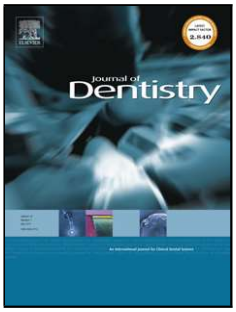
Journal Pre-proof



Detecting caries lesions of different radiographic extension on bitewings using deep learning

Anselmo Garcia Cantu, Sascha Gehrung, Joachim Krois, Akhilanand Chaurasia, Jesus Gomez Rossi, Robert Gaudin, Karim Elhennawy, Falk Schwendicke

PII: S0300-5712(20)30171-8

DOI: https://doi.org/10.1016/j.jdent.2020.103425

Reference: JJOD 103425

To appear in: *Journal of Dentistry*

Received Date: 19 May 2020

Revised Date: 30 June 2020

Accepted Date: 1 July 2020

Please cite this article as: Cantu AG, Gehrung S, Krois J, Chaurasia A, Rossi JG, Gaudin R, Elhennawy K, Schwendicke F, Detecting caries lesions of different radiographic extension on bitewings using deep learning, *Journal of Dentistry* (2020), doi: https://doi.org/10.1016/j.jdent.2020.103425

Detecting caries lesions of different radiographic extension on bitewings using deep learning

Anselmo Garcia Cantu[1*], Sascha Gehrung[1*], Joachim Krois[1], Akhilanand Chaurasia[2], Jesus Gomez Rossi[1], Robert Gaudin[3], Karim Elhennawy[4], Falk Schwendicke[1**]

[1] Department of Oral Diagnostics, Digital Health and Health Services Research, Charité - Universitätsmedizin Berlin, Germany

[2] Department of Oral Medicine and Radiology, King George's Medical University.,Lucknow, India

[3] Department of Oral- and Maxillofacial Surgery, Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health

[4] Department of Orthodontics, Dentofacial Orthopedics and Pedodontics, Charité - Universitätsmedizin Berlin, Germany

Short title: Deep Learning for Detecting Caries

* Shared first authorship.

****Correspondence to:**

Prof. Dr. Falk Schwendicke MDPH

Charité – Universitätsmedizin Berlin

Department of Oral Diagnostics, Digital Health and Health Services Research, Charité - Universitätsmedizin Berlin, Germany

Aßmannshauser Str. 4-6

14197 Berlin, Germany

Phone: 0049 30 450 62556

Fax: 0049 30 450 7562 556

falk.schwendicke@charite.de

Abstract

Objectives: We aimed to apply deep learning to detect caries lesions of different radiographic extension on bitewings, hypothesizing it to be significantly more accurate than individual dentists.

Methods: 3,686 bitewing radiographs were assessed by four experienced dentists. Caries lesions were marked in a pixelwise fashion. The union of all pixels was defined as reference test. The data was divided into a training (3,293), validation (252) and test dataset (141). We applied a convolutional neural network (U-Net) and used the Intersection-over-Union as validation metric. The performance of the trained neural network on the test dataset was compared against that of seven independent using tooth-level accuracy metrics. Stratification according to lesion depth (enamel lesions E1/2, dentin lesions into middle or inner third D2/3) was applied.

Results: The neural network showed an accuracy of 0.80; dentists' mean accuracy was significantly lower at 0.71 (min-max: 0.61-0.78, $p<0.05$). The neural network was significantly more sensitive than dentists (0.75 versus 0.36 (0.19-0.65; $p=0.006$), while its specificity was not significantly lower (0.83) than those of the dentists (0.91 (0.69-0.98; $p>0.05$); $p>0.05$). The neural network showed robust sensitivities at or above 0.70 for both initial and advanced lesions. Dentists largely showed low sensitivities for initial lesions (all except one dentist showed sensitivities below 0.25), while those for advanced ones were between 0.40 and 0.75.

Conclusions: To detect caries lesions on bitewing radiographs, a deep neural network was significantly more accurate than dentists.

Clinical significance: Deep learning may assist dentists to detect especially initial caries lesions on bitewings. The impact of using such models on decision-making should be explored.

Keywords: Artificial Intelligence; Caries; Digital imaging/radiology; Mathematical modeling; Radiography

Introduction

Caries is the most prevalent disease of humankind and places a significant burden on individuals' quality of life and healthcare systems [1]. To manage caries lesions, early detection and non- or micro-invasive therapy are recommended, with the conventional invasive/restorative therapy of caries being a last resort for managing advanced lesions rather than the standard it had been for a century [2]. To allow such early detection, the standard diagnostic tool, which is visual-tactile inspection, is insufficient on non-accessible (mainly proximal) surfaces [3]. A commonly applied additional method to detect caries lesions and assess their extension is bitewing radiography, which comes with higher sensitivity than visual-tactile inspection and similarly high specificity [4]. Caries lesion detection and assessment on bitewings shows limited reliability and validity; there is significant disagreement between examiners and a considerable proportion of false positive or false negative detections. A systematic review [4] found that, based on 117 studies and using data from 13,375 teeth, the detection of both initial and advanced lesions had a mean sensitivity ranging between 0.24 and 0.43 (i.e. up to 76% of lesions were missed), while the mean specificity ranged between 0.89 and 0.97 (i.e. false positive detections occurred far less often). There is evidence that more experienced examiners show an improved accuracy compared with less experienced ones; examiners with experience have an up to almost four times greater chance of correctly detecting proximal caries lesions than those examiners with low experience [5].

Automated assistance systems for dental radiographic imagery may overcome or alleviate these shortcomings, allowing more reliable and accurate assessment of caries lesions, for example on bitewings, especially in the hands of less experienced dentists. They may further save diagnostic time for assessment and reporting.

Machine learning (ML), a subfield of what is often coined "artificial intelligence", has shown to be a very powerful technique for computer-aided diagnostic support tasks. In machine learning, algorithms learn patterns and structures in data ("training") and may then be applied to make predictions on unseen data. A popular field in machine learning is "deep learning", where multi-layered (deep) neural networks are used to learn hierarchical features in the data. For complex cases such as imagery, so-called Convolutional Neural Networks, CNNs, are most commonly employed, learning features like edges, corners, shapes, and macroscopic patterns. Deep learning refers to the process of data (e.g. images) and corresponding labels (e.g. "carious tooth", or "specific area on an image where a caries lesion is present") being repetitively passed through the neural network during training, with the model parameters (so-called weights) being iteratively adjusted to improve of the model's accuracy [6, 7].

