Deep Learning for the Radiographic Detection of Apical Lesions

Thomas Ekert, MBA, HSG,[*†] *Joachim Krois, Dr. rer. nat.,*[*] *Leonie Meinhold, DDS,*[*] *Karim Elhennawy, Dr. med. dent.,*[*] *Ramy Emara, MSc,*[*] *Tatiana Golla, DDS,*[*] *and Falk Schwendicke, PhD*[*]

ABSTRACT

Introduction: We applied deep convolutional neural networks (CNNs) to detect apical lesions (ALs) on panoramic dental radiographs. **Methods:** Based on a synthesized data set of 2001 tooth segments from panoramic radiographs, a custom-made 7-layer deep neural network, parameterized by a total number of 4,299,651 weights, was trained and validated via 10 times repeated group shuffling. Hyperparameters were tuned using a grid search. Our reference test was the majority vote of 6 independent examiners who detected ALs on an ordinal scale (0, no AL; 1, widened periodontal ligament, uncertain AL; 2, clearly detectable lesion, certain AL). Metrics were the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and positive/negative predictive values. Subgroup analysis for tooth types was performed, and different margins of agreement of the reference test were applied (base case: 2; sensitivity analysis: 6). **Results:** The mean (standard deviation) tooth level prevalence of both uncertain and certain ALs was 0.16 (0.03) in the base case. The AUC of the CNN was 0.85 (0.04). Sensitivity and specificity were 0.65 (0.12) and 0.87 (0.04,) respectively. The resulting positive predictive value was 0.49 (0.10), and the negative predictive value was 0.93 (0.03). In molars, sensitivity was significantly higher than in other tooth types, whereas specificity was lower. When only certain ALs were assessed, the AUC was 0.89 (0.04). Increasing the margin of agreement to 6 significantly increased the AUC to 0.95 (0.02), mainly because the sensitivity increased to 0.74 (0.19). **Conclusions:** A moderately deep CNN trained on a limited amount of image data showed satisfying discriminatory ability to detect ALs on panoramic radiographs. *(J Endod 2019;45:917–922.)*

KEY WORDS

Artificial Intelligence; digital imaging/radiology; endodontics; mathematical modeling; radiography

SIGNIFICANCE

Dentists' diagnostic efforts for detecting ALs on panoramic radiographs may be reduced by applying CNNs. However, the CNN's sensitivity needs to be improved before clinical application.

Apical periodontitis is defined as an inflammatory process around the apex of the tooth root, mainly because of bacterial infection of the endodontic system and the subsequent response of the periapical bone tissue[1]. Apical periodontitis has been found to be highly prevalent (in 34%–61% of individuals and 3%–4% of teeth)[1–3]. Although root canal treatment may lead to healing of apical periodontitis in many cases, complete healing is not always achieved[1]. Apical periodontitis has been associated with systemic conditions such as diabetes mellitus[1].

Apical periodontitis is detected radiographically as periapical translucencies (a widened periodontal ligament or clearly detectable lesions). Such translucencies, also termed apical lesions (ALs), may be detected by targeted radiographic diagnostics (eg, after clinical determination of pulp sensitivity loss or by radiographic follow-up of endodontically treated teeth). However, oftentimes, the detection of ALs is the result of assessing radiographs taken for other purposes.

A range of options to detect ALs are available. The current standard for endodontic radiography, periapical radiographs, has only limited discriminatory ability to detect ALs when measured against a gold standard (ie, in cadaver or skull studies)[4]. Cone-beam computed tomographic (CBCT) imaging shows significantly higher discriminatory ability[5] but, considering its costs and the associated radiation dose, is only limitedly applied in general dental practice at present. Panoramic radiographs allow the assessment of not only 1 or a few teeth (as for periapical radiographs) but rather all teeth simultaneously while requiring significantly lower doses of radiation than CBCT imaging[6,7]. Although the detection of ALs on panoramic radiographs comes with limited sensitivity and negative predictive value (NPV), its specificity and positive

From the *Department of Operative and Preventive Dentistry, Charité-Universitätsmedizin Berlin, Germany; and †CODE University of Applied Sciences, Berlin, Germany

Address requests for reprints to Dr Falk Schwendicke, Department of Operative and Preventive Dentistry, Charité-Universitätsmedizin Berlin, Aßmannshauser Str 4-6, 14197 Berlin, Germany.
E-mail address: falk.schwendicke@charite.de
0099-2399/$ - see front matter

https://doi.org/10.1016/j.joen.2019.03.016

predictive value (PPV) are high, with an overall good diagnostic discriminatory ability[8–10]. Hence, detecting ALs on panoramic radiographs is a relevant daily task for dentists.

Regardless of their specific discriminatory ability, all radiographic examinations are prone to suffer from limited inter- and intraexaminer reliability[8,9]. This reliability further depends on the examiner's experience (eg, the reliability to detect ALs on CBCT imaging has been found to range between $\kappa = 0.28$ for students to $\kappa = 0.49$ for endodontic specialists)[11].

Automated assistance systems for dental radiographic image diagnostics may help to overcome issues of low reliability while achieving accuracies similar or higher of that of specialists[12]. Convolutional neural networks (CNNs) have been successfully applied for automated assessment of breast cancer in mammography[13], skin cancer in clinical skin screenings[14], or diabetic retinopathy in eye examinations[15]. In dentistry, CNNs have been applied to detect carious lesions on bitewing radiographs[16] or periodontal bone loss on periapical radiographs[17]. A detailed description of CNNs in general and their specific application in radiology have been provided elsewhere[18–20].

We aimed to apply CNNs on panoramic radiographs to detect ALs. We assumed the human effort to systematically, comprehensively, and reliably assess panoramic scans to be high; CNN application for this purpose may improve discriminatory ability and reliability while reducing this effort.

