Can a Computer Identify Carious Lesions in Dental X-Rays as Accurately as a Dentist?

An exploratory study comparing diagnostic assessments performed by humans and a specialized computer vision system

AUTHORS
Cambron Carter MEng
Nishita Sant MS
Rohit Annigeri MS
Nandakishore Puttashamachar MS
Kyle Stanley DDS

Abstract

Artificial Intelligence is rapidly revolutionizing the world around us, as driverless cars weave their way through traffic, computer speech becomes indistinguishable from human, and robotic players defeat world champions at complex games like chess and Go. Medical applications of computer vision have been especially surprising, with computer analyses of medical images, such as chest X-rays, equaling or surpassing the sensitivity and accuracy of experienced human clinicians. The powerful triad of computer vision, artificial intelligence and human clinical expertise is creating new opportunities to improve dental care and access. To ascertain whether parity between human and computer diagnostic abilities has really been achieved, a pilot study was conducted comparing the performance of three experienced dental radiograph readers with a computer vision/machine learning (CV/ML) system for identifying caries. The human and digital analysts annotated a sample of more than 10,000 dental X-rays, scoring them for the presence or absence of caries.

The study compared levels of diagnostic agreement and disagreement among the human readers, singly and in combination, and compared those with the computational results from the CV/ML system. The results demostrate the capabilities of the CV/ML system at predicting the existence of caries on the basis of radiographic images than the human readers and give credence to the promise of CV/ML systems capabable of augmenting the work of dentists, both by pre-screening images and identifying suspect areas, and by providing a second opinion that is demonstrably reliable. A CV/ML system, working in tandem with the human practitioner, can improve diagnostic accuracy, reduce costs by enhancing early detection, increase patient confidence in diagnoses, reduce liability exposure, and improve long-term outcomes – a win-win for patients and dentists alike.

Introduction

Artificial intelligence (AI) technologies are playing a growing role in healthcare. In recent years, computer-aided diagnostic (CAD) systems have learned to scan and interpret medical images, such as X-rays. These systems can detect numerous conditions and anomalies across various imaging modalities with accuracy rivaling that of human experts.[1,2,3,4,5,6] The branch of AI that makes computerized imaging diagnostics possible – computer vision (CV) –– has flourished thanks to recent advances in artificial neural network-based machine learning (ML) algorithms. Modeled on biological neural systems, these algorithms allow computers to learn in much the same way that humans do. This learning technology has enabled computers to defeat the best human players at extremely complex games like chess and go, and been the source of unprecedented disruption in many fields, including medicine.[7] In the domain of CV, a class of neural network known as Convolutional Neural Network (CNN) has been the primary force behind improved results in image and video classification, detection, segmentation, and augmentation––results which have made advances such as autonomous vehicle navigation a reality.[8] Like most digital computational systems, CNNs have a great advantage over humans in speed; they can process hundreds of images in the time it takes a human reader to interpret one, and with no loss of accuracy. What is more, CNNs are never subject to fatigue, stress, or environmental distraction. It is for these reasons that CNN-powered diagnostic systems are already helping to bring increased consistency and earlier detection to medical radiology.

Computer vision systems clearly have an application in dentistry, where radiography, intraoral scans, photographs and facial scans provide practitioners with a sometimes overwhelming quantity of unstructured data.[9] Like medical doctors, different dentists may draw different conclusions from radiographs, and these may affect the way a patient is treated.[10] Assisted by CV/ML diagnostic systems, clinicians could better diagnose and document – and therefore treat – their patients. Indeed, a supplemental CV/ML component might serve the same purpose as a second opinion, confirming a diagnosis or calling the clinician's attention to overlooked features.

To assess the performance of a CV/ML system in comparison with one or more humans, the authors completed an exploratory study comparing diagnostic assessments performed by humans and a CV/ML in a controlled environment. Three experienced professional clinicians and a CV/ML system examined a large set of bitewing and periapical images for the presence or absence of caries. What was measured was not the accuracy of their diagnoses but the level of agreement among the four participants. As this study shows, the CV/ML system was as good as a human expert.

Methodology

Three dental clinicians with professional experience ranging from 3 to 10 years in practice were asked to highlight carious lesions in 10,617 radiographs, composed of 4,147 bitewings and 6,440 periapicals. The online annotation tool used, created by Pearl (Pearl Inc., West Hollywood, California, USA), allowed them to identify suspected caries with a rectangular bounding box (Figure (2) and to categorize them as Enamel only, Into dentin, and Approaching or into pulp.

(It should be noted that the images used in this study were not among those used in the training or validation of the CV/ML system.)

