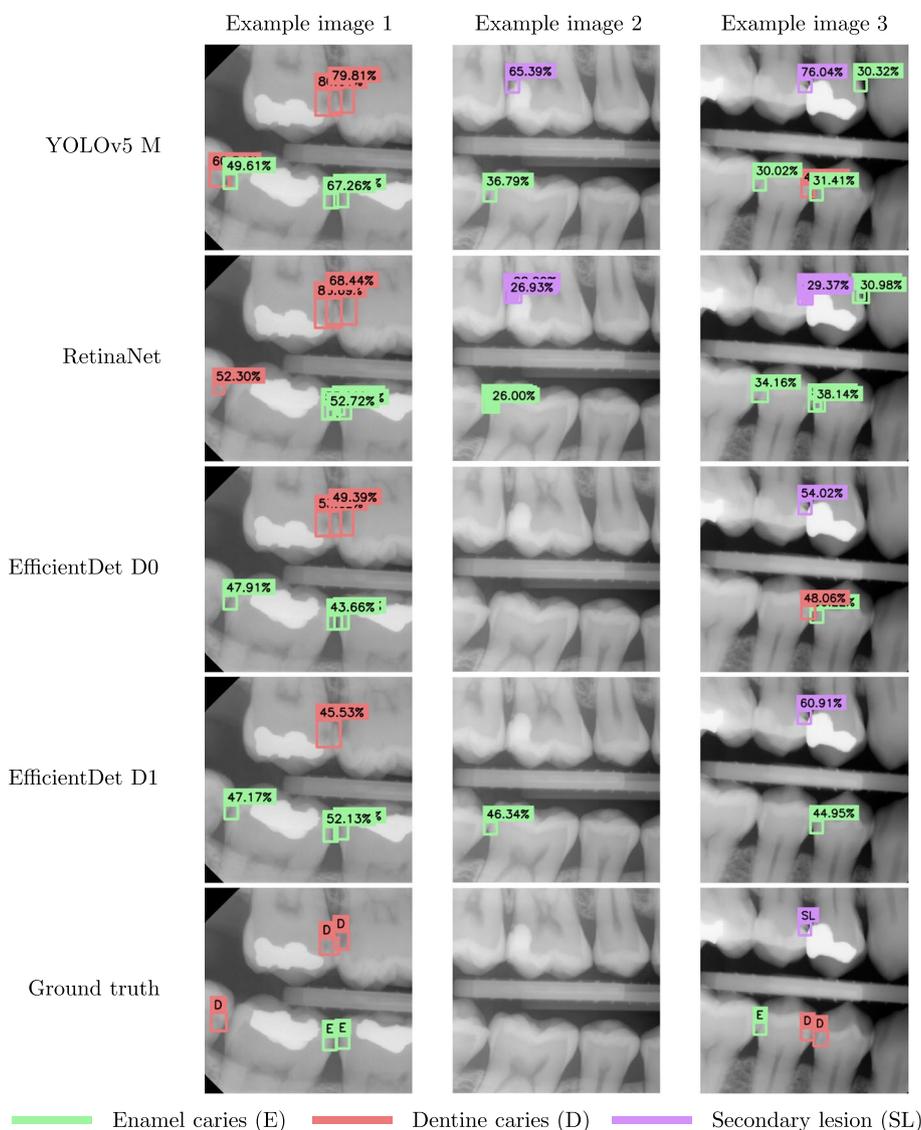


Fig. 5 Detail of bitewing images from the consensus test set with predictions given by the trained models. The ground truth is shown in the bottom row

and an ensemble strategy of the models should be thus be considered, to improve the robustness of the predictions. Figure 5 shows an example of the predictions given by each architecture on three different BW images, the ground truth is given for reference at the bottom row.

Compared to equivalent previously published studies, comparable in scope with the presented work, the performances of the models are lower than the values reported in [4, 5, 8–13, 15], although the values are not directly comparable as they are reported on different datasets. Unlike in these studies, the focus of this work was not to optimise and build a tailored object detection model, but to assess if the dataset was sufficient to obtain equivalent or better performance than dental clinicians, using state-of-the-art architectures. Indeed, as shown in “Results” section, the trained models achieved significantly higher performances in sum on all of the metrics. A combination of the models strengths and weaknesses could thus be a solid foundation for an assistive tool for carious lesion detection in clinical practice.

