



Application of deep machine learning for the radiographic diagnosis of periodontitis

Jennifer Chang¹ · Ming-Feng Chang^{2,3} · Nikola Angelov¹ · Chih-Yu Hsu² · Hsiu-Wan Meng¹ · Sally Sheng¹ · Aaron Glick⁴ · Kearny Chang¹ · Yun-Ru He² · Yi-Bing Lin^{2,3} · Bing-Yan Wang¹ · Srinivas Ayilavarapu¹

Received: 31 March 2022 / Accepted: 4 July 2022

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2022

Abstract

Objective Successful application of deep machine learning could reduce time-consuming and labor-intensive clinical work of calculating the amount of radiographic bone loss (RBL) in diagnosing and treatment planning for periodontitis. This study aimed to test the accuracy of RBL classification by machine learning.

Materials and methods A total of 236 patients with standardized full mouth radiographs were included. Each tooth from the periapical films was evaluated by three calibrated periodontists for categorization of RBL and radiographic defect morphology. Each image was pre-processed and augmented to ensure proper data balancing without data pollution, then a novel multitasking InceptionV3 model was applied.

Results The model demonstrated an average accuracy of 0.87 ± 0.01 in the categorization of mild ($< 15\%$) or severe ($\geq 15\%$) bone loss with fivefold cross-validation. Sensitivity, specificity, positive predictive, and negative predictive values of the model were 0.86 ± 0.03 , 0.88 ± 0.03 , 0.88 ± 0.03 , and 0.86 ± 0.02 , respectively.

Conclusions Application of deep machine learning for the detection of alveolar bone loss yielded promising results in this study. Additional data would be beneficial to enhance model construction and enable better machine learning performance for clinical implementation.

Clinical relevance Higher accuracy of radiographic bone loss classification by machine learning can be achieved with more clinical data and proper model construction for valuable clinical application.

Keywords Periodontitis · Computer-assisted radiographic image interpretation · Artificial intelligence · Machine learning · Deep learning

Introduction

Periodontitis is one of the most common chronic inflammatory disease affecting half of the adults in the USA [1, 2]. It is initiated by bacterial biofilm infection of the periodontal soft and hard tissues, including the gingiva, the cementum, the periodontal ligament, and the alveolar bone [3]. The host immune response is activated by the infection, leading to inflammation and clinical alteration of the periodontal tissues, increased periodontal pocket probing depth, clinical attachment loss, bleeding on probing, and bone loss around teeth. Without early diagnosis and proper management, progressing alveolar bone loss from severe periodontitis is one of the major causes of tooth loss [4]. Periodontitis is linked with several systemic diseases that can negatively impact patients' quality of lives, such as diabetes [5, 6], cardiovascular diseases [7, 8], respiratory diseases [9], adverse

✉ Jennifer Chang
Jennifer.chang@uth.tmc.edu

¹ Department of Periodontics and Dental Hygiene, The University of Texas Health Science Center at Houston School of Dentistry, Houston, TX, USA

² Institute of Computational Intelligence, National Yangming Chiaotung University, Taipei, Taiwan

³ Department of Computer Science, National Yangming Chiaotung University, Taipei, Taiwan

⁴ Department of General Practice and Dental Public Health, The University of Texas Health Science Center at Houston School of Dentistry, Houston, TX, USA

pregnancy outcomes [10], and cognitive impairments [11], among others. The first step of treating periodontitis is a proper diagnosis. According to the 2017 World Workshop on Classification of Periodontal and Peri-Implant Diseases and Conditions [12], a periodontitis patient should present with at least two teeth with detectable non-adjacent interdental attachment loss or non-interdental attachment loss of ≥ 3 mm with pocketing of > 3 mm. To further classify periodontitis for proper treatment planning, disease severity, complexity, and progression must all be taken into consideration. Classifying periodontitis severity requires identifying the worst interdental clinical attachment loss, evidence of radiographic bone loss (RBL), and amount of tooth loss due to periodontitis. The severity of periodontitis, based on the amount of RBL, are defined as stage I ($< 15\%$), stage II ($15\text{--}33\%$), and stage III or IV ($> 33\%$) of bone loss [13]. Stage III and IV can be further differentiated based on case complexity.

Calculating the percentage of RBL can be time-consuming and labor-intensive since clinicians need to correctly identify the anatomic landmarks, including the cemento-enamel junction (CEJ) to locate the physiologic healthy bone crest level, the base of the bone loss, and the root apex from the radiographs. In normal healthy status, the location of the physiologic bone crest level should be 1–2 mm apical to the CEJ. When there is bone loss, the location of the physiologic bone crest level will be estimated from the radiographs with the identification of the CEJ level [14]. RBL is then calculated according to the following formula [15] (Fig. 1):

The percentage of bone loss = distance α / distance β \times 100

Distance α = physiologic bone crest level to existing bone level

Distance β = physiologic bone crest level to root apex (Fig. 1)

Previous studies pointed out that even after careful calibrations, both intra- and inter-examiner reliabilities on dental radiographic measurements have clinical limitations [16–18].