MATERIALS AND METHODS

Image Data Set

This study was built on a data set of 2001 manually cropped image segments, each focusing one particular tooth, from 85 randomly chosen and anonymized digital panoramic dental radiographs collected using Orthophos XG (Sirona, Bensheim, Germany) according to manufacturer's instructions (considering patients' sex, age, etc). Panoramic radiographs had been taken in the central radiographic unit of the Charité-Universitätsmedizin Dental Center, Berlin, Germany; all dental patients regardless of the department managing them have their radiographs taken there. Positioning was performed according to the manufacturer's instruction by experienced medical-technical assistants specialized in dental radiography. The median (minimum/maximum) age of the 85 patients was 51 (15/91) years. The mean (median, minimum/maximum) number of teeth per patient (ie, per panoramic radiograph) was 19.5 (20, 1/30). There were 30.2%, 15.5%, 27.8%, and 26.5% incisors, canines, premolars, and molars, respectively.

Data collection was ethically approved (Charité Ethics Committee EA4/080/18). Only radiographs from dentate individuals were included, and 1 examiner prescreened the resulting tooth segments for accessibility. Overall, 249 tooth segments (mainly on anterior teeth) were excluded, most of them because they heavily overlapped with the vertebrae. No further inclusion or exclusion criteria were applied (ie, the quality of the panoramic images [contrast, hazing, positioning, etc] was not used to exclude images). However, sensitivity analyses (on tooth location and on ease of assessment) were performed (see later).

Reference Test

Our reference test was the majority vote of 6 independent and experienced dentists (clinical experience 3–10 years). They assessed the images for radiographically detectable ALs. ALs were ordinal scaled (0, no AL; 1, a widened periodontal ligament, uncertain AL; and 2, clearly detectable lesion, certain AL). Dentists had been informed about the study and the diagnostic task before performing classification. Images were viewed using Sidexis 4 (Sirona) in dimly lit rooms on diagnostic screens and standardized conditions. Both measurements and examinations allowed magnification and enhancement (contrast) tools to be used. Findings were recorded in a spreadsheet. For descriptive purposes, interrater reliability was computed via Fleiss kappa[21], assuming 0–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1 as almost perfect agreement[22]. Interrater reliability was 0.48 (ie, moderate). Dentists did not revisit their records, and intrarater reliability was not assessed (we focused on having a large number of examiners for reasons of robustness instead).

In the base case, a margin of agreement of 2 was applied (ie, we assigned a positive or, respectively, a negative label to an image segment if at least 4 out of 6 examiners agreed on the diagnosis; otherwise, the image segment was removed from the data set). In a sensitivity analysis, this margin was increased (see later). Details on the agreement between examiners are provided in the Supplemental Appendix S1 and Supplemental Figures S1 and S2 (available online at www.jendodon.com).

Modeling via CNNs

Neural networks are composed of layers of mathematical functions. In a feed-forward neural network, information is passed through the network from the first (input) layer to the final (output) layer[23]. CNNs, a specialized kind of neural networks, are particularly useful to extract hierarchal features from image data by applying convolution operations. Hence, these models can extract different features of an input image such as edges, corners, and spots or increasingly complex features such as shapes and macroscopic structures and patterns. The last few network layers of a CNN typically translate these feature-filtered images into class scores, which are translated into a particular class (eg, diseased or not).

We used the Keras framework for model development. We combined convolutional layers with rectified linear units and max pooling layers. We used batch normalization[24] and dropout for model regularization. A series of fully connected dense layers and a softmax classifier built the final model layers. Details on the model development are provided in Supplemental Appendix S1.

For hyperparameter tuning, we applied a grid search[25]. We considered the number and ordering of stacked layers, the number of units in the hidden layers, the number of convolutional filters, the kernel sizes, padding, pooling size, bias and activation functions, image preprocessing, different types of optimizers and learning rates, the usage of dropout and batch normalization layers and their parameterizations and positions, the batch size, and image augmentation parameters as hyperparameters. Details can be found in the Supplemental Appendix S1 and Supplemental Table S1 (available online at www.jendodon.com).

Before feeding image data into the CNN, data were digitally preprocessed as follows:

1. each image segment was transformed to gray scale,
2. segments from the upper jaw were flipped by 180° so that in all images the crowns faced upward and the roots downward,
3. all pixel values of each image segment were normalized to a fixed range (0, 1), and
4. all image segments were resized to 64 × 64 pixels. Further image augmentation techniques such as featurewise center, zero-phase component analysis whitening, rescaling, shearing, zooming, and rotating and flipping images horizontally and vertically were applied[26]. We found image rotation (±20°), shearing (from 0.8–1.2), and zooming (from 0.8–1.2) most appropriate. Data processing was performed using Python, and third party libraries such as NumPy, pandas, scikit-image, and scikit-learn.

We split the data set into training and validation set by applying group shuffling, assuring that the images of 1 particular patient were either included in the training or the validation set. In order to reduce the detrimental effect of class imbalance on model performance[27], we oversampled image instances from the minority class (in our case, these were images in which ALs were present). We applied 10-fold repetition to evaluate the robustness of the CNN performance.

Performance Metrics

Our primary performance metric was the area under the receiver operating characteristic (ROC) curve (AUC), which relates to the ability of a test (ie, a model) to make correct classifications (healthy/diseased). Secondary metrics were the sensitivity and specificity and the positive and negative predictive values (PPVs/NPVs). Note that the PPV/NPV are heavily affected by the prevalence of the condition of interest (ie, ALs); hence, they are useful to describe the diagnostic value of a test (a model) in a particular population but may vary heavily between populations. Details on the performance metrics are provided in Supplemental Appendix S1 (available online at www.jendodon.com).