As introduced in “Introduction” section, the exclusive use of BW images to identify carious lesions is under-par, as it requires a follow-up direct inspection and probing of the infected area. However, the presented deep learning models have the potential to improve the efficiency of the analysis of the bitewing images and aid in the detection of these lesions, helping to speed up and improve the detection and diagnosis of caries.

The architectures included in this study were not modified nor tailored for the used dataset or applications, unlike previously published works [4–13, 15]. Arranging the trained models in an ensemble fashion is expected to increase the overall performance, and the robustness of the predictions. Also, a patch-wise inference could further boost the performance by exposing the network to a closer view of the dental pieces, instead of working on the whole picture. Other augmentation techniques should be considered, such as gamma and brightness augmentations. Explainable AI techniques could be considered to better comprehend the decision process of the trained models, e.g., the features detected for each class. Finally, future work should provide information regarding the inference runtime, so as to assess if the detection models are suitable to be used in practice.

Conclusions

Detection and identification of caries on BW images entails several difficulties, including the monocular view of the dental structures, and hence, presence of artifacts due to the overlap of the dental pieces. Therefore, it is common practice to perform a visual inspection of the

lesions found in the medical images. In this study, it has been shown how AI-powered object detectors can ease the task of finding these lesions in the images, with better performance than dental clinicians. To support this statement, three state-of-the-art object detection architectures were trained on the HUNT4 Oral Health Study BW image dataset, and evaluated against expert dental clinicians. Out of the three architectures, YOLOv5 (medium size) yielded the best results, achieving significantly higher scores than the expert annotators. A combination of the presented models can be used as an assistive tool in the clinic, to speed up and improve the detection rate of carious lesions. The usefulness of such a tool will be assessed in a future clinical validation study.

Abbreviations

AI	Artificial intelligence
ML	Machine learning
DL	Deep learning
BW	Bitewing image
OPG	Panoramic X-ray image
IoU	Intersection over union
NSM	Non-maximum suppression algorithm
MDF	Mixture density function
CLAHE	Contrast limited adaptive histogram equalization
AP	Average precision
F1	F1-score
FNR	False negative rate
mAP	Mean average precision across classes
mF1	Mean F1-score across classes
mFNR	Mean false negative rate across classes

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12903-024-04120-0>.

Supplementary Material 1.

Acknowledgements

The authors would like to express their gratitude to the dental clinicians that helped with the annotations of the BW images: Trine Matheson Bye, Gunnar Lyngstad, Odd-Arne Opland, Harald Solem, and Mats Säll. Also to Theodor Remman, CEO of Boneprox A.S., and project manager of the AI-Dentify project. Furthermore, we would like to thank Hedda Høvik, Astrid J. Feuerherm, and Patrik Cetrelli working at TkMidt for helping in data processing, logistics, and making resources available.

The Trøndelag Health Study (HUNT) is a collaboration between HUNT Research Centre (Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology NTNU), Trøndelag County Council, Central Norway Regional Health Authority, and the Norwegian Institute of Public Health.

Authors' contributions

Conceptualization: JPdF, RHH, SD, AS; Methodology: JPdF, RHH, SD, AS; Data acquisition: JPdF, LCN; Formal analysis and investigation: JPdF, RHH, SD, LCN; Writing - original draft preparation: JPdF, RHH, SD; Writing - review and editing: JPdF, RHH, SD, LCN, AS, TL; Funding acquisition: TL, TR, AS; Resources: TL, TR, AS; Supervision: TL, TR, AS.

Funding

Open access funding provided by SINTEF The AI-Dentify project (project number 321408-IPNÆRINGSLIV20) is funded by the Research Council of Norway, under the scope of the Innovation Project for the Industrial Sector.