In healthcare, and in particular for medical image analysis, deep learning using CNNs has been successfully employed in diagnostic assistance systems, for example, to detect breast cancer in mammographies [8], skin cancer in clinical skin screenings [9], or diabetic retinopathy in eye

examinations [10]. In dentistry, CNNs have been applied to detect periodontal bone loss on peri-apical and panoramic radiographs, and apical lesions and caries lesions on peri-apical radiographs, all with acceptable to high accuracies [11]. So far, deep learning has only sparsely been applied for caries detection on bitewings, and data on deep learning of lesions of different radiographic extension are sparse. The latter is relevant from a decision-making and health economic perspective.

We aimed to apply deep learning on bitewings to detect proximal caries lesions of different radiographic extensions. We hypothesized that deep neural networks show a significantly superior accuracy compared with average dental experts for detecting caries lesions.

Data and methods

Study design

The present study compared the performance of a group of individual dentists and a deep neural network to detect caries lesions on bitewing images. Reporting of this study follows the STARD (Standards for Reporting of Diagnostic Accuracy Studies) guideline [12] and the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) [13].

In the present study we applied U-Net, a particular type of convolutional neural network, first published by Ronneberger et al (2015), which is widely applied for segmentation tasks in medical imaging [14]. With this type of classification model each pixel in an image can be assigned to a different class, as in our case into binary classes: either a pixel belongs to a carious lesion or it does not. As a consequence, the model outputs a binary mask that outlines and thereby highlights any carious lesion. Such pixel-wise results are beneficial in the sense that a domain expert can judge quite easily if the model output areas fit the image area with the suspected conditions (e.g. a carious lesion). However, for an automatized approach to evaluate the performance of the model, a metric needs to be defined that assesses the level of congruency between the truly affected and the predicted areas. For the present study we leverage the intersection over union score (IoU), a widely used metric which quantifies the similarity between the predicted area and the ground-truth area as the intersection divided by the union of the two areas. Note that for transporting information towards its medical usefulness, other metrics (mainly on tooth, not pixel level) are useful and were employed in the present study, as described below.

Sample size calculation

Our primary outcome, accuracy, was used for sample size estimation. Sample size estimation was performed under the assumption of using a paired two-sided t-test to compare intervention groups

(neural network and dentists) on the test dataset. A clustered design applied, with multiple teeth being assessed per radiograph. Hence, there was the need to account for this clustering, which was done via applying the Design Effect (DE), which is calculated via DE=1+(m-1)*ICC, with m being the cluster size and ICC being the Intraclass Correlation Coefficient. The ICC reflects on the clustering of teeth in each radiograph and has been found to be around 0.2 in this population [15]. Given the expected differences in accuracy between the neural network (accuracy: 0.80) and the dentists (in mean 0.75) [4], and conservatively assuming a standard deviation of 0.4, any study powered with 1-beta=0.80 to detect a difference with alpha=0.05 requires a total of 442 teeth to be assessed when conservatively assuming a limited correlation (0.01) between the groups. With a cluster size of n=8 (i.e. teeth per radiograph), the DE=1+(8-1)*0.2=2.4. Hence, the overall number of teeth to be assessed was 442*2.4=1061, or 133 radiographs (each with 8 teeth). The final test set consisted of 141 radiographs (see below).

Reference data set

Our imagery dataset consisted of a total of 3,686 bitewings originating from routine care provided at the dental clinic at xxx. The data were collected between 2016 and 2018. A comprehensive sample of these two years was drawn. The patients' ethnicity can be assumed to be predominantly Caucasian, while it is worth noting that xxx hosts a large Turkish and Arabic community, too (totaling approx. 9% of the city's total population). Images and metadata such as age, gender and date of image generation, among others, were available. Metadata were used for descriptive analyses only. Data collection was ethically approved (xxx ethics committee EA4/080/18). Only bitewings of permanent teeth, with at least the crowns of one dental arch being detectable, were included. Bitewings of primary teeth or those where any assessment was deemed impossible were excluded. There were 52% male and 48% female patients in the dataset. The mean (SD, min-max) age was 36.5 (14.3, 14-89) years. Most of the data (72%) were generated using radiographic machines from the manufacturer Dentsply Sirona (Bensheim, Germany), mainly Orthophos XG; 28% were generated using machines from Dürr Dental (Bietigheim-Bissingen, Germany). On all bitewings, each tooth was segmented and labelled using the FDI scheme by one dentist; this was checked by another one.

In the absence of a "hard" reference test [11], images were pixelwise labeled by three expert dentists independently and in triplicate, using an in-house custom-built annotation tool [16]. All of the labels were reviewed and revised (addition, deletion, confirmation) by a fourth expert dentist. All experts were employed at a specialist clinic for oral diagnostics and operative and preventive dentistry, mainly cariology, with a minimum clinical experience of 3 years. The examiners were instructed in person and calibrated using a handbook (describing how to use the annotation tool and how to annotate caries lesions, but also how to discriminate them from other entities) before the labeling tasks. No formal metrics of agreement (e.g. inter- and intra-examiner agreement) were collected. The annotators

independently assessed all available bitewings in a pixelwise fashion in dimly lit rooms on diagnostic screens and under standardized conditions. No triangulation with any clinical records etc. was performed. The reference test ("gold standard") was eventually constructed from all annotated pixels, i.e. the union of all annotated areas on the image.

Lesion staging

Each annotation on the test set (see below) was further classified by two independent dentists according to their stage, involving the outer or inner half of enamel (E1/2) and outer, middle or inner third of dentin (D1-3). In case of disagreements, consensus was reached by discussion. For analytic purposes, two sub-samples of stages were defined, gauging the accuracy of the model and the dentists on initial lesions (n=32 teeth), including only E1/2, and advanced lesions (n=67 teeth), including only D2/3. Further stratification into lesions on the mesial and the distal surface was applied.

The model

As described, U-Net, a fully convolutional neural network was used [14]. The U-Net architecture is characterized by a U-shaped alignment of convolutional network layers and skip connections between them. The left part of the "U" is referred to as encoder and the right part of the "U" is referred to as decoder. In the encoder part, the model condenses the input and the contextual information increases while the exact positional information about objects decreases. In the decoder part the contextual information is expanded and combined with precise information about object locations through the skip connections between the encoder and decoder layers. In this study, an EfficientNet-B5 network was chosen as the encoder [17]. Details on model architecture and implementation details are provided in the appendix (A1).

Model training and data preparation

The U-Net model was trained on 3,293 labeled images. Several models were trained with different training strategies, loss functions (the model error quantifier) and combinations of the parameters controlling the model optimization process. The outcomes of these experiments were evaluated using the mean-IoU score on the validation set consisting of 252 annotated images. Among the models, the one showing the highest mean IoU was used to perform predictions on the set of 141 test images.