Deep machine learning is a developing branch of computational algorithms designed to apply artificial intelligence to solve problems by imitating human intelligence and learning from the environments [19]. With computer software and hardware advancements, machine learning has long been introduced outside of the healthcare industry for complex problem solving that were not possible before. Convolutional neural network (CNN) is a subdivision of machine learning that is most applicable for image analyses [20]. It is essentially composed of a set of algorithms resembling the complicated neurons of the human brain. A CNN inputs a fixed size image and processes the input through different layers/neurons until reaching a targeted

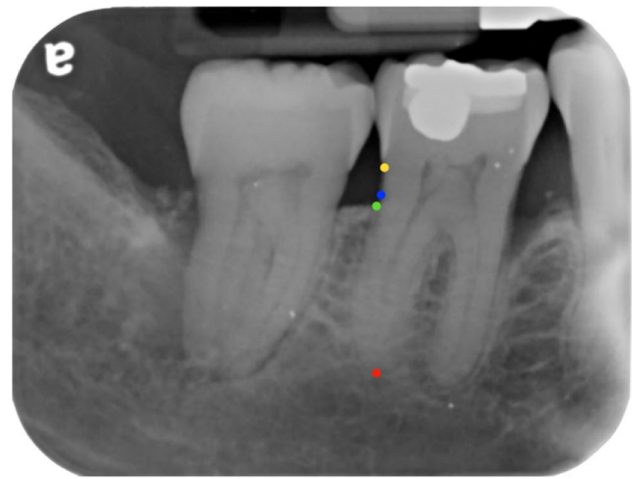


Fig. 1 Illustration of radiographic bone loss. How to calculate the percentage of radiographic bone loss. The percentage of bone loss = physiologic bone crest level to existing bone level/ physiologic bone crest level to root apex \times 100. Color coding of dots: yellow is the location of the CEJ, blue is the physiologic bone crest level, green is the existing bone level, and red is the root apex

output [21]. For any image analysis, the constructed CNN first learns from the training dataset with human-labeled outputs until the model can best produce the objective output on its own. Then, the performance of the CNN will be verified with the testing dataset to check the accuracy of the CNN. Recently, CNN has been applied to medical and dental images [22, 23], including image segmentation, disease detection, and classification. In dentistry, CNN have shown abilities to identify radiographic dental conditions such as dental caries and bone loss [24–26]. Studies applying machine learning on bone loss have been conducted in recent years [25–28]. However, most previous studies were done using panoramic X-rays, which is not the gold standard approach for radiographic diagnosis of periodontitis. Therefore, the objective of this study is to apply deep machine learning for classifying severity of RBL from periodontitis using periapical radiographs.

Materials and methods

Radiographic data collection

The study was approved by the Committee for Protection of Human Subjects of the University of Texas Health Science Center at Houston (HSC-DB-19-0994). All methods performed were in accordance with the relevant guidelines and regulations of the school. Patients that visited the School of Dentistry for treatment signed informed consent forms which explained the use of de-identified patient

data for research. Patients with electronic dental records from the school between 01/01/2010 and 12/31/2020 were screened. Only patients with full mouth standardized radiographs from the school were included in the study. Each standardized radiograph was 503 dots per inch in size, meaning that each pixel of the standardized radiograph was equivalent to 0.05 mm. The standardized peri-apical radiographs were captured by bisecting or parallel techniques using a standardized film holder with the distance between focal spot to position indicating device 9 or 12 inches. The digital images were captured using either CCD or PSP sensor in size #1 or 2. Radiographs that were not in the standard format were excluded from the study.

Three 1-h calibration sessions were carried out to calibrate three board-certified periodontists before the start of the study. Each tooth from the collected periapical films was evaluated by the three calibrated periodontists to calculate the percentage of radiographic bone loss (RBL) and categorized, as healthy (no RBL), stage I (< 15%), stage II (15–33%), and stage III/IV (> 33%) RBL, based on the criteria defined by the 2017 classification of periodontitis [13]. Additionally, radiographic defect morphology was classified as suprabony, intrabony, and severe intrabony (defects with greater than 3 mm of depth), as presented in Fig. 2. Teeth were excluded if the radiograph was lacking diagnostic quality: (1) improper angulation causing severe elongation, foreshortening, or overlapping of anatomic landmarks inhibiting RBL classification; or (2) anatomic landmarks not captured in the image, such as root apex of a tooth. Disagreements of radiographic categorization were resolved by group discussion to reach final consensus,

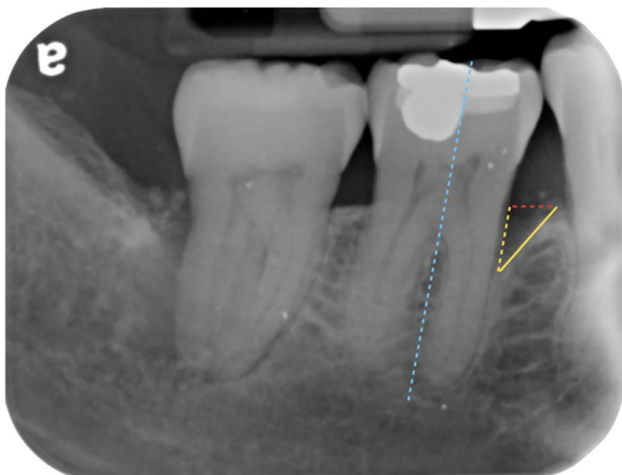


Fig. 2 Intrabony defects were associated with more severe bone loss. Intrabony defect is defined as the configuration of bone loss (solid yellow line) is not perpendicular to the long axis of the tooth (dotted blue line). The depth of the defect is measured from the bone crest level (dotted red line) to the base of bone loss. The current study noticed intrabony defects were associated with more severe bone loss compared to suprabony defects

which served as classification standards for deep machine learning analysis. The study process is illustrated in Fig. 3.