Availability of data and materials

The HUNT Research Centre is authorised by the Norwegian Data Inspectorate to securely store and manage the data. De-identified data are shared with researchers upon approval of their research protocol by the Regional Ethical Committee and HUNT Research Centre. To safeguard participant privacy, HUNT Research Centre minimises data storage outside its data bank and refrains from depositing data in open repositories. Detailed records of all exported data for various projects are maintained in the HUNT data bank, and the centre can reproduce this information upon request. Data export is unrestricted, subject to approved applications submitted to HUNT Research Centre. All the data in the this manuscript are available from TkmMidt (contact: Abhijit Sen, abhijit.sen@ntnu.no) on reasonable request. The code and trained models can be provided upon reasonable request to Boneprox A.B. (contact: Shreya Desai, shreya.desai@boneprox.se).

Declarations

Ethics approval and consent to participate

Ethical approval has already been granted by the Regional Ethical Committee (REK) based in central Norway (project number 64645), and also had approval from Norsk Senter for Forskningsdata (reference number 718269). In the HUNT4 Oral Health Study, written and signed consent was acquired from the participants [19].

Consent for publication

Not applicable.

Competing interests

The authors declare the following financial interest/personal relationships that may be considered as potential competing interests: SD is employee at Boneprox A.B., and TR is CEO of Boneprox A.S., and is co-founder of Boneprox A.S.

Received: 30 September 2023 Accepted: 7 March 2024

Published online: 18 March 2024

References

1. Organisation WH. Global oral health status report: towards universal health coverage for oral health by 2030. Geneva: World Health Organization; 2022.
2. Schwendicke F, Tzschoppe M, Paris S. Accuracy of dental radiographs for caries detection. *Evid-Based Dent*. 2016;17(2):43. <https://doi.org/10.1038/sj.ebd.6401166>.
3. Schwendicke F, Göstemeyer G. Conventional bitewing radiography. *Clin Dent Rev*. 2020;4(1):22. <https://doi.org/10.1007/s41894-020-00086-8>.
4. Devito KL, de Souza Barbosa F, Filho WNF. An artificial multilayer perceptron neural network for diagnosis of proximal dental caries. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod*. 2008;106(6):879–84. <https://doi.org/10.1016/J.TRIPLEO.2008.03.002>.
5. Berdouses ED, Koutsouri GD, Tripoliti EE, Matsopoulos GK, Oulis CJ, Fotiadis DI. A computer-aided automated methodology for the detection and classification of occlusal caries from photographic color images. *Comput Biol Med*. 2015;62:119–35. <https://doi.org/10.1016/J.COMPBIOMED.2015.04.016>.
6. Singh P, Sehgal P. Automated caries detection based on Radon transformation and DCT. 8th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2017. 2017. <https://doi.org/10.1109/ICCCNT.2017.8204030>.
7. Hwang JJ, Jung YH, Cho BH, Heo MS. An overview of deep learning in the field of dentistry. *Imaging Sci Dent*. 2019;49(1):1. <https://doi.org/10.5624/isd.2019.49.1.1>.
8. Prados-Privado M, García Villalón J, Martínez-Martínez CH, Ivorra C, Prados-Frutos JC. Dental Caries Diagnosis and Detection Using Neural Networks: A Systematic Review. *J Clin Med*. 2020;9(11):3579. <https://doi.org/10.3390/jcm9113579>.
9. Schwendicke F, Golla T, Dreher M, Krois J. Convolutional neural networks for dental image diagnostics: A scoping review. *J Dent*. 2019;91:103226. <https://doi.org/10.1016/j.jdent.2019.103226>.
10. Choi J, Eun H, Kim C. Boosting Proximal Dental Caries Detection via Combination of Variational Methods and Convolutional Neural Network. *J Signal*

