Prediction of pulp exposure before caries excavation using artificial intelligence
Deep learning-based image data versus standard dental radiographs
Ramezanzade, Shaqayeq; Dascalu, Tudor Laurentiu; Ibragimov, Bulat; Bakhshandeh, Azam; Bjørndal, Lars

Published in:
Journal of Dentistry

DOI:
10.1016/j.jdent.2023.104732

Publication date:
2023

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Prediction of pulp exposure before caries excavation using artificial intelligence: Deep learning-based image data versus standard dental radiographs

Shaqayeq Ramezanzade a,1,*, Tudor Laurentiu Dascalu b,1, Bulat Ibragimov b, Azam Bakhshandeh c, Lars Bjørndal c

Abstract

Objectives: The objective was to examine the effect of giving Artificial Intelligence (AI)-based radiographic information versus standard radiographic and clinical information to dental students on their pulp exposure prediction ability.

Methods: 292 preoperative bitewing radiographs from patients previously treated were used. A multi-path neural network was implemented. The first path was a convolutional neural network (CNN) based on ResNet-50 architecture. The second path was a neural network trained on the distance between the pulp and lesion extracted from X-ray segmentations. Both paths merged and were followed by fully connected layers that predicted the probability of pulp exposure. A trial concerning the prediction of pulp exposure based on radiographic input and information on age and pain was conducted, involving 25 dental students. The data displayed was divided into 4 groups (G): G X-ray, G X-ray-clinical data, G X-ray+AI, G X-ray-clinical data+AI.

Results: The results showed that AI surpassed the performance of students in all groups with an F1-score of 0.71 (P < 0.001). The students’ F1-score in G X-ray+AI and G X-ray-clinical data+AI with model prediction (0.61 and 0.61 respectively) was slightly higher than the F1-score in G X-ray and G X-ray-clinical data (0.58 and 0.59 respectively) with a borderline statistical significance of P = 0.054.

Conclusions: Although the AI model had much better performance than all groups, the participants when given AI prediction, benefited only ‘slightly’. AI technology seems promising, but more explainable AI predictions along with a ‘learning curve’ are warranted.

1. Introduction

When intervening for deep caries without severe symptoms, pulp vitality preservation is crucial for successful and cost-effective treatment [1]. In 2019, the European Society of Endodontology (ESE) guideline introduced a radiographic threshold as a predictor of pulp tissue vitality. There have been several treatment options to treat deep caries, including non-selective (previously addressed as complete excavation), selective carious removal and stepwise excavation, in an attempt to avoid pulp exposure. The risk of bacterial pulp penetration is low for deep lesions, while the presence of bacteria in the pulp is apparent in the extremely deep lesions requiring pulp invasive treatments [3]. Recently deep caries scenarios have further been subdivided into three categories [4]: (a) equal to the inner 1/3–1/4 of dentine thickness, (b) more than the inner ¼ but with a visible radiopaque dentine zone before the pulp, (c) while extremely deep lesions expand the entire thickness without any radiopaque zone separating the lesion from the pulp. The lack of clarity about the lesion depth can lead to various challenges. The literature has indicated that a high proportion of...
dentists utilize unnecessarily invasive interventions for deep carious lesions in permanent teeth instead of less invasive treatments and such decision-making for deep lesions relies on dentists’ comprehension of the scientific rationale underlying different excavation strategies [5].

Focus is raised on a more precise preoperative radiographic assessment of penetration depth, along with clinical data, in choosing the proper treatment modality for the removal of carious tissue [1]. Preoperative radiographs are known to contain valuable visual information needed for advanced caries decision-making, including the lesion depth and the presence or absence of a radiopaque zone separating the lesion from the pulp [2]. Some factors might affect the accuracy of radiographic assessments and impair accurate visual assessment by dentists, including technical issues with optimal angulation, making the clinical reality of carious lesions different from the radiographic view. Also, the experience and expertise of the dentist affect the accuracy of visual inspection of carious lesion depth in radiographs [6,7].

Therefore, computerized technological advancements like computer-assisted diagnosis (CAD) systems are needed to improve the diagnostic capability of dentists and extract more radiographic information than the naked eye of experts can detect [8].