Sensitivity Analyses

In the base case, the CNN was trained and tested for detecting both uncertain (a widened periodontal ligament) and certain (a clearly detectable lesion) ALs on all accessible teeth, with the majority vote of examiners constituting the reference test (see earlier). We additionally evaluated the model's discrimination ability for certain ALs only as well as for different tooth types (molars, premolars, canines, and incisors). The rationale for these analyses was that diagnostic uncertainty may (or may not) be reflected in the CNN performance too and that because of the radiographic image generation process, different tooth types are differently difficult to assess. We further assessed how the diagnostic agreement of dentists as a proxy for image accessibility and ease of interpretation would impact the CNN's performance by increasing the threshold of agreement for the reference test (to 6 dentists; ie, all dentists needing to agree on their classification; otherwise, the image was discarded). Reporting of this study follows the Standards for Reporting Diagnostic accuracy studies guideline[28].

RESULTS

The mean (standard deviation [SD]) prevalence of both certain and uncertain ALs in the base case (margin of agreement of 2) was 0.16 (0.03). An AL was more prevalent in molars than premolars, canines, or incisors (Table 1).

We used a 7-layer feed-forward CNN with a total number of 4,299,651 trainable weights to predict ALs (Fig. 1). The base case model was trained on average (SD) on 2238 (56) images and validated on 341[24] images, respectively.

In the base case, the mean (SD) AUC of the CNN was 0.85 (0.04) at a sensitivity of 0.65 (0.12) and a specificity of 0.87 (0.04). The resulting PPV was 0.49 (0.10), and the NPV was 0.93 (0.03). The corresponding ROC curves are shown in Figure 2.

The CNN was trained on all teeth, and in a sensitivity analysis, we applied it only to a particular tooth type (Table 1). The AUC for incisors, canines, premolars, and molars was 0.82 (0.06), 0.86 (0.08), 0.85 (0.06), and 0.84 (0.06), respectively. In molars, sensitivity was significantly higher than in other tooth types, whereas specificity was lower. However, the resulting PPV remained limited between 0.42 (0.19) for incisors and 0.56 (0.20) for canines. Conversely, the NPV was high (>0.90) for all tooth types.

By including only images that all examiners agreed on (margin of agreement = 6), the number of available images decreased (Table 1). The AUC increased significantly to 0.95 (0.02), mainly because the sensitivity increased to 0.74 (0.19). Consequently, the PPV increased significantly to 0.67 (0.14). The corresponding ROC curves are shown in Supplemental Figure S3 (available online at www.jendodon.com). There were only limited differences in the tooth type–wise analysis compared with the base case (ie, sensitivity was higher in molars, whereas specificity was lower).

When assessing only certain ALs, the prevalence decreased to 0.07 (0.02). In this case, the AUC significantly increased to 0.89 (0.04) (Supplemental Table S2 and Supplemental Fig. S4 [ROC curve] are available online at www.jendodon.com). This was caused by a significant increase to 0.71 (0.14) in sensitivity, whereas specificity slightly decreased to 0.84 (0.07). As a result of the low prevalence, PPV decreased to values between 0.21 and 0.34 depending on the tooth type (tooth type–wise analyses yielded similar results as those described for the base case).

DISCUSSION

The radiographic detection of ALs is subject to large variation between examiners (which we confirm in this study); moreover, the discriminatory ability is highly dependent on examiners' experience. The application of CNNs to assist in the detection of ALs may

TABLE 1 - The Detection of Both Uncertain and Certain Apical Lesions

Teeth	Reference test	Prevalence validation set	Images training set	Images validation set	AUC	Sensitivity	Specificity	PPV	NPV
All teeth	Majority (2)	0.16 ± 0.03	2238 ± 56	341 ± 24	0.85 ± 0.04	0.65 ± 0.12	0.87 ± 0.04	0.49 ± 0.10	0.93 ± 0.03
Incisors	Majority (2)	0.08 ± 0.04	2238 ± 56	102 ± 8	0.82 ± 0.06	0.55 ± 0.18	0.92 ± 0.05	0.42 ± 0.19	0.96 ± 0.03
Canines	Majority (2)	0.10 ± 0.03	2238 ± 56	51 ± 5	0.86 ± 0.08	0.52 ± 0.22	0.96 ± 0.03	0.56 ± 0.20	0.95 ± 0.03
Premolars	Majority (2)	0.16 ± 0.05	2238 ± 56	93 ± 5	0.85 ± 0.06	0.50 ± 0.12	0.90 ± 0.04	0.49 ± 0.17	0.90 ± 0.04
Molars	Majority (2)	0.27 ± 0.04	2238 ± 56	94 ± 9	0.84 ± 0.06	0.80 ± 0.14	0.70 ± 0.08	0.51 ± 0.07	0.90 ± 0.07
All teeth	Majority (6)	0.13 ± 0.04	1331 ± 38	195 ± 17	0.95 ± 0.02	0.74 ± 0.19	0.94 ± 0.04	0.67 ± 0.14	0.95 ± 0.04
Incisors	Majority (6)	0.07 ± 0.03	1331 ± 38	63 ± 5	0.92 ± 0.08	0.64 ± 0.30	0.96 ± 0.04	0.64 ± 0.27	0.97 ± 0.03
Canines	Majority (6)	0.09 ± 0.06	1331 ± 38	37 ± 3	0.96 ± 0.03	0.52 ± 0.31	0.98 ± 0.03	0.73 ± 0.33	0.95 ± 0.04
Premolars	Majority (6)	0.11 ± 0.05	1331 ± 38	55 ± 7	0.95 ± 0.04	0.63 ± 0.31	0.95 ± 0.03	0.55 ± 0.21	0.94 ± 0.05
Molars	Majority (6)	0.31 ± 0.08	1331 ± 38	41 ± 6	0.94 ± 0.03	0.87 ± 0.14	0.84 ± 0.13	0.74 ± 0.15	0.94 ± 0.07

AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value.
Majority refers to majority voting scheme assigning apical lesions as present/not present based on 6 independent examiners. A majority of 2 or 6 votes was demanded. The AUC, sensitivity, specificity, and PPVs and NPVs as well as the standard deviation values are shown. In the base case, all teeth were included during training, and both uncertain and certain apical lesions included in our analyses. For sensitivity analyses, the model was applied to specific subsets of teeth, and only images with full agreement of examiners (margin of agreement of 6) were included. Note that the number of images in the training set varies because of oversampling of the minority (prevalent) class.