- Process Syst. 2018;90(1):87–97. <https://doi.org/10.1007/S11265-016-1214-6/FIGURES/10>.
11. Srivastava MM, Kumar P, Pradhan L, Varadarajan S. Detection of Tooth caries in Bitewing Radiographs using Deep Learning. 2017. arXiv preprint [arXiv:1711.07312](https://arxiv.org/abs/1711.07312).
 12. Lee JH, Kim DH, Jeong SN, Choi SH. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. *J Dent*. 2018;77:106–11. <https://doi.org/10.1016/J.JDENT.2018.07.015>.
 13. Lee S, Oh Si, Jo J, Kang S, Shin Y, Park Jw. Deep learning for early dental caries detection in bitewing radiographs. *Sci Rep*. 2021;11(1):16807. <https://doi.org/10.1038/s41598-021-96368-7>.
 14. Cantu AG, Gehrung S, Krois J, Chaurasia A, Rossi JG, Gaudin R, et al. Detecting caries lesions of different radiographic extension on bitewings using deep learning. *J Dent*. 2020;100:103425. <https://doi.org/10.1016/j.jdent.2020.103425>.
 15. Park EY, Cho H, Kang S, Jeong S, Kim EK. Caries detection with tooth surface segmentation on intraoral photographic images using deep learning. *BMC Oral Health*. 2022;22(1):573. <https://doi.org/10.1186/s12903-022-02589-1>.
 16. Godfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016. <http://www.deeplearningbook.org/>.
 17. Krokstad S, Langhammer A, Hveem K, Holmen T, Midtthjell K, Stene T, et al. Cohort Profile: The HUNT Study. *Norway Int J Epidemiol*. 2012;42(4):968–77. <https://doi.org/10.1093/ije/dys095>.
 18. Støtle IH, Verket A, Høvik H, Sen A, Koldsland OC. Prevalence of periodontitis based on the 2017 classification in a Norwegian population: The HUNT study. *J Clin Periodontol*. 2021;48(9):1189–99. <https://doi.org/10.1111/jcpe.13507>.
 19. Rødseth SC, Høvik H, Schuller AA, Skudutyte-Rysstad R. Dental caries in a Norwegian adult population, the HUNT4 oral health study; prevalence, distribution and 45-year trends. *Acta Odontol Scand*. 2022;81(3):202–10. <https://doi.org/10.1080/00016357.2022.2117735>.
 20. Smistad E, Østvik A, Lovstakken L. Annotation Web - An open-source web-based annotation tool for ultrasound images. 2021. p. 1–4. <https://doi.org/10.1109/IUS52206.2021.9593336>.
 21. Westberg TE, Døving LM, Bjørg A. Kliniske rutiner- Kariologi. 2010. <https://www.odont.uio.no/iko/om/organisasjon/fagavd/kariologi-gerodontologi/rutiner-metoder/>. Accessed 12/07/2021.
 22. Hansson HH, Espelid I. Kan vi stole på kariesregistreringen? Validering av to visuelle indekser for registrering av okklusalkaries basert på ekstraherte tenner. *Nor Tannlegeforen Tid*. 2012;(122):676–682. <https://doi.org/10.56373/2012-9-11>.
 23. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection. 2017. arXiv preprint [arXiv:1708.02002](https://arxiv.org/abs/1708.02002).
 24. Jocher G, Chaurasia A, Stoken A, Borovec J. NanoCode012, Kwon Y, et al. ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation. 2022. <https://doi.org/10.5281/ZENODO.7347926>.
 25. Tan M, Pang R, Le QV. EfficientDet: Scalable and efficient object detection. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2020. p. 10778–10787. <https://doi.org/10.1109/CVPR42600.2020.01079>.
 26. Georgieva VM, Mihaylova AD, Petrov PP. An application of dental X-ray image enhancement. 2017 13th International Conference on Advanced Technologies, Systems and Services in Telecommunications, TELSIKS. 2017. p. 447–450. <https://doi.org/10.1109/TELSIKS.2017.8246321>.
 27. Davison AC, Hinkley DV. Bootstrap methods and their applications. New York: Cambridge University Press; 1997.
 28. Padilla R, Passos WL, Dias TLB, Netto SL, Da Silva EAB. A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics*. 2021;10(3):279. <https://doi.org/10.3390/ELECTRONICS10030279>.

Publisher's Note

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