Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

PerioAID: An AI-Assisted Quality Assessment and Monitoring of Marginal Bone Levels in Bitewing Radiographs

Maria Garoff
Umeå University
Huei-Min Chiang
Umeå University
Marko Nygård
Umeå University
Jakob Marklund
Umeå University
Fredrik Öhberg
Umeå University
Karolina Karlsson
The Sahlgrenska Academy at University of Gothenburg
Pernilla Lundberg
pernilla.lundberg@umu.se

Umeå University

Research Article

Keywords: artificial intelligence, keypoint detection, radiography, bitewing, periodontitis

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11243988/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

PerioAID: An AI-Assisted Quality Assessment and Monitoring of Marginal Bone Levels in Bitewing Radiographs

Authors: Maria Garoff[1,*], Huei-Min Chiang[2,*], Marko Nygård[3], Jakob Marklund[3], Fredrik Öhberg[3,] Karolina Karlsson[4], Pernilla Lundberg[2]

* Contributed equally

Affiliations

[1] Department of Odontology, Oral and Maxillofacial Radiology, Umeå University, Umeå, Sweden

[2] Department of Odontology, Section for Molecular Periodontology, Umeå University, Umeå Sweden

[3] Department of Diagnostics and Intervention, Biomedical Engineering and Radio Physics, Umeå University, Umeå, Sweden

[4] Department of Periodontology, Institute of Odontology, The Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden

Corresponding Author

Pernilla Lundberg, pernilla.lundberg@umu.se

ORCID: 0000-0002-8244-5200

Abstract

Aim: This study developed and evaluated PerioAID (Periodontal Artificial Intelligence-assisted Diagnostics), a framework for assessing bitewing radiograph quality, measuring marginal bone levels, monitoring radiographic bone loss, and assessing generalisability through external validation.
Materials and Methods: A total of 1000 sensor-based bitewing radiographs from 500 patients were annotated for cemento–enamel junction (CEJ) and marginal bone crests by two dental specialists. A You Only Look Once (YOLO) model was trained, tested, and externally validated on 100 phosphor plate-based radiographs. Performance was evaluated for image quality assessment, tooth detection, keypoint localisation, measurement accuracy, and longitudinal bone-level comparison.
Results: The model detected 95% of teeth and achieved 95% agreement for keypoint visibility. Measurements of CEJ-to-bone distances were comparable to interobserver variability and were maintained in the external validation cohort, demonstrating high diagnostic accuracy and generalisability. The model also identified radiographs unsuitable for bone-level assessment. In a longitudinal analysis, the framework enabled the comparison of marginal bone levels.
Conclusions: The PerioAID model accurately assessed image quality and marginal bone levels. Following further validation, it may serve as a clinical decision-support tool by flagging radiographs unsuitable for reliable bone-level assessment, facilitating early and longitudinal detection of marginal bone loss, and improving workflow efficiency.

Keywords: artificial intelligence, keypoint detection, radiography, bitewing, periodontitis

Introduction

Periodontitis is a prevalent inflammatory disease characterised by progressive destruction of the tooth-supporting tissues, including alveolar bone, and represents a major cause of tooth loss in adults [1, 2]. The clinical diagnosis is based on findings indicative of periodontal inflammation and tissue breakdown, such as bleeding on probing and increased probing depths, whereas radiographic examination is used to assess alveolar bone levels and characterise disease extent and severity.

In routine general dental practice, bitewing radiographs are widely used to assess posterior teeth and marginal bone levels [3]. Radiographic assessment of marginal bone levels is typically based on measuring the distance between the cemento–enamel junction (CEJ) and the alveolar bone crest [3]. Evaluation requires accurate visualisation of the alveolar bone crest, which is not always achieved due to technical limitations during image acquisition and/or advanced attachment loss [4-6]. An evaluation of Swedish public dental services revealed that most bitewing examinations fail to meet diagnostic requirements, with only 15% fulfilling the criteria for adequate depiction of the alveolar bone crest [6]. Consequently, attachment loss may go unnoticed, allowing the disease to progress undetected [2, 7].