Data grouping for fivefold cross-validation

To ensure validity of the constructed machine learning model, fivefold cross-validation was applied to confirm the performance of the machine learning. In brief, the collected dataset was divided into 5 groups. For each fold, three groups were used as the training datasets (to train the parameters of the CNN), one group as the validation dataset (to select the trained CNN with the highest accuracy), and the remaining last group as the testing dataset (to evaluate the performance of the selected CNN). Such fivefold cross-validation could more objectively evaluate the accuracy of the constructed model [29]. Example of how fivefold cross-validation was incorporated for the project is presented in step (2) of Fig. 4.

Data pollution was avoided by ensuring that the same subject was not used multiple times in different datasets when dividing the collected data into 5 groups. In addition, the prevalence of the severe stage III/IV category RBL is not as common as patients with less severe categories of RBL in the collected radiographs, similar to findings from previous studies [25–28]. Only 10% of the images were in the most severe stage III/IV category with > 33% RBL. To compensate for such imbalance and avoid data pollution during the 5-group division process, the included subjects were randomly divided into five groups to ensure the number of stage III/IV teeth was a similar amount in each group. Similarly, healthy, stage I, and stage II images were randomly divided into these five groups.

Image pre-processing and augmentation

Radiographic images were pre-processed before the application of deep machine learning. First step of image pre-processing was manual segmentation using an open access annotation tool, LabelMe [30]. Manual segmentation was done by drawing a polygon around the entire tooth and its adjacent interdental bone level. The smallest rectangle that could cover the polygon was used to crop each tooth from the radiographs. To focus on the interdental region of bone loss, the rectangles were shifted both left and right by 33% of its width before the images within the rectangles were cropped. Two images for each tooth were created after shifting. Each segmented image was flipped, rotated $\pm 10^\circ$, and contrast-enhanced using contrast limited adaptive histogram equalization (CLAHE) to overcome data imbalance according to methods described by Reza [31] with clipLimit of 2.0 and tileGridSize of 8×8 pixels to obtain proper data augmentation. The CLAHE limited the contrast amplification to reduce issues of noise amplification in near-constant regions

(1) Data collection					
(2) Data grouping, image pre-processing and augmentation before CNNs application					
	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Group 1	Training data	Training data	Training data	Validation data	Testing data
Group 2	Training data	Training data	Validation data	Testing data	Training data
Group 3	Training data	Validation data	Testing data	Training data	Training data
Group 4	Validation data	Testing data	Training data	Training data	Training data
Group 5	Testing data	Training data	Training data	Training data	Validation data

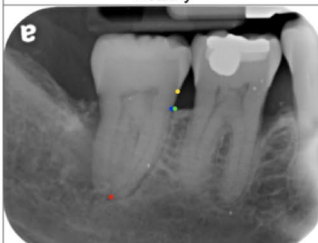
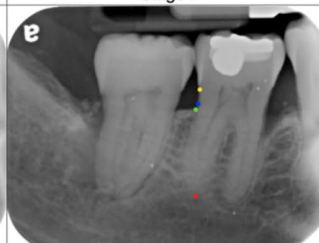
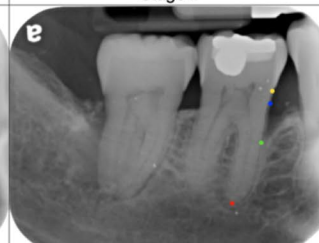

(3) RBL classification outcomes					
Mild			Severe		
Healthy	Stage I		Stage II	Stage III	
					

Fig. 3 Study process. Flowchart of the study process. Step (1) Data Collection: 6,219 proximal surfaces from 1,832 periapical radiographs of 236 patients were included in the study. Step (2) Data grouping, image pre-processing and augmentation before CNNs application: To ensure validity of the constructed machine learning model, 5-fold cross-validation was applied to confirm the performance of the machine learning. Step (3) RBL classification out-

comes: Machine learning classified images into mild (healthy to stage I <15% of bone loss) and severe (Stage II and III \geq 15% of bone loss) RBL, to overcome smaller sample size after data balancing. Color coding of dots: yellow is the location of the CEJ, blue is the physiologic bone crest level, green is the existing bone level, and red is the root apex. Notice in healthy situation, the existing bone level is the physiologic bone crest level

[32]. All images, from healthy to stage III/IV, were augmented and randomly sampled so that the number of images within each category was equal to achieve data balancing.

Application of deep machine learning

InceptionV3 [33] was used to determine the severity of RBL. InceptionV3 was designed to reduce the computation cost of deep CNN using 1×1 convolution, and stacked $1 \times n$ with $n \times 1$ convolutions. Multiple convolutions using different filter sizes were performed in a single layer to reduce the layers of the network. To achieve a more reliable diagnosis, a multitasking InceptionV3 model was developed to determine the defect morphology and RBL severity categories at the same time. All the cropped inter-dental images were in 8-bit grey level. The images were resized to 100×180 pixels before application of the multitasking InceptionV3 model. The multitasking InceptionV3 model, performed using Keras 2.3.0 on a Windows 10 computer and a Nvidia 2080TI graphic card, had a random initialized weights with the last fully connected layer deleted. The deleted layer was replaced by a Global Max Pooling layer, a 1024-node fully connected layer with the rectified linear (ReLU) activation function. The outputs of ReLU were linked to two parallel fully connected layers, one for RBL severity and the other for radiographic