During the model training process all images were resized (width=512 px, height=416 px). To speed-up and improve the optimization process, the model weights were initialized with the values obtained from training this model for caries segmentation on panoramic radiographs (there we used approx. 3000 annotated panoramic radiographs; unpublished). In order to improve model generalization, the set of images was augmented by applying random transformations, both geometric (image flipping, center

cropping, xy-translation and rotations) and on pixel level (gaussian-blur, sharpening, contrast and brightness).

During the training process, the model weights were updated to minimize a loss function of the binary-focal type using an Adam optimizer. First, the model was trained for 10 training cycles (epochs) holding the encoder weights constant and starting with a learning rate value of $5e^{-3}$. The training was further extended for 190 epochs, now allowing the optimization of weights on all of the network layers, with a batch size of 2 (due of memory constraints) and an initial learning rate of $5e^{-4}$. The model weight updates obtained after every epoch were stored, provided these showed an improvement in the mean IoU on the validation set compared with previous epochs. At the end of the training process, the best set of weights was selected. For the final predictions, we applied test time augmentation, which means that we predicted on the original image as well as on transformed (vertical and horizontal flips) versions of the image. The output of predictions (in the [0, 1]-interval) was binarized using an optimized cutoff value. The setup of the model architecture and optimization process were carried out using the Deep Learning library Keras and the Python programming language. The model was trained on a GeForce GTX 1080 Ti GPU using CUDA 10.

Comparator dentists

A cohort of seven dentists, all of them working in dental practices or clinics for min-max 3-14 years were used as comparator group to allow gauging the relative performance of the neural network against individual dentists. Each of the participants independently completed the described segmentation task on a set of 141 images, from which 128 images showed caries at different degrees of severity; 13 images were negative cases.

Evaluation metrics

To compare the neural network with the dentists, the accuracy of both was evaluated on tooth level. The prediction cutoff value used to identify the true positive (TP), false positive (FP), true negative (TN) and false negative (FN) detections for the tooth level evaluation corresponded to the one that maximized the F1 score on the validation set. Details on the tooth-level formulation are provided in the appendix (A2). The tooth-level performances of the model and the dentists were evaluated using an ensemble of seven different scores, i.e. accuracy, sensitivity, specificity, positive/negative predicted value (PPV/NPV), Matthew's correlation coefficient (MCC) and F1-score), each capturing different aspects of classification quality. The MCC is an overall measure of the similarity between the detected and actual cases and provides a fair evaluation in case of class imbalance; it is defined as $\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$, with true positive (TP), false positive (FP), true negative (TN) and false negative (FN) detections being used for calculation. The F1-score is the harmonic mean of PPV and sensitivity, defined as

$\frac{2\times PPV\times Sensitivity}{PPV+Sensitivity}$. Comparisons between the model and the dentists' performance were undertaken using a two-sided paired t-test. A p-value of $p<0.05$ was considered to be significant.

Results

Table 1 summarizes the performances of the neural network and the dentists on all lesions and stratified for lesions on the mesial and distal surfaces, respectively. The neural network showed an overall accuracy of 0.80; the dentists' mean accuracy was significantly lower at 0.71 (min-max: 0.71; 0.61-0.78; $p<0.05$). The neural network was significantly and substantively more sensitive than dentists (0.75 versus 0.36; 0.19-0.65; $p=0.006$), while its specificity was not significantly lower (0.83) than that of the dentists (0.91; 0.69-0.98; $p>0.05$). Although the PPV was not significantly different between the neural network and dentists ($p>0.40$), the NPV of the neural network (0.86) was higher than that of the dentists (0.72; 0.68-0.82; $p<0.001$). The F1-score of the neural network (0.73) was nearly twice as high as that of the dentists (0.41; 0.26-0.63; $p<0.001$). Also, the MCC was significantly higher for the neural network (0.57) than the dentists (0.35; 0.14-0.51; $p<0.001$). Generally, differences between surfaces were minimal. Notably, individual dentists showed a large variance in performance (Fig. 1), all of them showing poorer sensitivities than the model regardless of the specificity level. Exemplary bitewings with predictions and dentists' segmentations are shown in Figure 2.

If assessing the neural network's and dentists' sensitivities to detect initial and advanced lesions, the model showed robust sensitivities at or above 0.70 for both strata (Fig. 3). Dentists, in contrasts, largely showed low sensitivities for initial lesions (all except one dentist <0.25), while those on advanced ones were between 0.40 and 0.75.

Discussion

Caries detection is marred by limited and varying accuracy of individual examiners leading to inconsistent decisions and suboptimal care. To date, deep learning has very rarely been employed to interpret bitewing radiographs [18], which are the main diagnostic source for caries detection. Moreover, deep learning models have not been regularly compared against dentists and their performance, and there is no knowledge on how such models (neural networks) perform on lesions of different depths in comparisons with dentists. The latter, i.e. lesion-specific performance, has implications for clinical decision making.

The present study compared a neural network with a group of seven experienced dentists and found the model to show significantly higher overall accuracy (but also F1-score and MCC) than dentists. We

hence accept our hypothesis. Notably, dentists rarely outperformed the neural network for specificity, while for sensitivity they performed significantly and substantively worse. Especially for initial caries lesions, the risk of under-detection by dentists was very high. The neural network, in contrast, showed robust sensitivity regardless of the lesion depth.

This study comes with a range of strengths and limitations. First, a relatively large dataset (at least within the realm of dentistry) was used. The dataset was largely balanced with regards to the teeth-level prevalence of caries lesions, something which is different from previous studies in the field (e.g. on apical lesions or periodontal bone loss) [19, 20], likely to the benefit of the neural network, as discussed before. Second, the output of our neural network was a highlighted area where the network predicted a caries lesion; this allowed us to identify the areas the neural network assigned as carious. Our networks hence show some kind of "explainability" and can specifically guide dentists to the area of interest. Third, and as mentioned, we compared the neural network with individual dentists' performances on a hold-out test dataset. The yielded accuracies of the dentists are in line with previous data [4], which is assuring and confirms that these seven dentists are, at least to some degree, representative for the breadth of diagnostic performance one would encounter in clinical dentistry. Last, we used a clinically useful set of metrics which can inform subsequent studies, for example employing decision models, to explore the potential of deep learning to assist decision making, and its impact on treatment benefit or harm. As a limitation, both the dataset underlying our study and the dentists, as discussed, cannot claim full generalizability. Validation of our neural network on an external test set should be sought. Second, the constructed reference test was built on "fuzzy" labeling [21], i.e. on the manifold of provided annotations from different annotators. Other forms of constructing the reference may come to slightly different final labels. If possible, and especially for the test dataset, such labelling should be triangulated with data from other diagnostic methods (visual, tactile, transillumination) or a hard gold standard (like histology). Moreover, we only implemented the analysis of one bitewing image, while one could also make use of the correlation between bitewing pairs (left and right) from the same patient (clustering of lesions etc.) [22]. Last, our neural network is so far not implemented into a clinical tool, and it is prudent to say that we do not know at present what impact it may have when truly applied in patient care.