Detecting disease progression in radiographs requires the comparison of sequential images obtained with similar projection geometry to avoid misinterpretation, a process that is both time-consuming and dependent on operator expertise [3]. Retrospective evaluations in specialist periodontal care have shown that progressive marginal bone loss is often visible on dental radiographs years before a clinical diagnosis is established in general practice [8, 9]. Because treatment outcomes are substantially more favourable in the early stages of periodontitis than in advanced disease [9], there is a need for objective methods that can improve the identification of early mild bone loss and the monitoring of bone level changes over time.

Advances in artificial intelligence (AI) have shown considerable potential for dental image analysis across a range of diagnostic applications [10-12]. Recent systematic reviews report high accuracy for quantitative measurement in the assessment of periodontal bone loss. However, substantial methodological heterogeneity, dataset variability, and the frequent lack of external validation limit confidence in the robustness and generalisability of existing models, and their clinical applicability remains insufficiently established [13-15]. To address this limitation, the present study included external validation using an independent dataset acquired in a different Swedish region using a different imaging system.

Most studies have focused on panoramic or periapical radiographs, whereas few have evaluated bitewing radiographs despite their superior resolution and limited distortion for the assessment of marginal bone changes [12-16]. Furthermore, AI models generally perform better in detecting severe rather than mild periodontal bone loss in dental radiographs. Minor bone changes may be obscured by anatomical variation and differences in radiographic projection geometry [13-15]. Longitudinal analysis of sequential bitewing

radiographs may help overcome these limitations by identifying cumulative bone changes over time and thereby facilitate earlier detection of periodontal disease.

The aim of this study was to develop and evaluate an AI-based framework, i.e., Periodontal Artificial Intelligence-assisted Diagnostics (PerioAID), for the automated analysis of bitewing radiographs, incorporating: (i) detection of insufficient visualisation of marginal bone; (ii) quantitative measurement of the CEJ–bone distance; and (iii) a pilot evaluation of its performance in monitoring longitudinal marginal bone level changes. The proposed system is intended as a clinical decision-support tool to facilitate the early detection of periodontal disease and support more consistent assessment in general dental practice.

Materials and Methods

Datasets
This retrospective cross-sectional study included anonymised bitewing radiographs from adult patients (≥18 years old) obtained from two Swedish regions: Region Västerbotten and Region Västra Götaland.

A total of 1000 bitewing radiographs from 500 patients treated in 2019 at public general dental practices in Region Västerbotten, Sweden, were included in the main dataset. In addition, 21 radiographs were used for initial observer alignment, and a separate set of 102 radiographs was used for the calibration and assessment of interobserver agreement. The radiographs were derived from a previously established study population, in which patients were randomly selected from patients attending routine dental examinations. The radiographs were randomly sampled from 3000 radiographic examinations and automatically retrieved from the patients' dental records (T4, Carestream Dental, Atlanta, GA, USA; Onepix, Unident, Falkenberg, Sweden). All radiographic examinations had been performed using corded CMOS sensors (Schick 33; Dentsply Sirona, Charlotte, NC, USA). Only images acquired using the largest sensor size (capture area 25.6 × 36 mm; outer dimensions 31.2 × 43.0 × 7.5 mm) were included. The pixel size was 0.03 × 0.03 mm [6].

External validation was performed using 100 bitewing radiographs from public general dental practices in Region Västra Götaland, Sweden. These radiographs were derived from a database comprising intra-oral radiographs collected between 2017 and 2022 from approximately 1200 individuals aged 20–80 years. The external dataset was retrieved from the patients' dental records (T4, Carestream Dental, USA; Planmeca Romexis®, Plandent AB, Stockholm, Sweden). Images captured on a 30 × 40-mm phosphor plate (VistaScan, Dürr Dental AG, Bietigheim-Bissingen, Germany) were included. The pixel size was 0.05 × 0.05 mm. A random sample of 200 bitewing radiographs was initially generated from this database using Excel, from which 100 images from 50 patients were further randomly selected for inclusion in the final external validation dataset.