defect morphology classifications. RBL classification was performed by a two-node fully connected layer with the softmax activation function for mild (< 15%) or severe (\geq 15%) RBL. Defect morphology classification was performed by an additional three-node fully connected layer with the softmax activation function for suprabony or intrabony defects. The total number of parameters for this multi-tasking model was 23,906,534, of which 23,872,102 were trainable. To minimize the loss function of the categorical cross entropy, the RMSprop optimizer was used with a learning rate of 0.001 and a batch size of 64. After a series of experiments, the loss function of the multi-tasking model was the weighted sum of the loss from the periodontal bone loss classification (weight 0.7) and the loss from the morphology classification (weight 0.15) that obtained the optimal performance. The overall model is depicted in Fig. 4. The primary outcome of the study was to evaluate the performance of machine learning on RBL, using final consensus results as the classification standards.

Statistical analyses

The confusion matrix, test accuracy, sensitivity, specificity, positive, and negative predictive values of the constructed

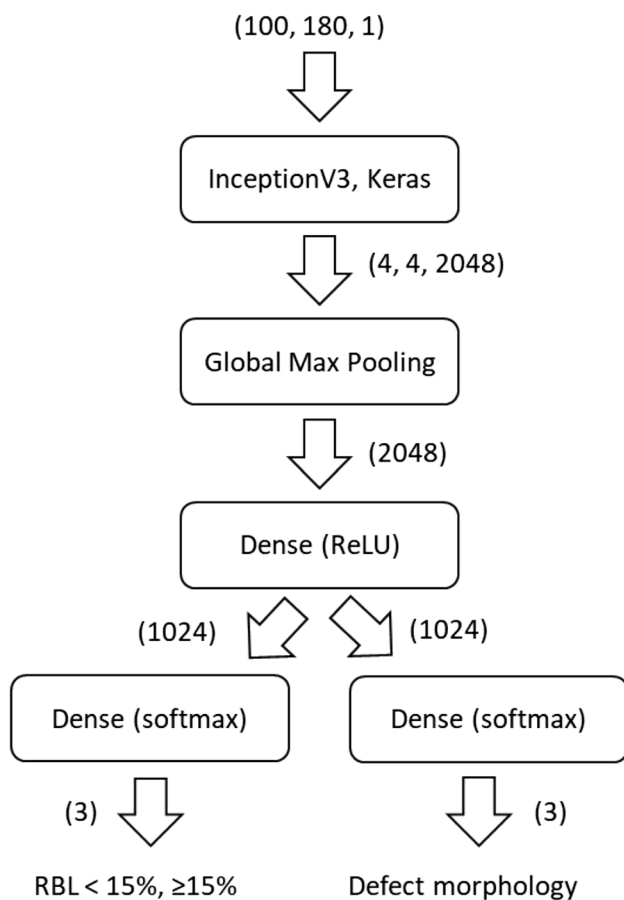


Fig. 4 The multi-tasking CNN model. (100, 180, 1) represents that each input image was 100X180 pixels in grey scale. (4, 4, 2048) represents that each image was 4X4 with 2048 channels. (2048), (1024), and (3) represent the dimension of the data outputs

multitasking InceptionV3 model against the classification standards were assessed. Additionally, the multitasking InceptionV3 model was evaluated against classification standards from periodontists as reference tests in true and false positive rates to analyze its receiver operating characteristic. Chi-square was used to compare differences between fivefolds. Comparisons between mild and severe RBL were made using two-sample *t* tests. A 5% significance level was used for all tests.

Results

Overall, 2362 periapical radiographs of 236 patients were screened for the study. After excluding 22% of non-standardized X-rays or images without proper diagnostic quality (such as anatomic landmarks for RBL classifications were not captured in the image, or improper angulation causing severe elongation, foreshortening, or overlapping), 6219

interproximal surfaces from 1836 X-rays were included and classified by the calibrated periodontists. Of which, 29.4%, 9.5%, 32.8%, and 28.3% were respectively incisors, canines, premolars, and molars. Regarding radiographic bone loss, 17.3% had none, 49.9% had < 15%, 22.8% had 15–33%, and 10% had > 33% of bone loss, while 7.4% of the bone loss configuration were classified as intrabony. Additionally, 0% of healthy, 4.1% of stage I, 14.9% of stage II, and 31.9% of stage III/IV were classified as intrabony defects, indicating that such defect was associated with more severe bone loss. Data augmentation yielded 300,800 images for multitasking InceptionV3 model training and testing. Occurrence frequency of different tooth numbers, RBL, and bone loss configuration classifications are reported in Table 1.

Ability of deep machine learning on categorizing severity of radiographic bone loss

Tables 2 and 3 summarize the results of multitasking InceptionV3 to categorize RBL into either mild or severe groups with fivefold cross-validation. The accuracy of deep machine learning application in different folds ranged from 0.86 to 0.88 (mean 0.87 ± 0.01). Accuracy of fold 1 was 0.87, fold 2 was 0.86, fold 3 was 0.88, fold 4 was 0.87, and fold 5 was 0.86. The mean accuracy of the model in the mild RBL group was 0.88 ± 0.03 and 0.86 ± 0.03 in the severe RBL group. No significant difference in accuracy was found in between two groups ($p = 0.20$). The mean test sensitivity of machine learning on RBL was 0.86 ± 0.03 across fivefolds, mean test specificity was 0.88 ± 0.03 , mean positive predictive value was 0.88 ± 0.03 , and the mean negative predictive value was 0.86 ± 0.02 (Table 3). Figure 5 illustrates the receiver operating characteristic (ROC) curves of the constructed multitasking InceptionV3 model, which the model was evaluated against classification standards as reference tests in true and false positive rates. The area under ROC curve across fivefolds was ranged 0.90–0.94 (mean 0.92 ± 0.02).