Recently, AI has revolutionized CAD systems [9]. Deep learning models for radiographic diagnosis have achieved performance comparable to that of dentists and in some instances, have surpassed their abilities [10–13]. Zheng et al. [12] assessed 844 X-rays of cases with extensive caries to design a CNN model. The model demonstrated greater accuracy than the dentists with 5–10 years of clinical experience (0.82 vs 0.79). A randomized crossover controlled trial was conducted to assess the impact of an AI-based diagnostic-support software for proximal caries detection using young dentists as participants [14]. AI increased the participants’ diagnostic accuracy but at the cost of increasing the invasiveness of the treatment. A few studies have developed a neural network model to predict pulp exposure and compared its performance with dentists [12,15,16]. However, the current literature lacks research on the benefits and drawbacks of an AI-dentist collaboration for predicting the occurrence of pulp exposure when performing deep caries excavation. The challenges of clinical applicability and adoption in practice have remained unaddressed and further research is warranted.

The aims of this paper were to: (1) Develop and employ a standardized automatic solution for predicting pulp exposure after excavation in cases with extensive caries undergoing stepwise or non-selective excavation. (2) Investigate the effect of providing AI-based radiographic information versus standard radiographic information to dental students and assess how this would affect their predictions of pulp exposure following the removal of carious tissue. The null hypothesis was that the model performance is similar to dental students and there is no difference in the diagnostic performance of dental students with and without AI-based information for prediction of pulp exposure in advanced caries.

2. Method

2.1. Protocol registration and ethical considerations

This research follows the Checklist for AI in Dental Research (Table S1) [17]. This project has received ethical approval from the related university (504-0342/22-5000), and the study complies with data protection rules (514-0847/23-3000). The university ethics committee (03-004/03) ethically approved the Data collection for the dataset [18].

2.2. Summary of the trial

In this study, we developed and employed a standardized automatic solution for predicting pulp exposure after excavation in teeth with extensive caries randomly undergoing stepwise or non-selective excavation. The patients with deep caries eligible for excavation treatment were previously treated in several dental centers. Data collected from patients included pretreatment radiographs, age, preoperative pain level, and the type of allocated treatment (stepwise excavation or non-selective excavation) along with the outcome of treatment (pulp exposure/no pulp exposure). After segmentation, labeling and preprocessing of data, an AI model was applied that was able to predict pulp exposure. Then, a website was designed on which pretreatment radiographs were uploaded. Each case was randomly presented in 4 different groups (G). Gx-ray + clinical data showed radiograph and clinical data covering age and pretreatment pain while AI prediction was hidden from the participant. Gx-ray + AI presented radiographs and AI prediction while clinical data (age and pain) were hidden from the participants. Gx-ray + clinical data - AI presented radiograph and clinical data (age and pain) and AI prediction. 4th- and 5th-year dental students were granted website access to analyze data and make predictions about the occurrence of pulp exposure for each case, utilizing a randomly selected data group from the available four groups. The performance was compared in cases with and without AI-model predictions.

2.3. Dataset

This study used a dataset from a previously published randomized clinical trial [18] on treating deep carious lesions with well-defined deep carious lesions. The selected teeth were randomly allocated to either stepwise excavation (n = 143) or non-selective excavation (n = 149). The data collection was multicenter-based including two centers at Denmark and four centers in Sweden. The radiographic data were generated using the analog radiographic technique. The selection criteria included individuals who reported no pain or had mild to moderate preoperative pain (cases with prolonged unbearable pain and/or pain disturbing night sleep; no response to cold and electrical pulp testing; attachment loss > 5 mm were excluded).

The age of patients was 33 years ± 11 years (mean ± standard deviation (SD)). 189 out of 292 cases reported no pain before excavation treatment while 103 cases reported mild to moderate pre-treatment pain. The distribution of cases with pulp exposure and no pulp exposure was skewed, with a higher proportion of treatments resulting in a ‘no pulp exposure’ outcome (224/292) compared to those with pulp exposure (68/292). Out of the 68 teeth with pulp exposure, non-selective excavation was administered in 43 instances. Only 11 of 292 cases had occlusal caries while the majority of teeth had caries affecting both the occlusal and the proximal surface(s) of the tooth. The frequency of tooth type is displayed in the appendix (Table S2). The majority of teeth were premolars and molars and the rate of pulp exposure were 26.05% and 22.2% respectively.