Images containing dental implants or primary teeth were excluded from all datasets. All included images were exported in jpg format and subsequently imported into a web-based annotation platform (Roboflow, Roboflow Inc., Des Moines, IA, USA; accessed 2026) for

further processing and annotation. Annotation was performed by two observers: a maxillofacial radiologist (MG) and a periodontist (HMC).

Annotation and Calibration

Keypoint detection is suitable for finding and measuring distinct points on an object in an image. This consists of drawing bounding boxes around objects of interest, with a fixed number of keypoints contained in each bounding box.

In our case, bounding boxes and four keypoints (proximal cemento–enamel junctions [CEJs] and marginal bone crests) were annotated using the Keypoint Detection tool in Roboflow. A rectangular bounding box was drawn around each tooth and was defined to include all visible crown and root structures. Within each bounding box, up to four keypoints were annotated: the CEJs (mesial and distal) and the marginal bone crest (mesial and distal). Each keypoint was assigned a visibility flag (i.e., visible or not visible). In the presence of collimation artefacts, bounding boxes were restricted to the visible portions of the teeth. In cases with restorations, the CEJ keypoints were placed at the most proximal apical margin of the restoration and no distinction was made in the presence of secondary caries. No annotation was performed when anatomical landmarks were not visible, including cases of primary caries where the CEJ could not be identified, pronounced overlap between adjacent crowns, and unerupted or partially erupted teeth. An example of an annotated image is shown in **Figure 1**.

Before the full-scale annotation process began, the two observers independently annotated and discussed 21 bitewing radiographs to align their approaches to tooth classification, keypoint placement, and visibility scoring. Each observer then annotated the same calibration subset of 102 images to assess interobserver agreement. After calibration, each observer independently annotated 500 unique images, resulting in 1000 annotated images in total. All 100 images in the external dataset were annotated by a single observer (HMC).

Intersection over Union (IoU)

To assess the agreement between observers in localising teeth, the overlap between their bounding box annotations was quantified. The intersection over union (IoU) is a measure of the overlap between two areas. In the case of rectangular bounding boxes, it can be defined as

$$IoU = \frac{A_I}{A_U}$$

where A_I is the area of the intersection between the boxes and A_U is the area of the union of the boxes.

PerioAID Model Optimisation

In this study, the You Only Look Once (YOLO) architecture for pose estimation was chosen as the basic deep learning framework. Three different YOLO versions were trained and evaluated: YOLOV8, YOLOV11, and YOLO26. The model uses pose estimation to identify objects in images and simultaneously localise predefined points of interest within those objects. Each tooth was considered an object, and the keypoints corresponded to the CEJs and marginal bone levels described in the “Annotation and Calibration” section. As part of model optimisation, a parametric sweep was performed to systematically evaluate

different parameter settings and identify the optimal configuration. The evaluated parameters can be found in **Supplementary Table 1**.

To further optimise model performance, different confidence thresholds (0.3–0.8 in increments of 0.1) were evaluated for bounding boxes and keypoints. For each parameter combination, the median distance between the manually annotated and predicted keypoints (*D*) and the bounding box recall (defined as the fraction of manually assessed bounding boxes accurately detected by the model with an IoU ≥ 0.3) was calculated. The optimal parameter combination was selected by minimising the score function, as follows:

$$S = \left(\frac{median(D)}{recall}\right).$$

The main dataset was split into three subsets: 600 images for training, 200 for validation, and 200 for testing. The pre-processing consisted of resizing the images to 256 × 256 pixels. During training, augmentation was applied to the images in the training subset using the default augmentation settings, including geometric transformations (e.g., scaling, translation, and flipping) and intensity variations. No augmentation was applied to the validation and test sets. The external dataset was pre-processed in the same way, and no augmentations were applied. The manually annotated data provided by the observers were used as ground truth (GT) for training and evaluation.