Our findings need discussion. First, dentists were far less sensitive, but slightly more specific than the neural network, and especially for initial lesions, dentists' accuracy was very limited. While for dentists, the detection of initial, e.g. small lesions is obviously a more difficult task than detecting advanced, clearly visible ones, this was not the case for the neural network. It would be interesting the visualize the activation maps of the neural network to assess why it showed this non-differential pattern [22], but also to follow-up dentists' decisions when mapping (or not) initial caries lesions. Possibly, dentists were more prudent in their decisions to mark initially carious areas (fearing over-detection and treatment) while the neural network did not show such behavior. We propose that such neural network could be

used specifically to assist dentists in detecting initial lesions, allowing early non- or micro-invasive care before these lesions progress. Dentists may use their inherent higher specificity to double check any findings and thereby avoid invasive over-treatment [23]. Moreover, we demonstrate substantial variance in the accuracy of dentists, confirming previous studies [24-26]. A neural network trained on dental bitewing radiographs could assist certain dentists with limited performance, leading to an overall acceptable diagnostic level and help to standardize their diagnostic performance, hence enhancing diagnostic and treatment quality.

In line with the above said, we suggest that future studies employ trained neural networks such as ours in a prospective and randomized design, and explore not only the accuracy resulting from its use, but also how dentists use and adopt the tool, how they interact with and how it changes their diagnostic and treatment decisions, including treatment intensity and patterns, efficiency and cost-effectiveness. Before entering clinical care, neural networks should be scrutinized along the criteria of evidence-based practice with a comprehensive set of outcomes, in a variety of settings, for their robustness, generalizability and clinical consequences [27].

In conclusion and within the limitations of this study, the trained neural network performed significantly more accurate than the majority of dentists to detect caries lesions on bitewings. Notably, though, dentists rather under-detected lesions, while the neural network over-detected to a limited degree. Especially for initial lesions, dentists' accuracy was far below that of the network. The generalizability of the network and the impact of using it on treatment decisions should be further explored.

Authors' contributions

The study was conceived by JK, SG, FS, who also setup the experimental setup. JK, SG, AGC performed the experiments. AGC conceived and implemented the teeth-level metric employed through the analysis. All authors analyzed, interpreted the data. JK, SG, AGC, FS wrote the manuscript. All authors read and approved the manuscript.

Conflict of interest

JK, FS, RG are founders of dentalXrai GmbH, a spin-out of the Charité developing AI solutions for dentistry. dentalXrai GmbH did not have any role in writing this paper. The authors are solely responsible for the contents of this paper.

Acknowledgements

We thank the dentists for their effort of labeling the image data. JK and FS are supported by a grant of the Berlin Institute of Health (DHA 2018). JK, FS, RG are co-founders of a spin-out of the Charité developing AI solutions for dentistry.

References

[1] E. Bernabe, W. Marcenes, C.R. Hernandez, J. Bailey, L.G. Abreu, V. Alipour, S. Amini, J. Arabloo, Z. Arefi, A. Arora, M.A. Ayanore, T.W. Barnighausen, A. Bijani, D.Y. Cho, D.T. Chu, C.S. Crowe, G.T. Demoz, D.G. Demsie, Z.S. Dibaji Forooshani, M. Du, M. El Tantawi, F. Fischer, M.O. Folayan, N.D. Futran, Y.C.D. Geramo, A. Haj-Mirzaian, N. Hariyani, A. Hasanzadeh, S. Hassanipour, S.I. Hay, M.K. Hole, S. Hostiuc, M.D. Ilic, S.L. James, R. Kalhor, L. Kemmer, M. Keramati, Y.S. Khader, S. Kisa, A. Kisa, A. Koyanagi, R. Lalloo, Q. Le Nguyen, S.D. London, N.D. Manohar, B.B. Massenburg, M.R. Mathur, H.G. Meles, T. Mestrovic, A. Mohammadian-Hafshejani, R. Mohammadpourhodki, A.H. Mokdad, S.D. Morrison, J. Nazari, T.H. Nguyen, C.T. Nguyen, M.R. Nixon, T.O. Olagunju, K. Pakshir, M. Pathak, N. Rabiee, A. Rafiei, K. Ramezanzadeh, M.J. Rios-Blancas, E.M. Roro, S. Sabour, A.M. Samy, M. Sawhney, F. Schwendicke, F. Shaahmadi, M.A. Shaikh, C. Stein, M.R. Tovani-Palone, B.X. Tran, B. Unnikrishnan, G.T. Vu, A. Vukovic, T.S.S. Warouw, Z. Zaidi, Z.J. Zhang, N.J. Kassebaum, Global, Regional, and National Levels and Trends in Burden of Oral Conditions from 1990 to 2017: A Systematic Analysis for the Global Burden of Disease 2017 Study, J Dent Res 99(4) (2020) 362-373.
[2] N.P.T. Innes, C.H. Chu, M. Fontana, E.C.M. Lo, W.M. Thomson, S. Uribe, M. Heiland, S. Jepsen, F. Schwendicke, A Century of Change towards Prevention and Minimal Intervention in Cariology, J Dent Res 98(6) (2019) 611-617.
[3] T. Gimenez, C. Piovesan, M.M. Braga, D.P. Raggio, C. Deery, D.N. Ricketts, K.R. Ekstrand, F.M. Mendes, Visual Inspection for Caries Detection: A Systematic Review and Meta-analysis, J Dent Res 94(7) (2015) 895-904.
[4] F. Schwendicke, M. Tzschoppe, S. Paris, Radiographic caries detection: A systematic review and meta-analysis, J Dent 43(8) (2015) 924-33.
[5] M.A. Geibel, S. Carstens, U. Braisch, A. Rahman, M. Herz, A. Jablonski-Momeni, Radiographic diagnosis of proximal caries-influence of experience and gender of the dental staff, Clin Oral Investig 21(9) (2017) 2761-2770.
[6] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521(7553) (2015) 436-44.
[7] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, Cambridge, MA, 2016.
[8] A.S. Becker, M. Marcon, S. Ghafoor, M.C. Wurnig, T. Frauenfelder, A. Boss, Deep Learning in Mammography: Diagnostic Accuracy of a Multipurpose Image Analysis Software in the Detection of Breast Cancer, Invest Radiol 52(7) (2017) 434-440.
[9] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542(7639) (2017) 115-118.
[10] V. Gulshan, L. Peng, M. Coram, M.C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P.C. Nelson, J.L. Mega, D.R. Webster, Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs, JAMA 316(22) (2016) 2402-2410.
[11] F. Schwendicke, T. Golla, M. Dreher, J. Krois, Convolutional neural networks for dental image diagnostics: A scoping review, J Dent 91 (2019) 103226.
[12] P.M. Bossuyt, J.B. Reitsma, D.E. Bruns, C.A. Gatsonis, P.P. Glasziou, L. Irwig, J.G. Lijmer, D. Moher, D. Rennie, H.C. de Vet, H.Y. Kressel, N. Rifai, R.M. Golub, D.G. Altman, L. Hooft, D.A. Korevaar, J.F. Cohen, STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies, BMJ 351 (2015) h5527.
[13] J. Mongan, L. Moy, C.E. Kahn, Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers, Radiology: Artificial Intelligence 2(2) (2020) e200029.