2.4. AI framework development

The original radiographs were downsampled by a factor of two and subsequently zero-padded to achieve a uniform bounding box size of 958 × 873 pixels (height x width). Following this, the pixel intensity values were normalized to the 0–1 range. The radiographs were manually segmented and labeled using the Seg3D software by one dentist (>3 years of clinical experience) and supervised by two experienced clinicians (<25 years of experience) as a prerequisite for AI framework development.

During training, the dataset underwent diverse data augmentation techniques to improve the model’s resilience. These techniques included random horizontal flipping for orientation variations, random affine transformations for translation and scaling within specific ranges, and random perspective changes to simulate different viewing conditions. To generate AI predictions across our entire dataset, we implemented a stratified 10-fold cross-validation method aligned with the outcome of pulp exposure. The training data within each fold was split further,
reserving 10 % of the data for validation purposes. In each fold, the test data was strictly used for inference. All data for a given patient was included in the same split.

The model was trained utilizing the RMSprop optimization algorithm, configured with an initial learning rate of $10^{-4}$ and a weight decay parameter of $10^{-8}$. We employed a learning rate scheduler to improve training efficiency. This scheduler reduces the learning rate when the model’s performance plateaus, with a waiting period of 10 epochs before implementing a reduction. Model selection was performed based on the validation set performance, with the best-performing model chosen within a training span of 50 epochs. The reference test was the ground truth reflecting the actual outcome of the treatment. The final outcomes were classified as ‘pulp exposure’ and ‘no pulp exposure’ [18]. The model was trained on a machine featuring a single Titan X GPU and 4 CPU cores, each accompanied by 1GB of memory. The AI framework represents a multi-path neural network, where the first path analyzes the dental radiograph, while the second path analyses numerical clinical features. The first path was a CNN based on ResNet-50 architecture. This network was pre-trained on the ImageNet database and then fine-tuned using the dental images. The second path was a neural network trained on the distance between the pulp and lesion extracted from X-ray segmentations, treatment type, pain and age of the patient. Both paths merged and were subsequently followed by fully connected layers, which predicted the probability of pulp exposure after the treatment. The code is available as supplementary material on GitHub https://github.com/tudordascalu/pulp-exposure-classification.

2.5. Website design, data randomization and intervention

A secured website was made in which each participant could log in using the assigned password. All the image datasets were uploaded with information about the type and location of the targeted tooth. The students were presented with all 4 groups using computerized randomization. The ‘radiograph’ cases were considered the control group in the platform. Assessment of each radiograph was allowed once and the response time of students to each case was recorded. Fig. 1 shows the website environment with an example of a GX-ray, clinical data + AI case including X-ray, clinical data and AI prediction. In principle, the website environment for all groups is illustrated in the appendix (Fig. S1–3). The Performance of the model was evaluated by comparing it with two senior dentists before running the trial with the dental students. The model demonstrated higher performance to the senior dentists, achieving an F1-score of 0.71 compared to the senior dentists’ F1-score of 0.59.

2.6. Participants

4th- and 5th-year dental students were invited to participate. The students had taken mandatory courses on deep caries pathology and its clinical management. Each student received information for participation in the trial both written and verbally and signed the consent letter. The participants were given a 10 min. oral presentation on different parts of the website, the accuracy of AI model prediction and technically, how to use the platform when their prediction was decided (pulp exposure/ no pulp exposure). The trial was carried out while at least one author was available for any technical questions (Sh. R, T.D). The test was conducted in small groups of 3–5 students. In addition, the students were trained to stop the timing for rest times during the trial.

2.7. Statistical considerations

Data analysis was conducted using Python. Due to the presence of a class imbalance in our dataset, as having more cases without the occurrence of pulp exposure than cases with pulp exposure, the primary metric was F1-score. The accuracy, the area under the Receiver-Operating Characteristics curve (ROCAUC), sensitivity, and specificity were also calculated. The mean and standard deviation were calculated for each evaluation metric. The paired t-test was performed to see the difference between AI performance and dental students’ performance in all groups and to compare dental students’ performance with and without AI guidance at a significance level of $P < 0.05$.