Each of the three clinicians reviewed a large subset of the data. Clinicians 1 and 2 reviewed 9,051 images in common; clinicians 2 and 3 8,770, and clinicians 1 and 3 9,638. 8,767 images were reviewed by all three clinicians.

Clinicians	Caries Exist	%
C1	1644	15.9
C2	783	8.6
C3	514	5.3

Table 1: Number of caries-positive images identified by each of the three clinicians.

Clinicians	Common Images Reviewed
C1-C2	9051
C2-C3	8770
C3-C1	9638

Table 2: Number of images which were reviewed by both clinicians in a clinicians-pair.

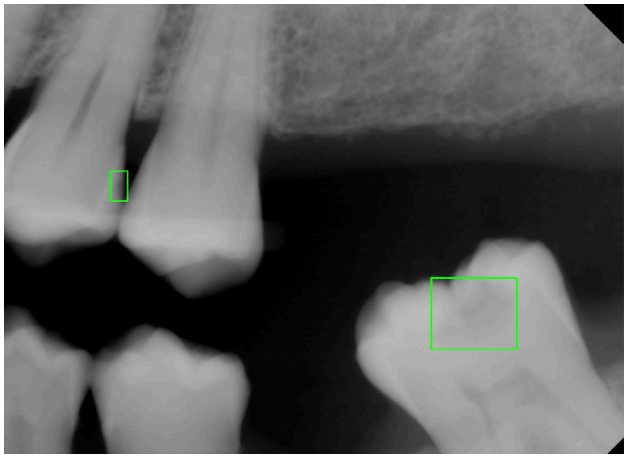

Figure 1: Result of a clinician drawing bounding boxes around carious regions.

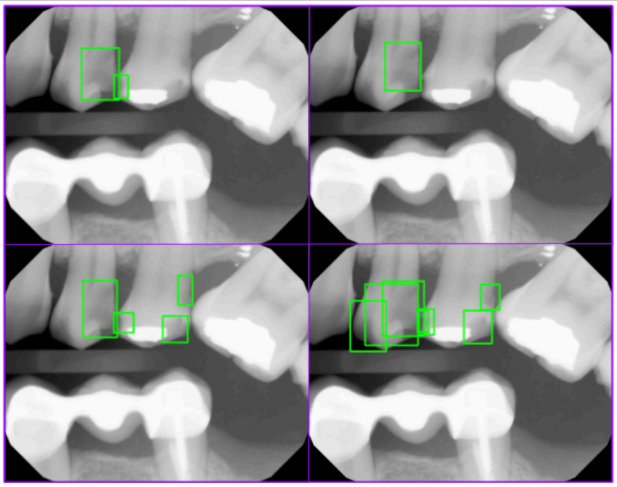

Figure 2: Output of three clinicians and one CV/ML model annotating a radiograph for caries. (Top left) Clinician 1: C1. (Top right) Clinician 2: C2. (Bottom left) Clinician 3: C3. (Bottom right) CV/ML.

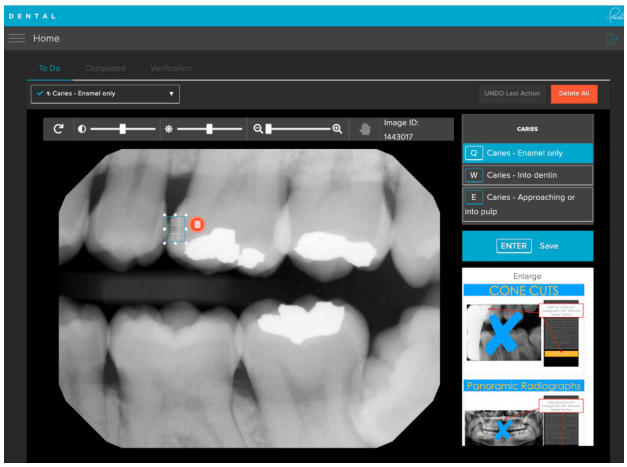

Figure (3). Pearl online annotation tool

Performance metrics of human readers

The table below presents the numbers of images in which pairs of readers agreed that caries existed, along with the percentage of commonly reviewed images that readers agree contain caries.

Clinician	Caries Exist	%
C1-C2	658	7.2
C2-C3	379	4.3
C3-C1	505	5.2

Table 3: Agreement between clinicians on the existence of carious lesions at the image level, i.e. this image does contain at least one carious lesion.