[14] O. Ronneberger, P. Fischer, T. Brox, Dental X-ray Image segmentation using a U-shaped Deep convolutional network, Medical Image Computing and Computer-Assisted Intervention (MICCAI) 9351 (2015) 234-243.
[15] L. Meinhold, J. Krois, R. Jordan, N. Nestler, F. Schwendicke, Clustering effects of oral conditions based on clinical and radiographic examinations, Clin Oral Investig 10.1007/s00784-019-03164-9 (2019).
[16] T. Ekert, J. Krois, F. Schwendicke, Building a mass online annotation tool for dental radiographic imagery, MIDL https://openreview.net/forum?id=SkcgYEoiG (2018(
[17] M. Tan, Q. Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural NetworksMingxing, Proceedings of the 36th International Conference on Machine Learning, https://arxiv.org/pdf/1905.11946.pdf (2019).
[18] M.M. Srivastava, P. Kumar, L. Pradhan, S. Varadarajan, Detection of Tooth caries in Bitewing Radiographs using Deep Learning, https://arxiv.org/abs/1711.07312 (2017).
[19] J. Krois, T. Ekert, L. Meinhold, T. Golla, B. Kharbot, A. Wittemeier, C. Dorfer, F. Schwendicke, Deep Learning for the Radiographic Detection of Periodontal Bone Loss, Sci Rep 9(1) (2019) 8495.
[20] T. Ekert, J. Krois, L. Meinhold, K. Elhennawy, R. Emara, T. Golla, F. Schwendicke, Deep Learning for the Radiographic Detection of Apical Lesions, J Endod 45(7) (2019) 917-922.e5.
[21] T. Walsh, Fuzzy gold standards: Approaches to handling an imperfect reference standard, J Dent 74 Suppl 1 (2018) S47-s49.
[22] F. Schwendicke, W. Samek, J. Krois, Artificial Intelligence in Dentistry: Chances and Challenges, J Dent Res 99 (7) (2020) 22034520915714.
[23] F. Schwendicke, S. Paris, M. Stolpe, Detection and treatment of proximal caries lesions: Milieu-specific cost-effectiveness analysis, J Dent 43(6) (2015) 647-55.
[24] R.P. da Silva, M.C. Meneghim, A.B. Correr, A.C. Pereira, G.M. Ambrosano, E.L. Mialhe, Variations in caries diagnoses and treatment recommendations and their impacts on the costs of oral health care, Community Dent Health 29(1) (2012) 25-8.
[25] I. Espelid, A.B. Tveit, P.J. Riordan, Radiographic caries diagnosis by clinicians in Norway and Western Australia, Community Dent Oral Epidemiol 22(4) (1994) 214-9.
[26] P. Mileman, D. Purdell-Lewis, L. van der Weele, Variation in radiographic caries diagnosis and treatment decisions among university teachers, Community Dent Oral Epidemiol 10(6) (1982) 329-334.
[27] M. Nagendran, Y. Chen, C.A. Lovejoy, A.C. Gordon, M. Komorowski, H. Harvey, E.J. Topol, J.P.A. Ioannidis, G.S. Collins, M. Maruthappu, Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies, BMJ 368 (2020) m689.