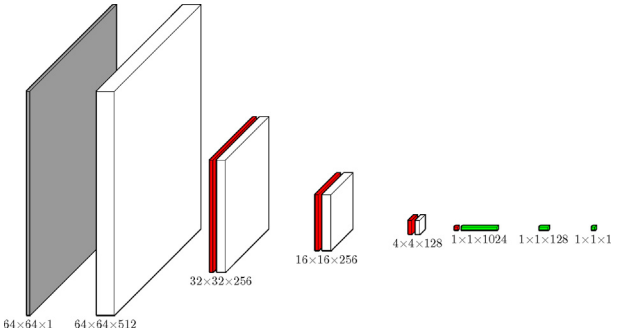


FIGURE 1 – Model architecture. The model consists of a series of chained layers (convolutional layers [*white*], max pooling layers [*red*], and fully connected layers [*green*]). The numbers indicate the width, height, and depth of each particular layer. Information is passed from the raw image data (*gray layer*) forward through the network. By stacking convolutional layers and applying a pooling operation, the extracted features become larger and more complex. The last few network layers cast the feature-filtered image data to votes (ALs being prevalent or not).

significantly improve reliability and allow all dentists to reach accuracies similar or superior to experienced specialists. Moreover, it may decrease the diagnostic efforts by saving assessment time and allowing semiautomated documentation. In the present study, we found a 7-layer CNN trained on a limited amount of labeled imagery data to have satisfying discriminatory ability (measured by AUC) for detecting ALs. The specific findings of our study need to be discussed.

The CNN's sensitivity was limited. Such limited sensitivity is related to the low prevalence of ALs in the data set, hence providing only limited chances of training the CNN. We countered this problem by oversampling the minority classes, and future studies using larger training data sets may overcome this issue. However, when applying the present CNN, a certain underdetection is likely. Notably, the associated NPV was robust, mainly because most images were truly unaffected (ie, negatively labeled).

However, the larger obstacle for a routine clinical application will be the limited PPV, which was the result of the low prevalence and occurred despite our CNN having relatively high specificity. The relevance of prevalence on the usefulness (expressed as PPV/NPV) of the CNN was shown in our sensitivity analysis in which only certain ALs were considered as the target condition. Although the CNN's overall discriminatory ability (AUC) increased (because of higher sensitivity), the PPV decreased in parallel to the decrease in prevalence. Our study highlights the need to consider the joint effects of discriminatory ability and prevalence on a test's clinical usefulness and to assess the effects of different detection thresholds on this usefulness.

We further confirmed the relevance of tooth types on discriminatory ability. Tooth type–specific accuracies have been found before[8,9] and are the result of different prevalence but also different discriminatory abilities of a CNN for different teeth (mainly as a result of structure overlapping, angulation, and so on). On molars, the CNN's sensitivity was higher than on other teeth, whereas the specificity was lower; both the PPV and the NPV remained nearly unaffected given the impact of changes (increases) in prevalence. Hence, the resulting usefulness of our model did only limitedly differ between tooth types.

A number of limitations and future directions need discussion. First, in general, deep neural networks are powerful but are also opaque prediction models. Their complex and nonlinear structure makes it difficult to reason about the inherent decision-making process. Visualizing what image features are operationalized by the CNN may enhance our

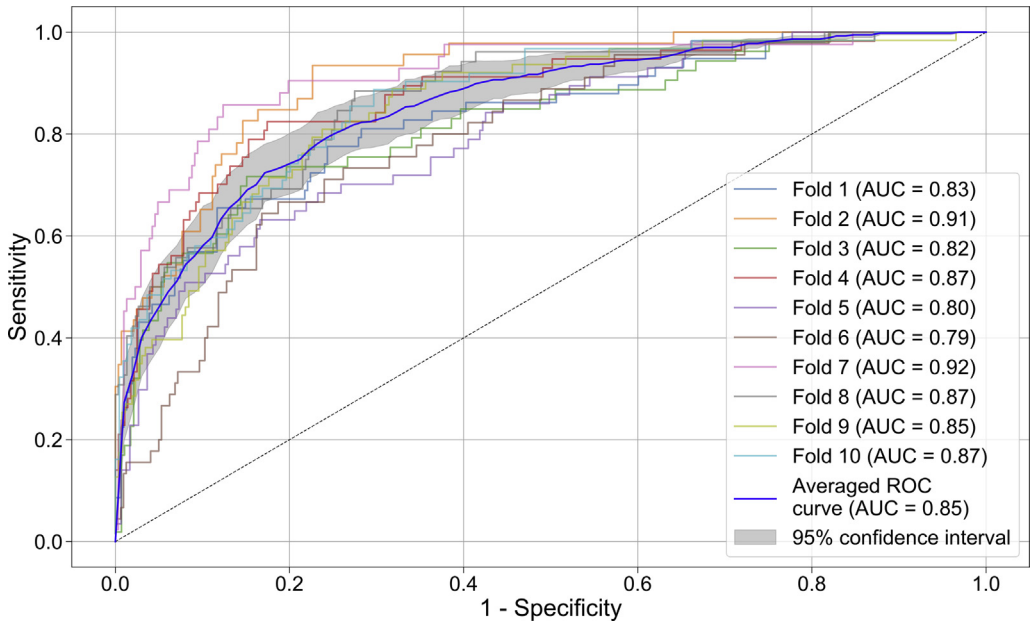


FIGURE 2 – ROC curves for the base case model. The CNN was evaluated against the reference test with respect to sensitivity (the proportion of positives that are correctly identified as such) and specificity (the proportion of negatives that are correctly identified as such). The colored curves indicate the discrimination ability in each validation fold. The *bold blue line* represents the mean discrimination ability; the *gray area* corresponds to the 95% confidence interval, respectively. The discrimination ability is further summarized by the AUC.