Discussion

Based on the 2017 World Workshop on Classification of Periodontal and Peri-Implant Diseases and Conditions, diagnosing severity of periodontitis depends on clinical examination of attachment loss, calculation of number of tooth loss due to the disease, and evidence of radiographic bone loss [12]. The criteria of radiographic bone loss over at least two non-adjacent teeth to diagnose periodontitis are particularly important since clinical attachment level can be biased by clinician skills and other local factors [34]. Previous studies reported even with strong efforts in calibrations

Table 1 Demographic summary of the collected data

Number of subjects	236
Number of radiographs	1832
Number of measurement locations	6219
Tooth type classifications	
Incisors	1826 (29.4%)
Canines	590 (9.5%)
Premolars	2041 (32.8%)
Molars	1762 (28.3%)
RBL classifications	
Healthy	1075 (17.3%)
Stage I (< 15%)	3105 (49.9%)
Stage II (15–33%)	1418 (22.8%)
Stage III/IV (> 33%)	621 (10.0%)
Configuration of bone loss (of the 5144 interproximal areas with bone loss)	
Suprabony	4763 (76.6%)
Intrabony	381 (7.4%)
	Within the 381 bone loss defects, 69 of them had an intrabony component greater than 3 mm

Table 2 Performance of multitasking InceptionV3 model versus agreements between three calibrated board-certified periodontists on categorizing severity of radiographic bone loss. Accuracy of fivefold cross-validation applying deep machine learning multitasking InceptionV3 model to categorize mild or severe periodontal radiographic

bone loss. The accuracy of deep machine learning applications in different folds within the testing data ranged 0.86–0.88 (0.87 ± 0.01). Mean accuracy of the model in the mild RBL group was 0.88 ± 0.03 and 0.86 ± 0.03 in the severe RBL group. No significant difference in accuracy was found in between two groups ($p=0.20$)

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Accuracy (Mean±SD)																																													
Validation data	0.88	0.88	0.87	0.88	0.86	0.87±0.01																																													
Testing data	0.87	0.86	0.88	0.87	0.86	0.87±0.01																																													
Mild group	0.90	0.90	0.90	0.89	0.82	0.88±0.03																																													
Severe group	0.85	0.82	0.87	0.84	0.90	0.86±0.03																																													
Confusion matrix	<table border="1"> <tr> <td>Severe</td> <td>3154</td> <td>17486</td> </tr> <tr> <td>Mild</td> <td>18559</td> <td>2081</td> </tr> <tr> <td></td> <td>Mild</td> <td>Severe</td> </tr> </table>	Severe	3154	17486	Mild	18559	2081		Mild	Severe	<table border="1"> <tr> <td>Severe</td> <td>3649</td> <td>16671</td> </tr> <tr> <td>Mild</td> <td>18231</td> <td>2089</td> </tr> <tr> <td></td> <td>Mild</td> <td>Severe</td> </tr> </table>	Severe	3649	16671	Mild	18231	2089		Mild	Severe	<table border="1"> <tr> <td>Severe</td> <td>2763</td> <td>17877</td> </tr> <tr> <td>Mild</td> <td>18563</td> <td>2077</td> </tr> <tr> <td></td> <td>Mild</td> <td>Severe</td> </tr> </table>	Severe	2763	17877	Mild	18563	2077		Mild	Severe	<table border="1"> <tr> <td>Severe</td> <td>2797</td> <td>15123</td> </tr> <tr> <td>Mild</td> <td>15936</td> <td>1984</td> </tr> <tr> <td></td> <td>Mild</td> <td>Severe</td> </tr> </table>	Severe	2797	15123	Mild	15936	1984		Mild	Severe	<table border="1"> <tr> <td>Severe</td> <td>2055</td> <td>18265</td> </tr> <tr> <td>Mild</td> <td>16715</td> <td>3604</td> </tr> <tr> <td></td> <td>Mild</td> <td>Severe</td> </tr> </table>	Severe	2055	18265	Mild	16715	3604		Mild	Severe	
Severe	3154	17486																																																	
Mild	18559	2081																																																	
	Mild	Severe																																																	
Severe	3649	16671																																																	
Mild	18231	2089																																																	
	Mild	Severe																																																	
Severe	2763	17877																																																	
Mild	18563	2077																																																	
	Mild	Severe																																																	
Severe	2797	15123																																																	
Mild	15936	1984																																																	
	Mild	Severe																																																	
Severe	2055	18265																																																	
Mild	16715	3604																																																	
	Mild	Severe																																																	

Table 3 Test sensitivity, specificity, positive, and negative predictive values of the multitasking InceptionV3 model

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Accuracy (Mean±SD)
Sensitivity	0.85	0.82	0.87	0.84	0.90	0.86±0.03
Specificity	0.90	0.90	0.90	0.89	0.82	0.88±0.03
Positive predictive value	0.89	0.89	0.90	0.88	0.84	0.88±0.03
Negative predictive value	0.86	0.83	0.87	0.85	0.89	0.86±0.02
F1 score	0.87	0.86	0.89	0.86	0.87	0.87±0.01

Sensitivity = true positive / (true positive + false negative)

Specificity = true negative / (true negative + false positive)

Positive predictive value = true positive / (true positive + false positive)

Negative predictive value = true negative / (true negative + false negative)

F1 score = mean of sensitivity and positive predictive value

of experienced clinicians; there would still be disagreements on periodontal RBL between and within examiners [17, 35, 36].