Task-Specific Analysis

The study aims were approached by identifying images with missing marginal bone (i) and deriving quantitative measurements of CEJ-to-bone distances (i and ii).

Task (i) was evaluated at both the keypoint and image levels. At the keypoint level, the manually assigned visibility flags of the marginal bone keypoints (GT annotation) were compared directly with the corresponding model predictions. At the image level, the analysis was restricted to tooth bounding boxes that, in the GT annotation, had at least one visible keypoint on each of the distal and mesial tooth sides. Tooth bounding boxes that did not meet this criterion were excluded from the image-quality analysis. This was done to avoid classifying teeth that were only partially captured in the radiographs. Each retained tooth bounding box was classified based on the visibility of the CEJ and marginal bone crest keypoints. Tooth bounding boxes were considered sufficient when the marginal bone crest was visible, regardless of CEJ visibility, and insufficient when the marginal bone crest was not visible. An image was classified as of sufficient quality if all retained proximal surfaces were classified as sufficient; otherwise, it was classified as insufficient.

Task (ii) was evaluated by calculating the Euclidean distance between the predicted and GT CEJ and marginal bone keypoints for each proximal tooth surface.

Task (iii) was evaluated by calculating the CEJ–marginal bone crest distances (ii) across multiple images from the same patient acquired at different time points, based on model predictions only.

Statistical Analysis

The model performance and annotation agreement were compared using the same metrics within predefined datasets. Model performance was evaluated against the ground truth using the 200-image test set. Furthermore, interobserver agreement and observer-versus-model comparisons were conducted separately using the 102-image calibration subset. The 21-image practice subset was excluded from all analyses.

For keypoint-level visibility flags, agreement was reported as the percentage and Cohen's kappa (κ). For the GT-filtered image-level quality classification in task (i), results were summarised using the confusion matrix, percentage agreement, κ, and mean number of proximal surfaces classified as of insufficient quality per image in GT and model predictions. Bounding-box agreement was summarised as the median intersection over union (IoU). Keypoint localisation and CEJ–bone distance differences were calculated as Euclidean distances and expressed in mm after conversion using a factor of 0.03 mm/pixel in the main dataset and 0.05 mm/pixel in the external dataset. The CEJ–marginal bone distance agreement was also visualised using cumulative threshold curves and Bland–Altman analysis.

Results

Interobserver Agreement

Interobserver agreement was high across the calibration subset. The median bounding box overlap was 0.94 and keypoint visibility agreement was 95.2% (κ 0.88). For keypoints flagged as visible by both observers, the median interobserver Euclidean difference was 0.20 mm, with 83.7% of keypoints within 0.5 mm. The median Euclidean CEJ–marginal bone distance difference was 0.24 mm, with 83.9% of measurements within 0.6 mm. These results indicate a high level of consistency between observers and provide a robust reference standard for evaluating model performance, within which the model demonstrated slightly lower but comparable agreement. Comparison between observers and model predictions revealed only minimal differences. A comparison of interobserver agreement and observer–model performance on the calibration subset is provided in **Supplementary Table 2**. An example image showing GT and model predictions is shown in **Figure 2**.

PerioAID Model Performance

The model correctly identified 1656 out of 1744 teeth (IoU ≥ 0.3) in the test set. The median bounding-box IoU was 0.95. The keypoint-level visibility agreement was 94.5% (κ 0.85), indicating excellent agreement with the GT (**Table 1**).

Using the GT-filtered image-level quality algorithm, 199 of 200 test images were found to be evaluable. One image was excluded because the tooth bounding boxes failed to meet the inclusion criterion of including at least one visible keypoint at each of both proximal surfaces. GT classified 93 images as of sufficient quality and 106 as of insufficient quality, whereas the model classified 68 and 131 images in these ways, respectively (**Figure 3**). Image-level agreement was 78.4% (κ 0.56), with a sensitivity of 91.5%, a specificity of 63.4%, a positive predictive value of 74.0%, and a negative predictive value of 86.8% for

detecting insufficient image quality. The mean number of proximal surfaces classified as of insufficient quality per image was 0.98 for GT and 1.42 for the model (**Table 2**).