In addition, the paired-sample t-test was performed to compare the time required for determining the outcome (pulp exposure/no pulp exposure) in all 4 groups at a significance level of $P < 0.05$.

3. Results

Twenty-five students agreed to participate in the trial. The metrics
for AI and dental students are presented in Table 1. The AI model metrics were; an F1-score of 0.71, accuracy of 0.78, sensitivity of 0.62, and specificity of 0.83, and AUC of 0.73. The AI model had better performance than students in all four groups of data (P < 0.001). Regarding the performance of students, the F1-scores in $G_{X-ray}$ and $G_{X-ray}$-clinical data-∅ (0.61 and 0.61 respectively) were slightly higher than the F1-score in $G_{X-ray}$ and $G_{X-ray}$-clinical data (0.58 and 0.59 respectively) with a borderline statistical significance of $P = 0.054$. Comparably, the accuracy for $G_{X-ray}$-∅ and $G_{X-ray}$-clinical data-∅ (0.68 and 0.67 respectively), were higher than the accuracy in $G_{X-ray}$ and $G_{X-ray}$-clinical data which were 0.65 and 0.66 respectively.

In $G_{X-ray}$-∅ and $G_{X-ray}$-clinical data-∅, the mean agreement between dental students’ answers and AI prediction was 0.76. In addition, the correlation between participants’ agreement as measured in groups with AI predictions ($G_{X-ray}$-∅ and $G_{X-ray}$-clinical data-∅) and their F1 score was 0.61. Concerning the treatment type, the AI model performed better on the stepwise excavation arm than the non-selective excavation arm (F1-score 0.77 vs. 0.66 respectively). In contrast, the participants’ performance at the non-selective excavation arm was better than the stepwise excavation arm (F1-score 0.61 vs. 0.55) (Table S3).

Table 2 illustrates the participant’s performance, their year of education, the excavation modality and the center at which the individual was treated. While the majority of the participants were 5th-year students (18 of 25), seven of them were 4th-year students. The 5th-year students performed slightly better than the 4th-year students with a F1-score of 0.60 versus 0.58. The average time spent by participants on each case was 12.2 s $\pm$ 6.0 s. (mean $\pm$ SD). There was a significant difference in time spent on correct vs. wrong answers (i.e., predicting the actual outcome as it clinically occurred). The dental students spent on average one more sec on suggesting the wrong outcome ($P < 0.05$). In addition, there was a highly significant difference between the time spent on correct and wrong answers in groups with AI prediction ($G_{X-ray}$-∅ and $G_{X-ray}$-clinical data-∅) with more time spent on the wrong answers compared to the correct ones ($P < 0.001$, Table S4). The performance of the students and AI on patients collected from the reference center was compared with the other centers from the excavation trial. The AI prediction was comparable across all centers, while the students predicted better on the data collected from the reference center vs. the rest of the centers (F1-score of 0.62 vs. F1-score of 0.55).

The area under the ROC curve was measured to see how well the model could distinguish between the ‘pulp exposure’ and ‘no pulp exposure’ classes (Fig. 2). The model’s performance, as assessed by both sensitivity and specificity, surpassed the mean performance of the dental students (Fig. 2). The variations were demonstrated in sensitivity and specificity when dental students did not have access to AI predictions ($G_{X-ray}$ and $G_{X-ray}$-clinical data) compared to when AI predictions were available ($G_{X-ray}$-∅ and $G_{X-ray}$-clinical data-∅) (Fig. 3). Approximately half of the students exhibited improvement in both sensitivity and specificity with the assistance of AI, while a subset of students did not show significant benefits from the AI predictions.