The successful application of CNN on classifying the severity of radiographic bone loss could be developed into a commercially available software to help reducing the clinical workload of precise anatomic landmark identifications and percentage of bone loss calculations for every single patient in daily clinical practice. Additional benefits of the application are that it can (1) avoid the problem of low agreement of novice dentists and serve as a learning aid for dental students in periodontal diagnosis; (2) save time for hygienists and general dentists allowing them to focus on proper clinical

treatments, including serving as a guidance for general dentists to determine periodontal disease severity, complexity, and proper timing of referral to a periodontal specialist; and (3) be useful for periodontists to assess radiographic bone level changes from disease progression, allowing more proactive case management, or treatment outcome, such as radiographic bone fill after surgical regenerative procedures. Several studies have attempted to apply convolutional neural networks (CNN) on similar tasks of classifying RBL binarily [25, 28, 29]. Krois et al. found that the accuracy of CNN was 0.81. The F1 score of DeNTNet by Kim et al. ranged 0.66–0.75. Moran et al. used ResNet and Inception models that had 0.74 and 0.81 accuracy, respectively. Previous studies consistently reported limitations of (1) using panoramic radiographs, which inherit lower accuracy on RBL detection; and (2) the relatively small size of dataset. Similar to previous reports, only 10% of the collected images were in the most severe stage III/IV category with > 33% RBL. With such limitation, the current results primarily focused on classifying between mild and severe bone loss that had the most clinical relevance.

For periodontal purposes, the clinical gold standard is to calculate the amount of bone loss from a periapical radiograph. A periapical radiograph is the most reliable way to capture the entire tooth to the root apex with minimal distortions. A prospective clinical study comparing the accuracy of conventional dental radiographs in assessment of periodontal bone loss concluded that periapical radiographs are more accurate than panoramic films in detecting of osseous destruction, irrespective to the location of the tooth (maxillary, mandibular, anterior, or posterior), and mesial or distal surfaces [37]. Compared to previous studies, one of the strengths of the current study is the inclusion of 6219 interproximal surfaces from 1836 standardized periapical radiographs for analyses. The large amount and accurate nature of standardized periapical radiographs significantly enhance the reliability of the study results. Even though different articles had different designs of the algorithms and

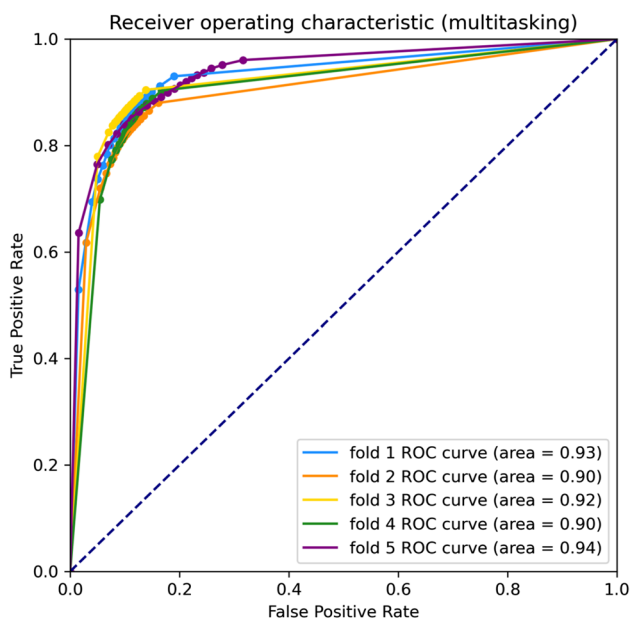


Fig. 5 Receiver operating characteristic (ROC) curves of the multitasking Inception V3 model. The multitasking InceptionV3 model was evaluated against classification standards from periodontists as reference tests in true and false positive rates. The five color coded curves represent five folds of the analyses

clinical data collection, direct comparison may not be feasible. The 0.87 average accuracy of the present study was higher than previously reported, indicating that the current results are more readily available to be translated for clinical application. Future study will attempt to apply machine learning models on non-standardized radiographs from multiple clinical settings.

Besides the larger sample size and more accurate radiographic resource, another possible reason for higher accuracy of the constructed model from our current study was the application of InceptionV3 model [33]. InceptionV3 is a CNN model that instead of stacking the layers, the model orders the layers to operate on the same level to increase efficiency. It has been reported to have superior performance in image classification in the medical field [24]. In addition, the current study found that intrabony defects were highly associated with more severe bone loss during the study process. Therefore, multitasking model factoring in both defect morphology and RBL category were adopted to enhance the accuracy of the present study. Fivefold cross-validation strengthens the results of the present study, indicating that the reported high accuracy did not happen by chance or human manipulation [29].