understanding of the decision made and may help to confirm or refute the medical logic behind it. Second, and as mentioned earlier, the reliability of the examiners constituting the reference test was limited (as can be expected as discussed). However, we accounted for that by having a relatively large number of examiners, increasing the robustness of our reference test. The relevance of examiner agreement on a CNN's discriminatory ability was confirmed in our sensitivity analysis in which only images with full agreement of the examiners were considered. In this case, the AUC was 0.95 (ie, very high), mainly because the sensitivity increased, whereas the prevalence remained similarly high at 13%. The issue of a fuzzy gold standard may additionally be overcome by triangulation of radiographic findings with clinical assessments[29]. Third, we applied a straightforward custom-made model architecture because we found that more complex, state-of-the-art architectures caused overfitting at the expense of discriminatory ability. Increasing the size of the image data set may help to overcome this issue. Fourth, our study used relatively broad inclusion criteria for the images to be assessed and did only very limitedly exclude images because of quality and so on. This was decided a priori because 1 relevant task of a CNN may be to direct dentists' diagnostic effort toward images that are hard to assess. Future studies should investigate the impact of factors like image projection or quality contrast on the discrimination ability. Notably, we had excluded images of too poor quality to assess by dentists (because assigning a reference test was found impossible in this case). It may be worthwhile to assess if using alternative data to construct such a reference test and applying the CNN to these data yields accurate predictions (ie, if the CNN can nevertheless evaluate such images) or not. Fifth, the impact of using a CNN during the diagnostic process on the subsequent decision making in practice should be explored. Last, we assessed manually cropped image segments from panoramic radiographs. Assessing the whole dentition instead may improve the discriminatory ability because most oral conditions are known to be clustered (correlated) within 1 mouth[30].

CONCLUSIONS

According to the methodology applied in this study, a moderately deep CNN trained on a limited amount of image data showed satisfying discriminatory ability to detect ALs on panoramic radiographs. For detecting ALs, applying a CNN may reduce dentists' diagnostic efforts; however, sensitivity needed to be improved before clinical application.

ACKNOWLEDGMENTS

Supported by the Berlin Institute of Health (Digital Health Accelerator 2018) (F.S. and J.K.).

The authors deny any conflicts of interest related to this study.

SUPPLEMENTARY MATERIAL

Supplementary material associated with this article can be found in the online version at www.jendodon.com (https://doi.org/10.1016/j.joen.2019.03.016).

REFERENCES

1. Segura-Egea JJ, Martin-Gonzalez J, Castellanos-Cosano L. Endodontic medicine: connections between apical periodontitis and systemic diseases. Int Endod J 2015;48:933–51.
2. Huumonen S, Suominen AL, Vehkalahti MM. Prevalence of apical periodontitis in root filled teeth: findings from a nationwide survey in Finland. Int Endod J 2017;50:229–36.
3. Connert T, Truckenmuller M, ElAyouti A, et al. Changes in periapical status, quality of root fillings and estimated endodontic treatment need in a similar urban German population 20 years later. Clin Oral Investig 2019;23:1373–82.
4. Kanagasingam S, Hussaini HM, Soo I, et al. Accuracy of single and parallax film and digital periapical radiographs in diagnosing apical periodontitis-a cadaver study. Int Endod J 2017;50:427–36.
5. Leonardi Dutra K, Haas L, Porporatti AL, et al. Diagnostic accuracy of cone-beam computed tomography and conventional radiography on apical periodontitis: a systematic review and meta-analysis. J Endod 2016;42:356–64.
6. Grandreviewresearch. Dental X-Ray Market Analysis By Product (Analog, Digital), By Type (Intraoral - Bitewing, Periapical, Occlusal; Extraoral - Panoramic, Cone Beam Computed Tomography), By Application (Medical, Cosmetic, Forensic), And Segment Forecasts, 2018 - 2024. 2016. Available at: https://www.grandviewresearch.com/industry-analysis/dental-x-ray-market. Accessed May 18, 2019.
7. KZBV. KZBV Jahrbuch 2017. Available at: www.kzbv.de. Accessed May 29, 2019.
8. Nardi C, Calistri L, Grazzini G, et al. Is panoramic radiography an accurate imaging technique for the detection of endodontically treated asymptomatic apical periodontitis? J Endod 2018;44:1500–8.
9. Nardi C, Calistri L, Pradella S, et al. Accuracy of orthopantomography for apical periodontitis without endodontic treatment. J Endod 2017;43:1640–6.
10. Ahlqwist M, Halling A, Hollender L. Rotational panoramic radiography in epidemiological studies of dental health. Comparison between panoramic radiographs and intraoral full mouth surveys. Swed Dent J 1986;10:73–84.

11. Parker JM, Mol A, Rivera EM, Tawil PZ. Cone-beam computed tomography uses in clinical endodontics: observer variability in detecting periapical lesions. J Endod 2017;43:184–7.

12. Lin PL, Huang PY, Huang PW. Automatic methods for alveolar bone loss degree measurement in periodontitis periapical radiographs. Comp Methods Programs Biomed 2017;148:1–11.

13. Becker AS, Marcon M, Ghafoor S, et al. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. Invest Radiol 2017;52:434–40.

14. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–8.

15. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402–10.

16. 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.

17. Lee JH, Kim DH, Jeong SN, Choi SH. Diagnosis and prediction of periodontally compromised teeth using a deep learning-based convolutional neural network algorithm. J Periodontal Implant Sci 2018;48:114–23.

18. Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge, MA: MIT Press; 2016.