The percentages are consistent with the statistically probable incidence of caries in any large random data set. We estimate the expected frequency of caries-positive bitewings and periapicals to be between 2-12%[11,12]. The relatively large difference between the pair of readers with the highest level of agreement and the pair with the lowest presumably reflects the difficulty of positively identifying a carious lesion in borderline cases.

The next table shows the converse: the level of agreement that no caries are present.

Clinician	Agreement: No Caries Exist	%
C1	7206	79.6
C2	7852	89.5
C3	7657	79.5

Table 4: Agreement between clinicians on the non-existence of carious lesions at the image level, i.e. this image does not contain any carious lesions.

Not surprisingly, pairs of clinicians found it easier to agree when no suspicion of a lesion was present.

The next table shows the levels of disagreement within each pair. No conclusion can be drawn as to the accuracy of their respective judgments, but it is noteworthy that clinician 1 was more than twice as likely to disagree with either of the other clinicians than they were to disagree with each other.

Clinicians	Discrepancies	%
C1-C2	1187	13.1
C2-C3	539	6.2
C3-C1	1476	15.3

Table 5: Disagreement between each clinician-pair, both positive and negative in the case of the presence of carious lesions.

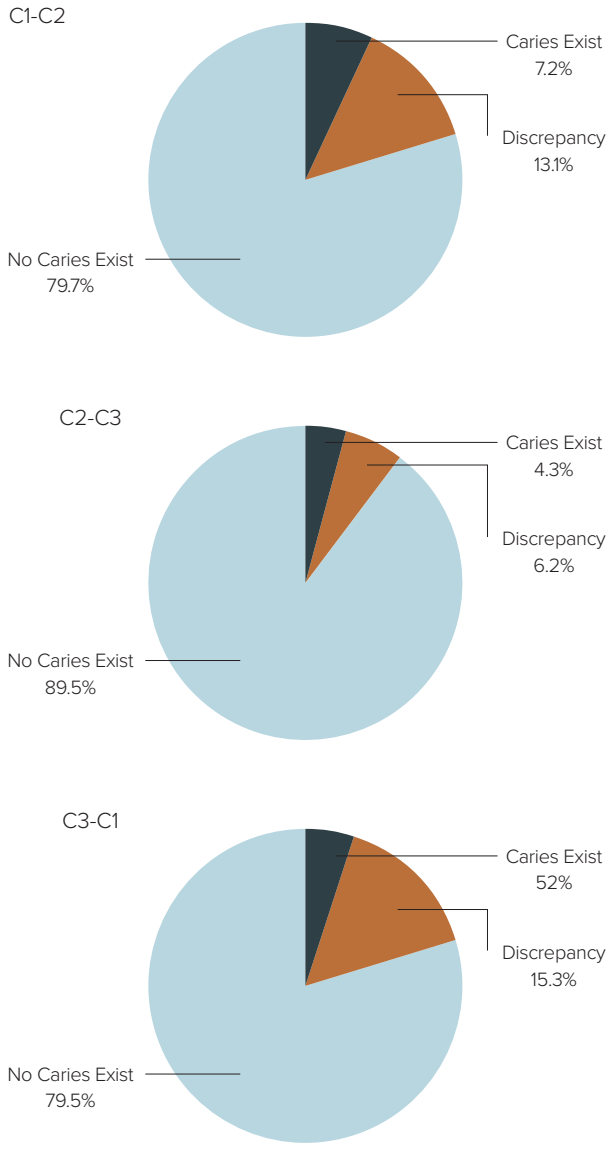

Figure 4: Visualization of inter-clinician agreement and disagreement. (Top) C1 compared with C2. (Middle) C2 compared with C3. (Bottom) C3 compared with C1.

The percentage of agreement and disagreement for *all three* clinicians is displayed in the table below.

Image-level C1-C2-C3	Count	%
Shared Reviews	8767	100
Positives: Unanimous	370	4.2
Negatives: Unanimous	6927	79.0
Positives: At least 1	1840	21.0
Negatives: At least 1	8397	95.8
Positives: At least 1 or 2	1470	16.8
Overall Agreement: Unanimous	7297	83.2

Table 6: Considering all three clinicians, measures of agreement and disagreement at the image level.

Note that in the above table, the categories Discrepancies + Unanimous agreement, Unanimous positives + At least 1 negative, and Unanimous negatives + At least one positive all add up to 100% of the Shared Reviews.