Figure legends

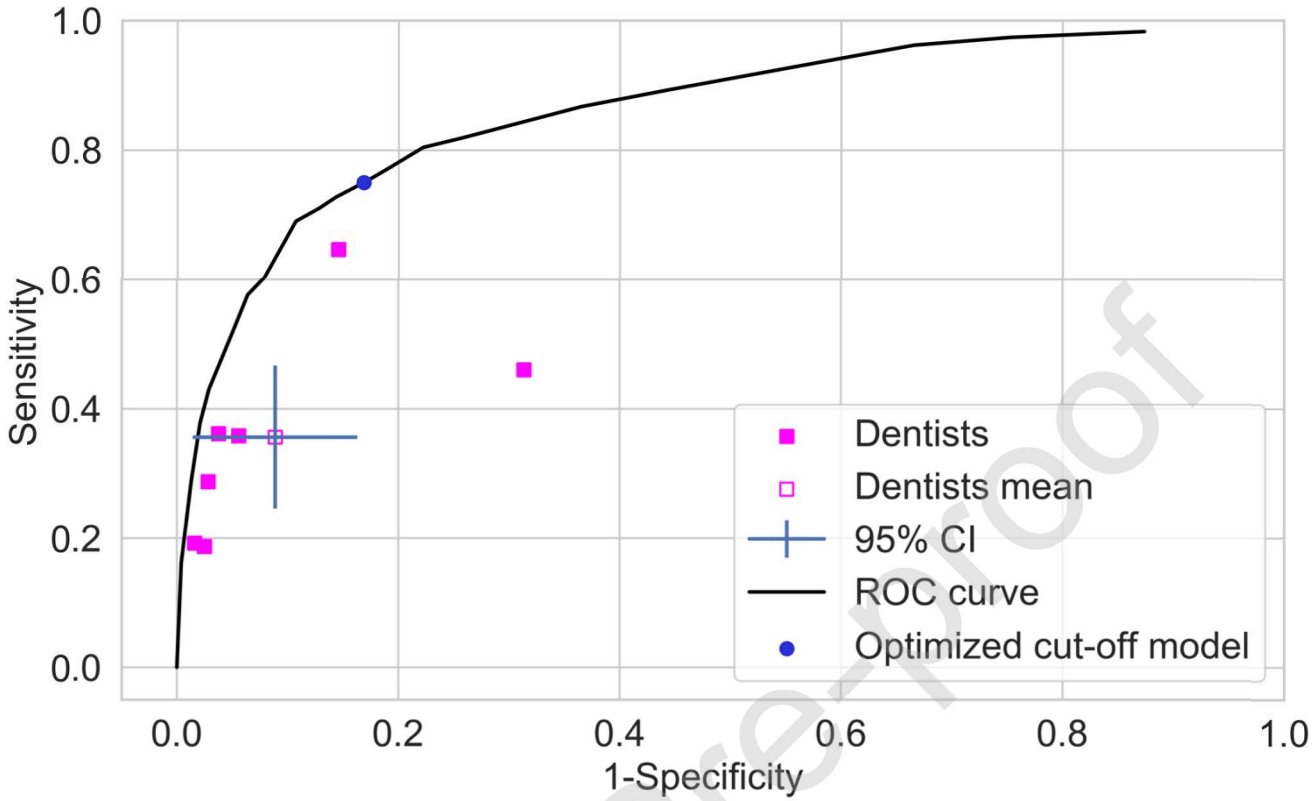


Figure 1. Receiver operating characteristic (ROC) curve. The solid line depicts the trajectory described by the neural network when evaluated (on the reference dataset) with respect to sensitivity and specificity at different cut-off values. Individual dentists' discrimination accuracy is represented by purple markers (operating points). None of the dentists outperformed the neural network. The network's sensitivity and specificity values at the optimal detection cut-off are shown by the green triangle. The cross indicates the mean and 95% of the dentists' sensitivities and specificities.

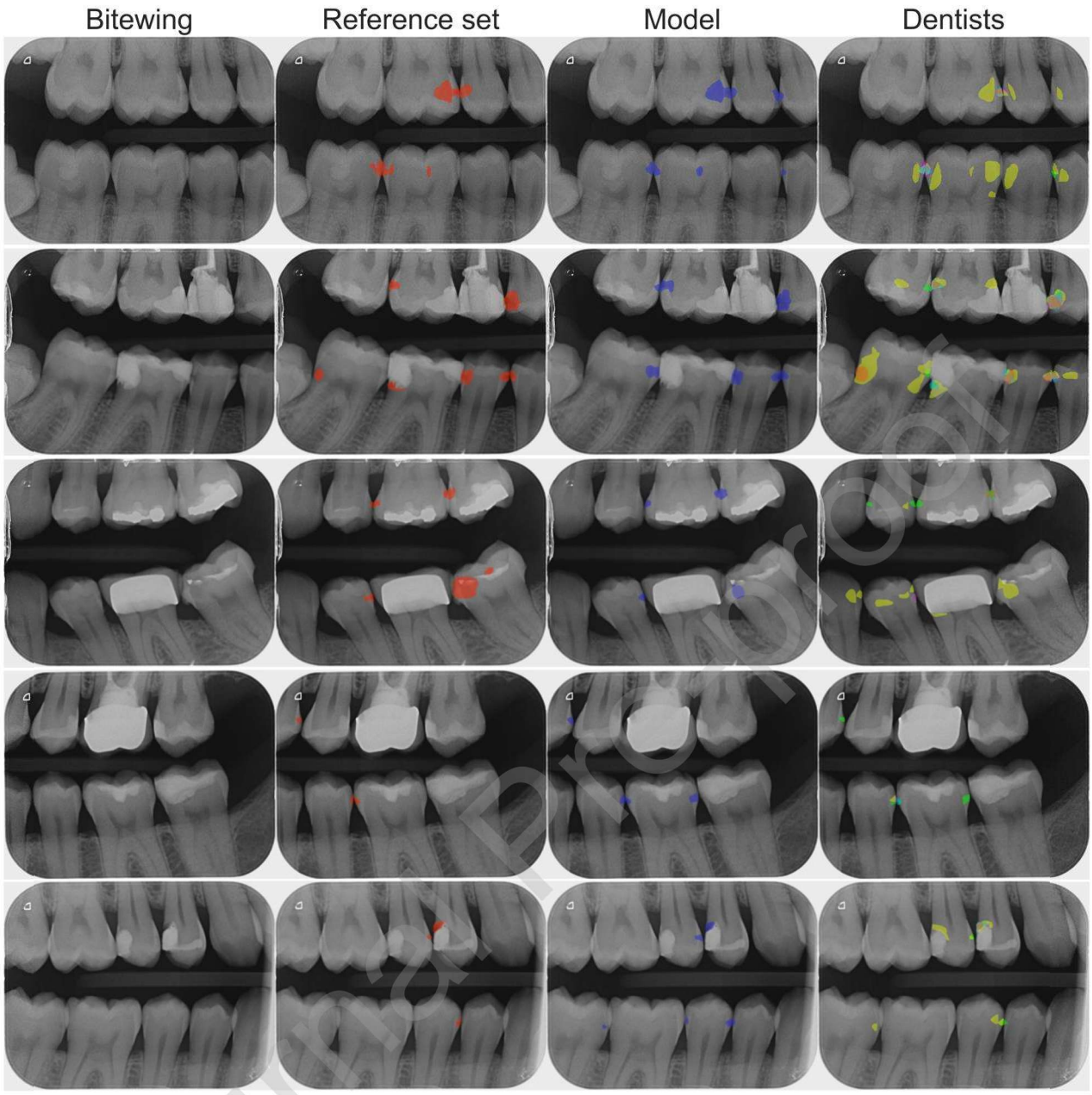


Figure 2. Exemplary neural network output and dentists' segmentation. The native bitewing, the reference test (red), the neural network's prediction (blue) and dentists' segmentation (each dentist with a different color) are shown.