Although a multitasking model yielded better performance, one of the limitations of the present study was that it focused on radiographic parameters used in the 2017 new classification for diagnosis of periodontitis. Incorporating other important factors determining the periodontal tooth prognosis and treatment plan, including detail information of intrabony defects (angle and walls of the defect), severity of furcation involvements, and presence of endodontic lesions, would be promising in enhancing performance of machine learning. However, the RBL classification standards were determined based on agreements from clinicians diagnosing using two-dimensional images. Such standards cannot avoid inherent limitations of the images and human errors. A more ideal gold standard on severity of bone loss, intrabony defect configuration, and severity of furcation defect might need to be obtained from three-dimensional cone beam computed tomography or surgical entry of the sites.

In conclusion, the current study yielded high accuracy in applying deep machine learning to categorize mild (0.88 ± 0.03) or severe (0.86 ± 0.03) periodontal bone loss without significant difference between mild or severe group ($p = 0.20$). Additionally, there were no statistically significant difference between fivefolds in test accuracy, sensitivity, specificity, positive and negative predictive values, F1 score, and area under receiver operating characteristic curve (all $p > 0.05$), indicating that the results were validated rather than circumstantial occurrence. Future study should focus on data collection for better model construction and training, including larger dataset, more severe RBL cases, and incorporate other important

diagnostic factors (angle and walls of intrabony defects, severity of furcation involvements, and presence of periodontal-endodontic lesions) from more reliable surgical entry or cone beam computed tomography analyses, to enable clinical application of machine learning to assist clinical periodontal diagnosis and treatment planning on a daily basis.

Acknowledgements The authors would like to acknowledge the supports from the University of Texas Health Science Center at Houston School of Dentistry and the Taiwan National Yangming Chiao Tung University.

Author contribution All authors have made substantial contributions to the study. JC, MFC, NA, AG, YBL, BYW, and SA contributed to the design of the study. JC, HWM, SS, and KC have been involved in clinical data collection. MFC, CYH, and YRH dedicated to convolutional neural networks construction and application. JC and MFC worked on data interpretation and manuscript preparation. All authors reviewed the manuscript.

Declarations

Ethical approval The study was approved by the Committee for Protection of Human Subjects of the University of Texas Health Science Center at Houston (HSC-DB-19-0994).

Informed consent Not applicable.

Conflict of interest The authors declare no competing interests.

References

1. Eke PI, Dye BA, Wei L, Thornton-Evans GO, Genco RJ (2012) Prevalence of periodontitis in adults in the United States: 2009 and 2010. *J Dent Res* 91(10):914–920. <https://doi.org/10.1177/0022034512457373>
2. Eke PI et al (2015) Update on prevalence of periodontitis in adults in the United States: NHANES 2009 to 2012. *J Periodontol* 86(5):611–622. <https://doi.org/10.1902/jop.2015.140520>
3. Page R, Schroeder H (1976) "Pathogenesis of inflammatory periodontal disease. A summary of current work." Laboratory investigation. *J Technical Methods Pathol* 34(3):235–249
4. Richards D (2014) Review finds that severe periodontitis affects 11% of the world population. *Evid Based Dent* 15(3):70–71. <https://doi.org/10.1038/sj.ebd.6401037>
5. Borgnakke WS, Ylostalo PV, Taylor GW, Genco RJ (2013) "Effect of periodontal disease on diabetes: systematic review of epidemiologic observational evidence." *J Periodontology* 84(4S):S135–S152. <https://doi.org/10.1902/jop.2013.1340013>
6. Chapple ILC, Genco R, on behalf of working group 2 of the joint E. workshop* (2013) "Diabetes and periodontal diseases: consensus report of the Joint EFP/AAP Workshop on Periodontitis and Systemic Diseases." *J Periodontology* 84(4S):S106–S112. <https://doi.org/10.1902/jop.2013.1340011>
7. Tonetti MS, Van Dyke TE, on behalf of working group 1 of the joint E. workshop* (2013) "Periodontitis and atherosclerotic cardiovascular disease: consensus report of the Joint EFP/