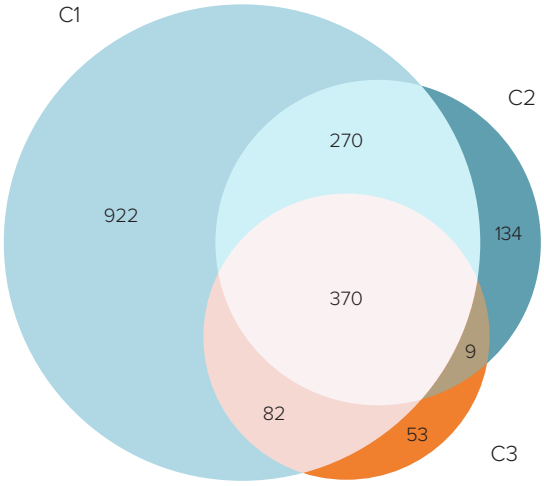

Figure 5: Visualization of agreement and disagreement between multiple annotators. Data is consistent with Table 6. The images not represented are 6927 images (79.0%) which are unanimously labeled as not containing caries.

Performance metrics of CV/ML system

Finally, the performance of a CV/ML system trained to analyze dental radiographs was benchmarked against the human clinicians, singly and in combination, in determining whether or not an image contained a carie. The test consisted of several iterations, in each of which the judgment of one human clinician was assumed to be correct, and the other humans and the computer were scored on their degree of agreement with this assumed ground truth (GT).

The metric used for calculating accuracy is the area under the curve (AUC) for the Receiver Operating Characteristic curve (ROC). ROCs were normalized to the unit square in order to allow comparisons to be made between the absolute yes/no judgments of human readers and the fractional confidence levels generated by the CV/ML system. The results are shown below in Table 7.

	C1	C2	C3	CV/ML
GT-C1	X	0.684	0.627	0.810
GT-C2	0.846	X	0.734	0.850
GT-C3	0.862	0.844	X	0.880

Table 7: Results of three clinicians and one computer vision model tested while holding each of the three humans as ground truth. The rows represent a different clinician being used as ground truth and the columns represent a different clinician and CV/ML being used as predictors.

An observation worth noting is the difference in the comparison of any two readers when one is held as ground truth and the other as the predictor. For example, Clinician 1 predicts with an AUC of 84.6% against Clinician 2 as ground truth. In the inverse situation, Clinician 2 only predicts with an AUC of 68.4%. The lack of symmetry is due to the effect of class imbalance on the ROC paradigm. If the likelihood of positives greatly outweighs the likelihood of negatives the penalty for a Type II error (false negative) will have a lesser impact on the AUC. Considering that bitewing and periapical radiographs naturally occur with a likelihood of being caries-negative, i.e. healthy, the penalty for Type II errors has a higher impact on the AUC. The consequence of this is that clinicians with an affinity for sensitivity will generally outperform those with an affinity for specificity. As can be seen in Figure 5 (venn diagram), C1 > C2 > C3 with respect to sensitivity.

	C1	C2	C3	CV/ML
GT-C1-C2	X	X	0.780	0.906
GT-C2-C3	0.912	X	X	0.928
GT-C3-C1	X	0.884	X	0.902

Table 8: Results when holding reader-pairs (unanimous) as ground truth and measuring the performance of the remaining clinician and computer vision model as predictors.

	CV/ML
GT-C1-C2-C3: Unanimous	0.927
GT-C1-C2-C3: At least 1	0.788

Table 9: Results when treating all three clinicians as ground truth and only computer vision as the predictor. Unanimous means all three clinicians must agree that a carie is present in the image. "At least 1" means that if any one clinician identified the image as caries-positive it is treated as such.

Discussion

As dental technology progresses, diagnosis, treatment planning, and clinical practice can be expected to evolve and improve. One pervasive problem, however, especially from the point of view patients, has been the lack of diagnostic unanimity among clinicians, which has been a source of documentented confusion and distrust among patients.

The results of this study indicated that three experienced dentists disagreed on the existence of caries for approximately 17% of bitewing and periapical radiographs. Two clinicians were more likely to agree, as can be observed by comparing Table 5 to Table 6. When considering all three clinicians together, the number of cases identified as displaying at least one carie is approximately 20%, whereas when considering the intersections of reader pairs the number of caries-positive cases drops to 4%. (These results should be calibrated with the expected frequency of caries-positive radiographs occurring in a random sample.)

The results of the study suggest that the CV/ML tool is superior to clinicians in predicting the existence of caries on the basis of radiographic images. This holds both for situations in which a single clinician is used for ground truth and in comparisons of CV/ML predictions with the intersection between two clinicians' annotations. It is observed that when two clinicians were treated as ground truth, the predictive performance of both CV/ML and humans improves. Both CV/ML and humans were more proficient at identifying caries-negative than caries-positive cases, and were equally sensitive to the distribution of naturally occurring caries in a randomly selected patient pool.