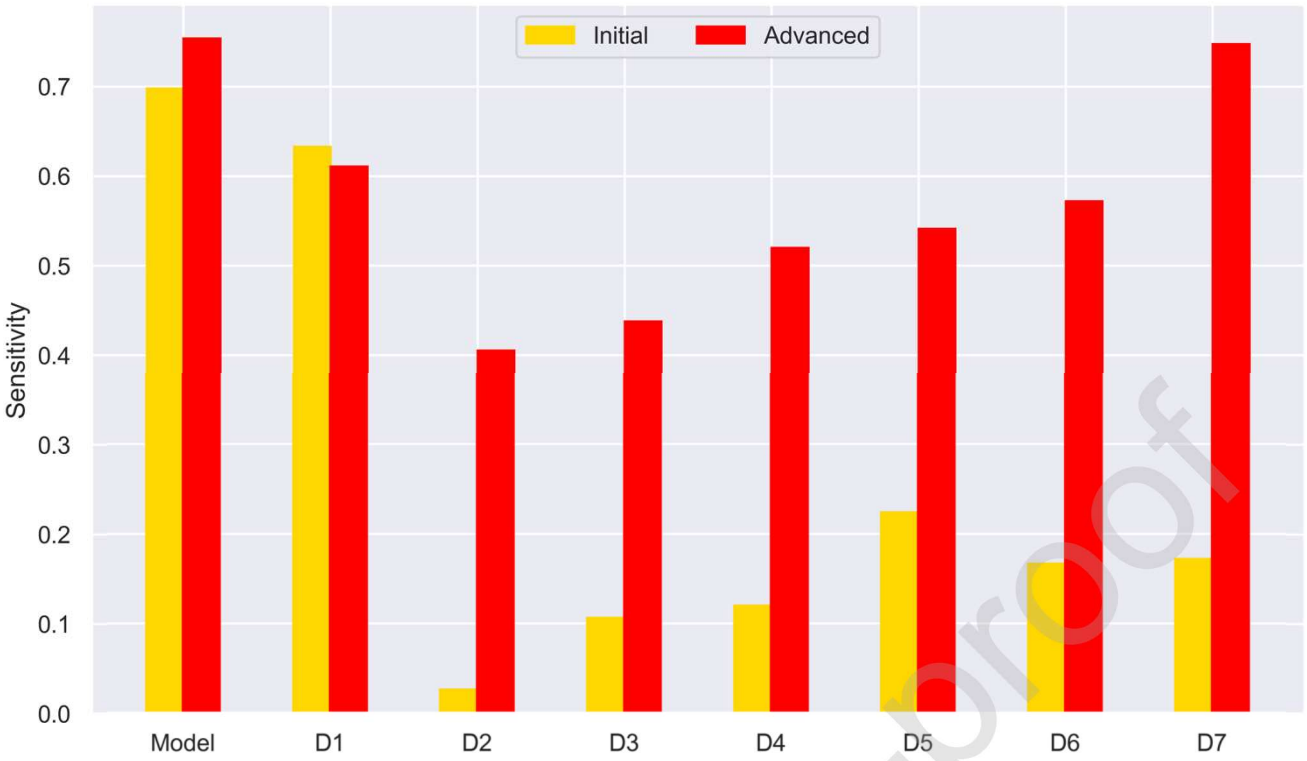


Figure 3. The sensitivities displayed by the neural network and individual dentists (D1-7) on initial (E1/2, yellow) and advanced (D2/3, red) lesions.

Appendix

A1. Model architecture

The U-Net architecture is characterized by the input and output layer of the same dimensions and skip connections between layers in the encoder (left part of the 'U') and layers in the decoder (right part of the 'U') (Fig. A1). In general, during the contractive phase the model condenses the input and the contextual information increases while the exact positional information about objects decreases. In the expanding phase contextual information flows upwards and is combined with precise information about object locations through the skip connections between encoder and decoder layers.

The U-Net version we applied is available through the git repository segmentation-models (https://github.com/qubvel/segmentation_models/tree/master/segmentation_models). The encoder (or backbone) consists of the EfficientNet-B5 network which was pretrained on the ImageNet dataset (http://www.image-net.org/). Using pretrained networks has proven to lead to faster convergence and better results especially when the amount of data is limited. The applied U-Net has more layers than the original implementation and EfficientNet-B5 was integrated into the U-Net architecture by removing the last fully connected layers.