19. Mazurowski M, Buda M, Saha A, Bashir M. Deep learning in radiology: an overview of the concepts and a survey of the state of the art. J Magn Reson Imaging 2019;49:939–954..

20. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.

21. Fleiss J. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76:378–82.

22. Landis J, Koch G. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.

23. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.

24. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv 2015: arXiv:1502.03167v3 [cs.LG], 2015.

25. Claesen M, De Moor B. Hyperparameter search in machine learning. Available at: https://arxiv.org/pdf/1502.02127v2.pdf 2015. Accessed May 18, 2019.

26. Hussain Z, Gimenez F, Yi D, Rubin D. Differential data augmentation techniques for medical imaging classification tasks. Annual Symposium proceedings AMIA Annu Symp Proc 2018;2017:979–84.

27. Buda M, Maki A, Mazurowski M. A systematic study of the class imbalance problem in convolutional neural network. Neural Netw 2018;106:249–59.

28. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351:h5527.

29. Walsh T. Fuzzy gold standards: Approaches to handling an imperfect reference standard. J Dent 2018;74(Suppl 1):S47–9.

30. Masood M, Masood Y, Newton JT. The clustering effects of surfaces within the tooth and teeth within individuals. J Dent Res 2015;94:281–8.

SUPPLEMENTAL APPENDIX S1. METHODS DETAILS

Reference Test

Six independent examiners assessed uncertain and certain ALs (Supplemental Fig. S1). The examiners showed moderate agreement at margin of agreement 2 (Fleiss kappa = 0.48), substantial agreement at margin of agreement 4 (0.61), and full agreement at margin of agreement 6 (1.0). In the majority of images, all 6 examiners agreed with each other (54.6%; Supplemental Fig. S2).

CNN Model Architecture

We trained a 7-layer CNN with 4,299,651 trainable weights. Its visual representation is shown in Figure 1 and details are given in Supplemental Table S1. The model architecture was determined using a grid search. For model architecture development, we evaluated the model candidate's performance with respect to the AUC. We evaluated different numbers of neuronal units (16–2048 in powers of 2) and the number of filters (16–2048 in powers of 2) for each particular convolutional layer. We further applied different kernel sizes (2×2 to 5×5) and evaluated different configurations of the max pooling layers (2×2 to 4×4). As activation functions, we used rectified linear units and sigmoid. Dropout layers were evaluated using dropout rates from 0.1–0.9 in steps of 0.1. Furthermore, we added batch normalization layers to evaluate their benefits with respect to overfitting and convergence. We used binary cross entropy as loss function together with the Adam or the RMSprop optimizer. We evaluated different learning rates (0.0002, 0.0001, and 0.001) and images sizes (64×64 and 128×128) and used batch sizes from 1 to 128 in powers of 2. The final model was trained with an Adam optimizer with a learning rate of 0.0001, an image size of 64×64, a batch size of 32, and for 100 epochs. Training was performed using Ubuntu 16.04 LTS and a Nvidia GTX 1080 TI GPU (Nvidia Corporation, Santa Clara, CA).

Performance metrics

We evaluated the interrater reliability by computing the Fleiss kappa (equation 1), which is an extension of the Scott pi[3], and assesses the reliability of agreement of nominal scale ratings among more than 2 examiners[1]. Hence, it is a measure for within-group reliability.

$$\kappa_{Fleiss} = \frac{P^* - P_e^*}{1 - P_e^*} \tag{1}$$

The terms P^* and P_e^* are the overall agreement probability and the probability of agreement caused by chance, respectively[2].

We used different model performance metrics such as the AUC; sensitivity (equation 2); specificity (equation 3), also referred to as recall; and the PPV (also referred to as precision) (equation 4) and NPV (equation 5)[4].

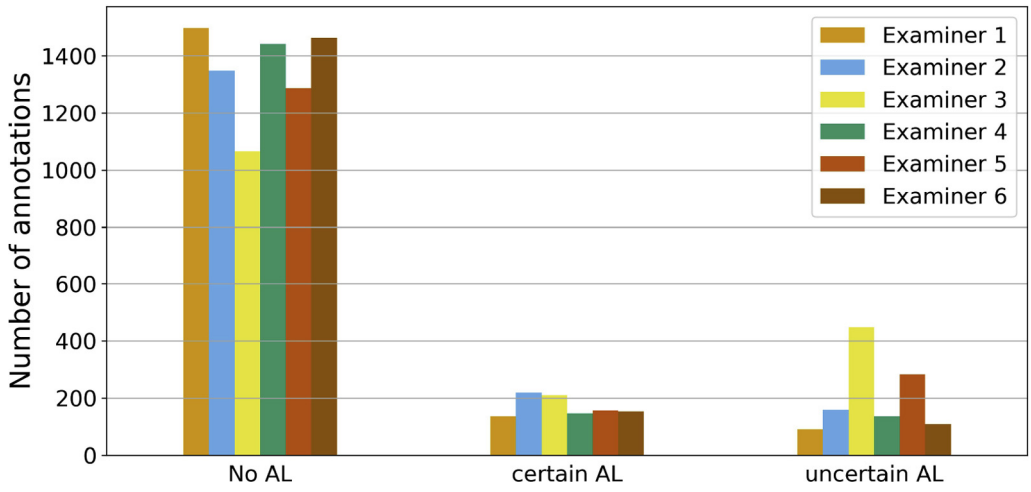
$$sensitivity\ (recall) = \frac{TP}{TP+FN} \tag{2}$$