- AAPWorkshop on Periodontitis and Systemic Diseases,." J Periodontology 84(4S):S24–29. <https://doi.org/10.1902/jop.2013.1340019>
8. Schenkein HA, Loos BG (2013) Inflammatory mechanisms linking periodontal diseases to cardiovascular diseases. J Periodontol 84(4S):S51–S69. <https://doi.org/10.1902/jop.2013.134006>
 9. Gomes-Filho IS et al (2020) Periodontitis and respiratory diseases: a systematic review with meta-analysis. Oral Dis 26(2):439–446. <https://doi.org/10.1111/odi.13228>
 10. Sanz M, Kornman K, on behalf of working group 3 of the joint E. workshop (2013) Periodontitis and adverse pregnancy outcomes: consensus report of the Joint EFP/AAP Workshop on Periodontitis and Systemic Diseases. J Periodontol 84(4S):S164–S169. <https://doi.org/10.1902/jop.2013.1340016>
 11. Noble JM, Borrell LN, Papapanou PN, Elkind MSV, Scarmeas N, Wright CB (2009) "Periodontitis is associated with cognitive impairment among older adults: analysis of NHANES-III,." J Neurol Neurosurg Psychiatry 80(11):1206. <https://doi.org/10.1136/jnnp.2009.174029>
 12. Chapple ILC et al (2018) Periodontal health and gingival diseases and conditions on an intact and a reduced periodontium: consensus report of workgroup 1 of the 2017 World Workshop on the Classification of Periodontal and Peri-Implant Diseases and Conditions. J Periodontol 89(S1):S74–S84. <https://doi.org/10.1002/JPER.17-0719>
 13. Tonetti MS, Greenwell H, Kornman KS (2018) Staging and grading of periodontitis: Framework and proposal of a new classification and case definition. J Periodontol 89(S1):S159–S172. <https://doi.org/10.1002/JPER.18-0006>
 14. Hausmann E, Allen K, Clerehugh V (1991) What alveolar crest level on a bite-wing radiograph represents bone loss? J Periodontol 62(9):570–572. <https://doi.org/10.1902/jop.1991.62.9.570>
 15. Schei O, Waerhaug J, Lovdal A, Arno A (1959) Alveolar bone loss as related to oral hygiene and age. J Periodontology 30(1):7–16. <https://doi.org/10.1902/jop.1959.30.1.7>
 16. Gelfand M, Sunderman EJ, Goldman M (1983) Reliability of radiographical interpretations. J Endodontics 9(2):71–75. [https://doi.org/10.1016/S0099-2399\(83\)80079-X](https://doi.org/10.1016/S0099-2399(83)80079-X)
 17. Valachovic RW, Douglass CW, Berkey CS, McNeil BJ, Chauncey HH (1986) Examiner reliability in dental radiography. J Dent Res 65(3):432–436. <https://doi.org/10.1177/00220345860650031201>
 18. Benn DK (1990) A review of the reliability of radiographic measurements in estimating alveolar bone changes. J Clin Periodontol 17(1):14–21. <https://doi.org/10.1111/j.1600-051X.1990.tb01041.x>
 19. Obermeyer Z, Emanuel EJ (2016) Predicting the future - big data, machine learning, and clinical medicine. N Engl J Med 375(13):1216–1219. <https://doi.org/10.1056/NEJMp1606181>
 20. McBee MP et al (2018) Deep learning in radiology. Acad Radiol 25(11):1472–1480. <https://doi.org/10.1016/j.acra.2018.02.018>
 21. Lo S-CB, Chan H-P, Lin J-S, Li H, Freedman MT, Mun SK (1995) Artificial convolution neural network for medical image pattern recognition. Neural Netw 8(7):1201–1214. [https://doi.org/10.1016/0893-6080\(95\)00061-5](https://doi.org/10.1016/0893-6080(95)00061-5)
 22. Y. Yu, (2016) "Machine learning for dental image analysis," ArXiv, abs/1611.09958.
 23. P. Rajpurkar et al., (2017) "CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning," CoRR, abs/1711.05225, [Online]. Available: <http://arxiv.org/abs/1711.05225>
 24. Lee J-H, Kim D-H, Jeong S-N, Choi S-H (2018) Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. J Dent 77:106–111. <https://doi.org/10.1016/j.jdent.2018.07.015>
 25. Krois J et al (2019) "Deep learning for the radiographic detection of periodontal bone loss,." Sci Rep 9(1):8495. <https://doi.org/10.1038/s41598-019-44839-3>
 26. Chang H-J et al (2020) "Deep learning hybrid method to automatically diagnose periodontal bone loss and stage periodontitis,." Scientific Rep 10(1):7531. <https://doi.org/10.1038/s41598-020-64509-z>
 27. Kim J, Lee H-S, Song I-S, Jung K-H (2019) "DeNTNet: Deep Neural Transfer Network for the detection of periodontal bone loss using panoramic dental radiographs,." Scientific Rep 9(1):17615. <https://doi.org/10.1038/s41598-019-53758-2>
 28. MBH. Moran, M. Faria, G. Giralaldi, L. Bastos, B. da S. Inacio, A. Conci, (2020) "On using convolutional neural networks to classify periodontal bone destruction in periapical radiographs,." in 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2036–2039. <https://doi.org/10.1109/BIBM49941.2020.9313501>
 29. Burman P (1989) "A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods,." Biometrika 76(3):503–514. <https://doi.org/10.1093/biomet/76.3.503>
 30. Russell (2008) LabelMe. <https://github.com/wkentaro/labelme/releases/tag/v4.5.7>. Accessed 15 March 2021.
 31. Reza AM (2004) Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement. J VLSI signal Process Syst signal, Image Video Technol 38(1):35–44. <https://doi.org/10.1023/B:VLSI.0000028532.53893.82>
 32. Pizer SM et al (1987) Adaptive histogram equalization and its variations. Computer Vision, Graphics, Image Process 39(3):355–368. [https://doi.org/10.1016/S0734-189X\(87\)80186-X](https://doi.org/10.1016/S0734-189X(87)80186-X)
 33. Szegegy, (2016) "Rethinking the inception architecture for computer vision,." roceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
 34. Caton JG et al (2018) A new classification scheme for periodontal and peri-implant diseases and conditions – Introduction and key changes from the 1999 classification. J Periodontol 89(S1):S1–S8. <https://doi.org/10.1002/JPER.18-0157>
 35. Saunders MB, Gulabivala K, Holt R, Kahan RS (2000) Reliability of radiographic observations recorded on a proforma measured using inter- and intra-observer variation: a preliminary study. Int Endod J 33(3):272–278. <https://doi.org/10.1046/j.1365-2591.1999.00304.x>
 36. McHugh M (2012) Interrater reliability: the kappa statistic. Biochem Med 22:276–281
 37. Pepelassi EA, Diamanti-Kipiotti A (1997) Selection of the most accurate method of conventional radiography for the assessment of periodontal osseous destruction. J Clin Periodontol 24(8):557–567. <https://doi.org/10.1111/j.1600-051X.1997.tb00229.x>

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