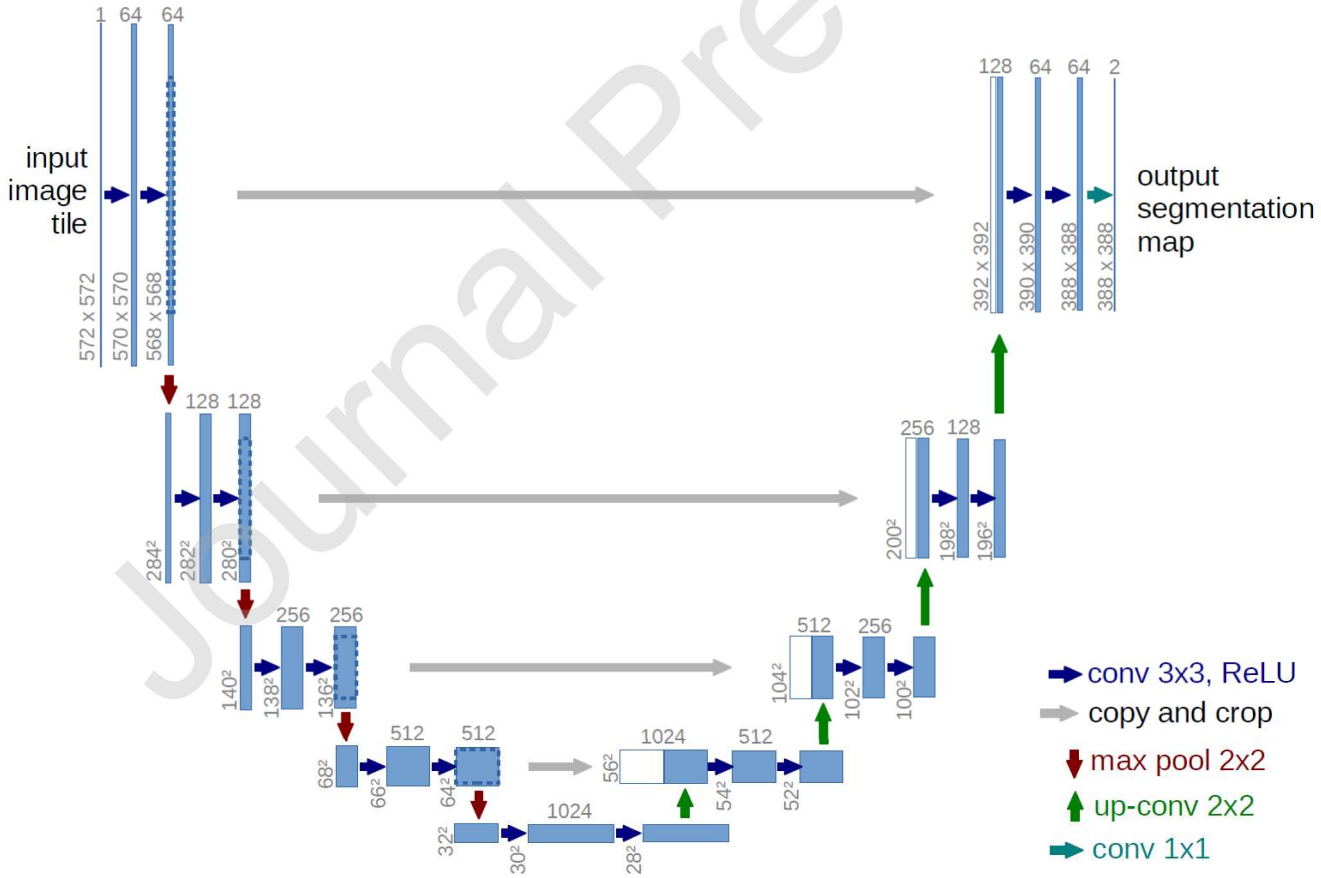


Figure A1. The schematic representation of the original U-Net model [14].

A2. Tooth level metrics

The presence of a caries lesion on a tooth is determined by the intersection of the corresponding pixel segmentation and a pixel mask of the corresponding tooth (the masks being generated by human annotation and by the output of a tooth segmentation model).

Quantification of teeth level metrics of caries segmentation was achieved by computing, for every tooth:

- $N(Ref \cap Pred)$: the total number of intersections between pixel blobs in the reference test and those corresponding to model predictions (or dentists' annotations).
- $N(Ref)$the number of overlapping free pixel blobs from the reference set.
- $N(Pred)$the number of overlapping free predicted pixel blobs.

Based on these quantities, the elements of the confusion matrix are defined as follows:

- True Positive:

$$TP = \frac{N(Ref \cap Pred)}{N(Ref) + N(Pred) + N(Ref \cap Pred)}$$

- False Positive:

$$FP = \frac{N(Pred)}{N(Ref) + N(Pred) + N(Ref \cap Pred)}$$

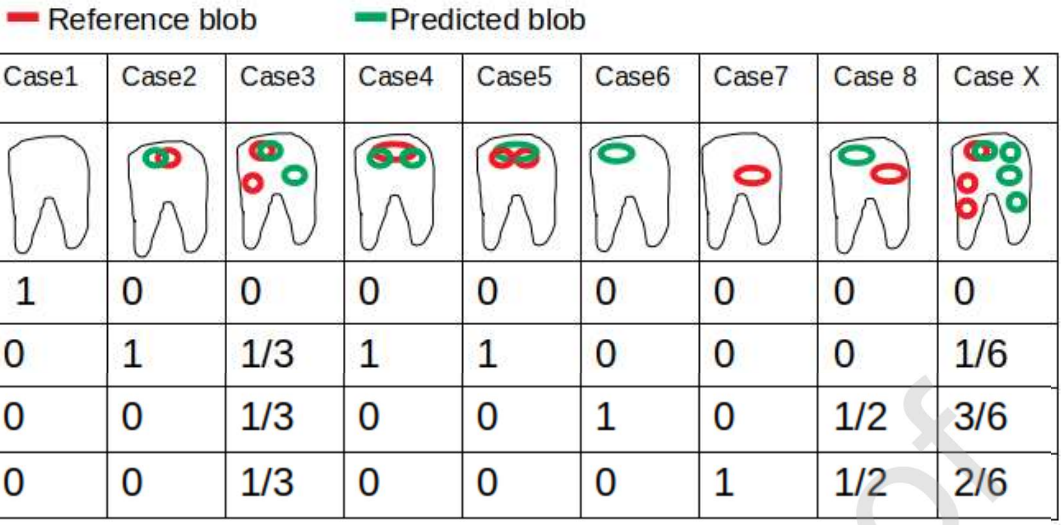
- False Negative:

$$FN = \frac{N(Ref)}{N(Ref) + N(Pred) + N(Ref \cap Pred)}$$

- True Negative:

$$TN = \left\{ \begin{matrix} 1 if N(Ref) = N(Pred) = 0 \\ 0 otherwise \end{matrix} \right\}$$

According to these definitions, a single tooth can exhibit positive values of TP, FN and FP, thereby capturing the different types of model error at the pixel blob level. The different teeth level scores are finally computed by taking the sum of the above defined quantities over all the teeth and images (see illustration in Fig. A2).



TCM	Case1	Case2	Case3	Case4	Case5	Case6	Case7	Case 8	Case X
TN	1	0	0	0	0	0	0	0	0
TP	0	1	1/3	1	1	0	0	0	1/6
FP	0	0	1/3	0	0	1	0	1/2	3/6
FN	0	0	1/3	0	0	0	1	1/2	2/6

Figure A2. Schematic illustration of the computation of teeth-level confusion matrix elements on simple representative cases of prediction and annotation blob occurrences.

Table 1. Tooth level metrics for the neural network and individual dentists. Total and surface-stratified results are shown. Note that surface-specific values are estimated based on the occurrences of positives and negatives, which may lead to score drift and differences compared with the overall mean.

Parameter	**Neural network**			**Dentists (mean; min-max)**		
	Total	Mesial	Distal	Total	Mesial	Distal
Accuracy	0.80	0.89	0.88	(0.71; 0.61-0.78)	(0.86; 0.8-0.91)	(0.85; 0.78-0.88)
Sensitivity	0.75	0.76	0.75	(0.36; 0.19-0.65)	(0.37; 0.19-0.68)	(0.37;0.20-0.67)
Specificity	0.83	0.93	0.91	(0.91; 0.69-0.98)	(0.96; 0.84-0.99)	(0.96; 0.84-0.99)
PPV	0.70	0.74	0.67	(0.75; 0.41-0.88)	(0.81; 0.44-0.99)	(0.75; 0.4-0.88)
NPV	0.86	0.93	0.94	(0.72; 0.68-0.82)	(0.85; 0.81-0.91)	(0.87; 0.84-0.9)
F1	0.73	0.75	0.71	(0.41; 0.26-0.63)	(0.47; 0.31-0.71)	(0.46; 0.31-0.69)
MCC	0.57	0.68	0.63	(0.35; 0.14-0.51)	(0.45; 0.31-0.63)	(0.44; 0.30-0.62)

Abbreviations: PPV/NPV: positive/negative predictive value, MCC Matthew's correlation coefficient. Please see the main text for the definition of the metrics.