$$specificity = \frac{TN}{TN+FP} \tag{3}$$

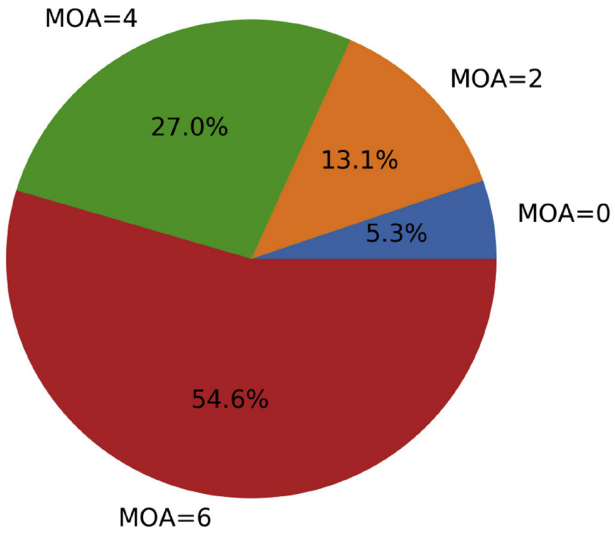
$$PPV\ (precision) = \frac{TP}{TP+FP} \tag{4}$$

$$NPV = \frac{TN}{TN+FN} \tag{5}$$

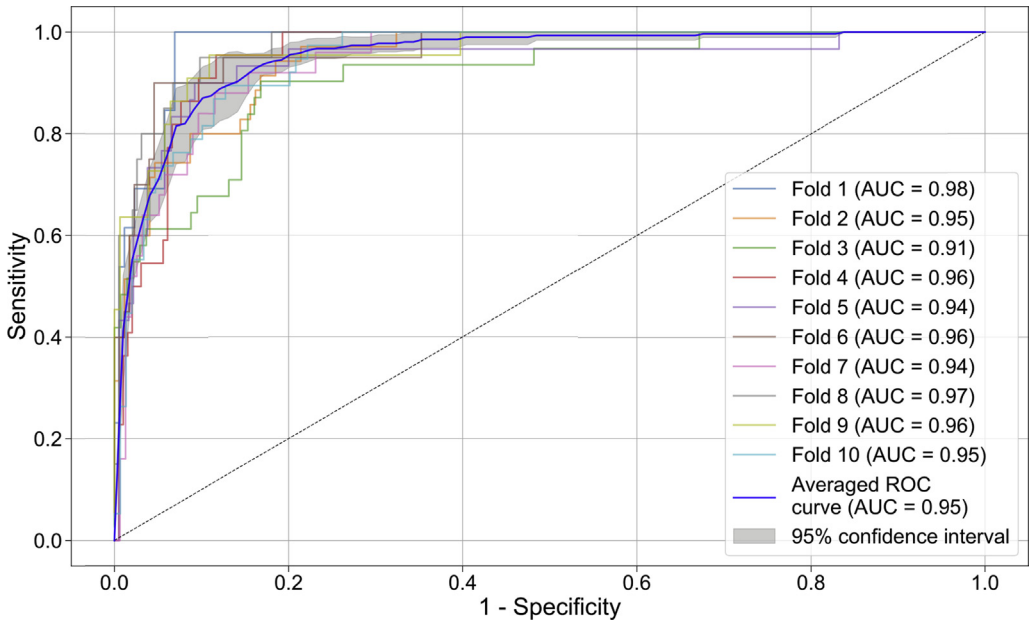
where *TP* and *TN* denotes true positive and true negative classifications and *FP* and *FN* refer to false-positive and false-negative classifications, respectively. The AUC relates to a classifier's ability to avoid false classification. The sensitivity (recall) accounts for the classifier's ability to identify positive labels, and specificity accounts for the classifier's ability to identify negative labels. The PPV (precision) accounts for the class agreement of the data labels with the positive labels given by the classifier, whereas the NPV accounts for the class agreement of the data labels with the negative labels given by the classifier[4].



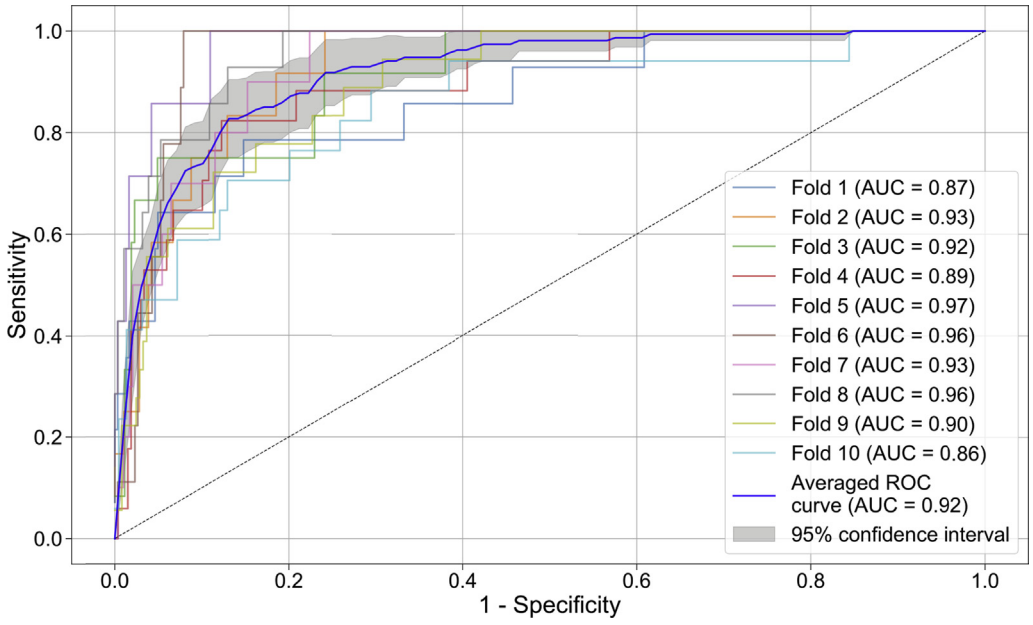
SUPPLEMENTAL FIGURE S1 – A histogram for ALs determined by 6 independent examiners.