CV/ML performance improved when benchmarked against unanimous ground truth between three clinicians as opposed to the union of each. This is not a surprising result but still a useful one. CV/ML system achieves greater than 90% AUC for caries whose existence is agreed upon by all three clinicians. Further, it is shown that computer vision is proficient, and often superior, at identifying caries-positive radiographs when benchmarked against each of the individual dental clinicians or the combination of two clinicians. The degradation in performance when considering instances where the human clinicians are not unanimous (as compared to when they are unanimous) is, in part, due to the ROC paradigm. The higher level of sensitivity required of the CV/ML system for it to be sensitive to the unique recall of each clinician, irrespective of the disagreement between them, precipitates in a higher false-positive rate. This false-positive rate reduces the overall AUC. Furthermore, the discrepancies which exist between "unanimous" and "at least 1" clinician agreement, also exist in data on which the CV/ML system is trained. Even with rigorous instruction and adjudication, expert annotators still disagree about what constitutes a carie in a radiograph. The CV/ML system is therefore obliged to train on data which contradicts itself some percentage of the time. When there is unanimity in identifying a carie, the visual features that are annotated positively are presumably far more consistent.

This research is limited in that only three dentists were used for ground truth annotation. Future research is needed to evaluate lesion-level results and to expand on the number of dentists used for ground truth annotation.

Conclusion

Takeaways for the industry:

- Typically, practices lose a significant number of their new patients every year and struggle to replace them. Key contributors to this dynamic are the lack of trust in diagnoses that are perceived as expensive, as well as concern among patients that they are being "sold" and not treated. The lack of diagnostic consensus among dentists validates the patient's initial reaction to try another dentist, because they will often receive a different or more favorable diagnosis.
- Because CV/ML is not only accurate but also more consistent, as an assistive aid to denists, it holds the promise of defining a new standard of care that will improve practice economics and increase the overall health of the patient population by increasing trust in the diagnostic process.
- Dental practitioners will limit their liability exposure by incorporating a CV/ML second opinion into their diagnosis and treatment planning sessions.
- The knock-on effects will also improve systemic health by increasing the overall patient population in treatment.

References

1 Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118 (2017).

2 Haenssle, H. A. et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann. Oncol. 29, 1836–1842 (2018).

3 Cheng, J.-Z. et al. Computer aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in CT scans. Sci. Rep. 6, 24454 (2016).

4 Cicero, M. et al. Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. Invest. Radiol. 52, 281–287 (2017).

5 Kooi, T. et al. Large scale deep learning for computer aided detection of mammographic lesions. Med. Image Anal. 35, 303–312 (2017).

6 Barreira, C. M. et al. Abstract WP61: Automated large artery occlusion detection in stroke imaging-paladin study. Stroke 49, AWP61 (2018).

7 Radiographics. 2017 Mar-Apr;37(2):505-515. doi: 10.1148/rg.2017160130. Epub 2017 Feb 17. Machine Learning for Medical Imaging.Erickson BJ1, Korfiatis P1, Akkus Z1, Kline TL1.

8 Front Neurorobot. 2019; 13: 12.Neural Network Based Uncertainty Prediction for Autonomous Vehicle Application Feihu Zhang,1,* Clara Marina Martinez,2 Daniel Clarke,3 Dongpu Cao,4 and Alois Knoll,5

9 Christian Coachman DDS, CDT Marcelo Alexandre Calamita DDS, MSD, PhD Francis Gray Coachman DDS Robert Gray Coachman DDS NewtonSesma DDS, MSD, PhD. Facially generated and cephalometric guided 3D digital design for complete mouth implant rehabilitation: A clinical report. The Journal of Prosthetic Dentistry. Volume 117, Issue 5, May 2017, Pages 577-586.

10 Radiology. 2005 Jan;234(1):274-83. Epub 2004 Nov 10. Pulmonary nodules on multi-detector row CT scans: performance comparison of radiologists and computer-aided detection. Rubin GD1, Lyo JK, Paik DS, Sherbondy AJ, Chow LC, Leung AN, Mindelzun R, Schraedley-Desmond PK, Zinck SE, Naidich DP, Napel S.

11 World Health Organization. Factsheet 355, Noncommunicable Diseases. Updated June 2017. Accessed online July 2017: http://www.who.int/mediacentre/factsheets/fs355/en/

12 World Health Organization. Projections of mortality and causes of death, 2015 and 2030. Online database ‘WHO Regions’ accessed 28 October 2016: http://www.who.int/healthinfo/global_burden_disease/projections/en/