### 4. Discussion

In this study, we developed and employed a standardized automatic solution for predicting pulp exposure after excavation in cases with extensive caries undergoing either stepwise or non-selective excavation. This study investigated the influence of supplying dental students with AI-generated radiographic data, as opposed to conventional radiographs and clinical information, on their proficiency in predicting pulp exposure. It showed that the implemented multi-path neural network outperformed participants in all four data groups, achieving a significantly higher F1-score ($P < 0.001$) in predicting pulp exposure. This suggested a promising potential for AI-driven diagnosis of predicting pulp exposure in deep caries management with regards to the two different excavation techniques. Interestingly, the students’ performance with AI prediction ($G_{X-ray}$-∅, $G_{X-ray}$-clinical data-∅) showed only a marginal improvement compared to their performance without the model predictions ($G_{X-ray}$, $G_{X-ray}$, clinical data), with a p-value of 0.054. Nevertheless, the more agreeable students were to AI predictions, the better their overall performance was. Taken together, the null hypothesis was partially rejected, as there was a significant difference in the diagnostic performance of AI versus dental students but no statistical difference between dental students with and without AI predictions.

Recently, the focus has been given to examine the categories of well-defined deep and extremely deep lesions [2,3]. The sole term ‘deep’ has been covering a huge range of different lesion penetrations, making a potential interpretation of lesion penetration impossible as a variable important for outcome assessment. However, the inaccuracy and erroneous estimation before treating deep caries might result in under- or over-treatment, limited success as well as increased cost of treatment [19]. Our superior findings of AI are similar to a diagnostic accuracy study comparing the performance of another multi-modal CNN of ResNet18-C with experienced dentists for the radiographic diagnosis of advanced caries and pulpitis [12]. A recent paper has assessed the risk of pulp exposure in apical radiograph images using the DenseNet model. Both the AI model and dentists had comparably high performance (AUC of 0.97 and 0.87 respectively) [20]. In contrast to their material, we used Bitewing radiographs only, which could make a difference in terms of interpreting proper penetration depths. In cases of pulpitis, we speculate that their database could have more obvious scenarios, giving the idea that their dataset had fewer challenging teeth for both the model and dentists than our research comprised. In addition, the absence of over-fitting mitigation strategies raises concerns regarding the robustness of the model [20].

A concern with AI models in dentistry is the possibility to improve accuracy at the cost of ‘increasing type I errors and overtreatment’ [14]. To ensure the model funnels higher sensitivity into better, but not necessarily more invasive treatment options, the specificity of the model was checked. The AI model’s specificity outperformed the participants in all four groups (specificity: 0.83). The AI model had fewer false-positive cases of pulp exposure than the participants. This is also consistent with the aforementioned research [12] as the model had significantly higher specificity than dentists did (specificity: 0.86 vs. 0.73). The present model showed the ability to reduce the cases of ‘over-treatment of deep caries’ dramatically. This supports the use of less

### Table 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>Subgroups</th>
<th>Participants</th>
<th>Accuracy</th>
<th>F1-score</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G_{X-ray}$</td>
<td></td>
<td></td>
<td>0.65</td>
<td>0.58</td>
<td>0.53</td>
<td>0.69</td>
<td>-</td>
</tr>
<tr>
<td>$G_{X-ray}$-clinical data</td>
<td></td>
<td></td>
<td>0.66</td>
<td>0.59</td>
<td>0.57</td>
<td>0.68</td>
<td>-</td>
</tr>
<tr>
<td>$G_{X-ray}$-∅</td>
<td></td>
<td></td>
<td>0.68</td>
<td>0.61</td>
<td>0.59</td>
<td>0.70</td>
<td>-</td>
</tr>
<tr>
<td>$G_{X-ray}$-clinical data-∅</td>
<td></td>
<td></td>
<td>0.67</td>
<td>0.61</td>
<td>0.60</td>
<td>0.68</td>
<td>-</td>
</tr>
<tr>
<td>AI model</td>
<td></td>
<td></td>
<td>0.78</td>
<td>0.71</td>
<td>0.62</td>
<td>0.83</td>
<td>0.73</td>
</tr>
</tbody>
</table>

* Center 6 acts as the reference center.

### Table 2

The participant’s performance based on their year of education, treatment type of teeth and the center on which the patients were treated.
The black-box nature (i.e., the opaque and unexplained events that occur within the system) of AI predictions led to some ‘lack of trust’ among participants. In some radiological research, explainable AI (XAI) provides some visualization but without providing detailed information about the underlying causes or implications of the findings [21]. In our case, XAI was limited in its ability to explain the logic behind pulp exposure prediction.