SUPPLEMENTAL FIGURE S2 – Margin of agreement when assessing ALs by 6 independent examiners. We calculate the margin of agreement by the difference of the vote counts. The figure shows that for 54.6% of all image segments all 6 examiners agreed on the diagnostic class ($6 - 0 = 6$). Moreover, on 27% of the images, 5 examiners agreed on the diagnostic class (prevalent or not), whereas 1 examiner disagreed ($5 - 1 = 4$). On 13.1% and 5.3% of the images, the vote count was 4 to 2 and ($4 - 2 = 2$) and 3 to 3 ($3 - 3 = 0$), respectively.



SUPPLEMENTAL FIGURE S3 – ROC curves for both uncertain and certain ALs with a margin of agreement of 6. The CNN was evaluated against the reference test with respect to sensitivity (the proportion of positives that are correctly identified as such) and specificity (the proportion of negatives that are correctly identified as such). The *colored curves* indicate the discrimination ability in each validation fold. The *bold blue line* represents the mean discrimination ability over 10 folds; the *gray area* corresponds to the 95% confidence interval, respectively. The discrimination ability is further summarized by the AUC.



SUPPLEMENTAL FIGURE S4 – ROC curves for certain AL with a margin of agreement of 6. The CNN was evaluated against the reference test with respect to sensitivity (the proportion of positives that are correctly identified as such) and specificity (the proportion of negatives that are correctly identified as such). The *colored curves* indicate the discrimination ability in each validation fold. The *bold blue line* represents the mean discrimination ability over 10 folds; the *gray area* corresponds to the 95% confidence interval, respectively. The discrimination ability is further summarized by the AUC.

SUPPLEMENTAL TABLE S1 - Model Layers and Hyperparameters

Layer count	Layer type	Output shape in final model	Kernel/pooling size in final model	No. of trainable weights in final model
	Input	64, 64, 1		0
1	Conv2D	64, 64, 512	3, 3	5120
	ReLU	64, 64, 512	—	0
	Max Pooling	32, 32, 512	2, 2	0
2	Conv2D	32, 32, 256	3, 3	1179904
	ReLU	32, 32, 256	—	0
	Max Pooling	16, 16, 256	2, 2	0
3	Conv2D	16, 16, 256	3, 3	590080
	ReLU	16, 16, 256	—	0
	Max Pooling	8, 8, 256	2, 2	0
4	Conv2D	8, 8, 128	3, 3	295040
	ReLU	8, 8 128	—	0
	Max Pooling	4, 4, 128	2, 2	0
	Flatten	2048	—	0
5	Dense	1024		2098176
	ReLU	1024		0
	DropOut (0.5)	1024	—	0
6	Dense	128		131200
	ReLU	128		0
	DropOut (0.7)	128	—	0
7	Dense	1		129
	Batch normalization	1	—	4
	Softmax	1	—	0
	DropOut (0.5)	1	—	0
	Output	1	—	0

SUPPLEMENTAL TABLE S2 - Certain Apical Lesions and Sensitivity Analyses

Teeth	Reference test	Prevalence validation set	Images training set	Images validation set	AUC	Sensitivity	Specificity	PPV	NPV
All teeth	Majority (2)	0.07 ± 0.02	2536 ± 54	353 ± 24	0.89 ± 0.04	0.71 ± 0.14	0.84 ± 0.07	0.29 ± 0.09	0.97 ± 0.01
Incisors	Majority (2)	0.03 ± 0.02	2536 ± 54	105 ± 8	0.86 ± 0.26	0.69 ± 0.38	0.91 ± 0.05	0.22 ± 0.16	0.99 ± 0.01
Canines	Majority (2)	0.04 ± 0.03	2536 ± 54	54 ± 4	0.97 ± 0.04	0.62 ± 0.44	0.94 ± 0.04	0.34 ± 0.26	0.99 ± 0.02
Premolars	Majority (2)	0.07 ± 0.02	2536 ± 54	95 ± 6	0.82 ± 0.05	0.50 ± 0.24	0.84 ± 0.09	0.21 ± 0.13	0.96 ± 0.02
Molars	Majority (2)	0.13 ± 0.03	2536 ± 54	99 ± 12	0.87 ± 0.04	0.82 ± 0.15	0.71 ± 0.13	0.33 ± 0.09	0.96 ± 0.03
All teeth	Majority (6)	0.05 ± 0.02	2162 ± 53	292 ± 24	0.92 ± 0.04	0.40 ± 0.24	0.96 ± 0.05	0.43 ± 0.20	0.97 ± 0.01
Incisors	Majority (6)	0.01 ± 0.01	2162 ± 53	91 ± 10	0.93 ± 0.09	0.10 ± 0.30	0.98 ± 0.02	0.01 ± 0.04	0.99 ± 0.01
Canines	Majority (6)	0.04 ± 0.03	2162 ± 53	48 ± 4	0.94 ± 0.05	0.15 ± 0.30	0.98 ± 0.02	0.12 ± 0.20	0.97 ± 0.03
Premolars	Majority (6)	0.04 ± 0.03	2162 ± 53	81 ± 5	0.89 ± 0.07	0.32 ± 0.33	0.96 ± 0.04	0.25 ± 0.31	0.97 ± 0.02
Molars	Majority (6)	0.10 ± 0.03	2162 ± 53	73 ± 9	0.88 ± 0.06	0.53 ± 0.28	0.90 ± 0.12	0.50 ± 0.21	0.95 ± 0.02

Majority refers to majority voting scheme assigning apical lesions as present/not present based on 6 independent examiners. A majority of 2 or 6 votes was demanded.

SUPPLEMENTAL APPENDIX S1 REFERENCES

1. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76:378–82.
2. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 2008;61:29–48.
3. Scott WA. Reliability of content analysis: the case of nominal scale coding. Public Opin Q 1955;19:321.
4. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag 2009;45:427–37.