Further assessments are needed to seek a solution for adding explainability to the designed model. In addition, while the majority of cases benefited from AI prediction being available, the presence of AI prediction did not detectably affect their performance (Fig. 3) for a few dental students. This finding is consistent with Mertens et al. [14] showing most but not all dentists have higher performance with AI. In general, it has been shown that dentists’ behavior and provision regarding extensive carious lesions are affected by their experience and familiarity [22].

Future models should identify patterns in user behavior and preference before guiding dentists in order to have the greatest gain from AI.

Based on the ESE guideline of deep caries management, patient signs/symptoms and clinical findings should be included as essential references in the diagnosis process in combination with radiographic caries depth assessment [2]. However, there was no significant difference in accuracy and F1-score of all groups with and without integrating clinical data (age, presence/absence of mild to moderate pretreatment pain). It was also shown that the last year dental students (5th-year)
performed slightly better than 4th-year students. This is understandable as the experience of the dentists affects their accuracy of visual inspection of caries depth in radiographs [6].

In the current research dataset, teeth were excluded if the carious lesion was an obvious exposed case, with the carious lesion reaching the pulp. Thus, the class imbalance, with ‘no pulp exposure’ cases outnumbering ‘pulp exposure’ cases, resulted from deliberate selection by dentists who aimed to choose eligible cases for deep caries excavation therapy aiming to avoid pulp exposure in their treatments. Taken together, we considered a careful assessment of F1-score, sensitivity, and specificity metrics, which are more informative than accuracy in this scenario. Additionally, to ensure that the performance evaluation didn’t favor cases with no pulp exposure as the majority class, we used the macro average F1-score (equivalent to computing the average of the F1-score for each class).

A few studies have assessed the collaboration of dentists and AI models in detecting radiographic features in dental radiographs [14,23]. The current study was the first to assess the collaboration of an AI model with a large group of dental students in predicting the occurrence of pulp exposure when doing different excavation strategies. Also, the ground truth was derived from actual treatment outcomes rather than relying on expert annotations [18]. Another strength was that the used dataset was collected from several dental centers in Denmark and Sweden. The diverse dataset along with data augmentation might have been beneficial regarding the generalizability of the model, although external validation with digital images is warranted in future work to ensure the generalizability to the real-world scenarios.

Notably, the current AI solution does not consider all factors affecting the treatment decisions in real-world dental practice, e.g., providers’ experience and equipment, patients’ expectations, costs, and regulations [14]. Considering such factors along with the successful implementation of XAI strategies, accurate evaluation of learning curves is essential in future research to enable clinicians to confidently utilize AI as a supportive decision-making tool [24].

5. Conclusion

AI holds the potential as a valuable instrument for predicting pulp exposure in vital teeth with deep caries and no severe symptoms. Although the AI model had much better performance than all groups, the participants when given AI prediction, benefited only ‘slightly’. AI technology seems promising but more explainable AI predictions along with a ‘learning curve’ are warranted.

Clinical significance

This study highlights the potential of an AI-based platform in enhancing the decision-making process of dental students when confronted with radiographs and clinical data of advanced carious teeth in dental practice. Although, the real benefit of such an AI platform in clinical practice is still questioned.

CRediT authorship contribution statement

Shaqayeq Ramezanzade: Conceptualization, Methodology, Funding acquisition, Formal analysis, Writing – original draft, Data curation. Tudor Laurentiu Dascalu: Conceptualization, Methodology, Funding acquisition, Formal analysis, Writing – original draft, Data curation, Writing – review & editing. Bulat Ibragimov: Conceptualization, Methodology, Data curation, Writing – review & editing. Azam Bakhsandeh: Conceptualization, Methodology, Data curation, Writing – review & editing. Lars Bjørndal: Conceptualization, Methodology, Data curation, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: This project has been selected for UCPH Data+ funding (Strategy 2023 funds) at the University of Copenhagen, Denmark. Also, the current research was funded by the Danish Endodontic Society. The authors had no other conflicting interests to declare.

Acknowledgments

None.

Supplementary materials


References