For visible keypoints, the median Euclidean distance was 0.21 mm, with 89.7% within 0.5 mm, indicating high localisation precision. For CEJ–bone distance measurements, the median Euclidean distance difference was 0.25 mm, with 86.0% of measurements within 0.6 mm, reflecting good agreement for clinically relevant measurements. Further keypoint-level and CEJ–bone distance details are provided in **Table 1**.

The cumulative distribution of absolute CEJ–bone distance differences for model versus GT on the test set recalled the interobserver curve (**Figure 4a**). Bland–Altman analysis of model versus GT CEJ–marginal bone distances showed minimal systematic bias (mean difference –0.01 mm) and 95% limits of agreement ranging from –0.89 to 0.87 mm (**Figure 4b**).

External validation showed slightly reduced but comparable performance, with a detection recall of 93.8%, a median IoU of 0.94, keypoint visibility agreement of 92.4% (κ = 0.79), and a median CEJ–bone distance difference of 0.33 mm (**Table 1**). Performance reduction was primarily observed for distance-based measurements, while detection and visibility agreement remained high.

As part of the pilot longitudinal analysis, the model sufficiently detected the CEJ and marginal bone level, enabling measurement of CEJ–bone crest distances in bitewing radiographs from the same individual obtained at different time points. The visualisations confirm consistent keypoint placement across time points and demonstrate the model's ability to track longitudinal changes in bone levels (**Figures 5a and 5b**). Overall, CEJ–bone distances increased progressively over time. In cases with double bone contours, the model consistently identified the most coronal bone level, as illustrated in **Figure 5b** (2015–2022). Minor differences in measured values (at the decimal level) were observed (**Figures 5a and 5b**), including apparent changes in marginal attachment, such as both loss and gain, reflect variations in radiographic projection between time points.

Discussion

The PerioAID framework demonstrated expert-level agreement for the localisation of periodontal landmarks and assessment of CEJ–bone distances, maintained performance following external validation, and showed potential for automated image-quality assessment and the longitudinal monitoring of marginal bone levels.

The high interobserver agreement obtained between the annotators established a robust reference standard for model development and evaluation. Agreement for visibility assessment was high (κ = 0.88), and differences in CEJ–bone distance measurements were small and within the range previously reported for trained clinicians [17]. Model-derived measurements showed only slightly greater deviations and remained within a similar submillimetre range. These findings suggest that the framework performs at a level comparable to expert assessment while offering the advantages of standardised and reproducible measurements.

Despite growing interest in AI-assisted periodontal diagnostics, studies based on bitewing radiographs remain scarce [12-15]. Existing investigations have largely focused on classifying bone loss severity and have seldom included external validation, quantitative measurement of marginal bone levels, or evaluation of longitudinal changes. Lee et al. demonstrated the successful localisation of CEJ and alveolar bone landmarks but focused primarily on the binary classification of severe versus non-severe periodontal bone loss and did not include external validation [18]. AlGhaihab et al. evaluated a commercially available AI system on a small independent test set comprising only 17 bitewing radiographs, but the model development process was not fully described and formal external validation was not reported [19]. Erturk et al. [20] classified bitewing radiographs according to periodontal bone loss severity, whereas the present framework provided quantitative CEJ–bone measurements at the individual tooth-surface level, enabling the detection of subtle changes and longitudinal comparison over time. More recently, Cassiano et al. applied a YOLO-based pose-estimation model for the localisation of CEJ and alveolar bone landmarks; however, evaluation was limited to technical detection metrics and did not include comparison with expert-derived measurements [21]. In contrast, the present study incorporated expert reference measurements, interobserver analyses, external validation using a separate cohort and imaging system, image-quality assessment, and pilot longitudinal monitoring. Nevertheless, collectively, these studies support the feasibility of AI-based assessment of periodontal bone levels using routinely acquired bitewing radiographs.

An important feature of PerioAID was its ability to identify and flag radiographs in which marginal bone levels were inadequately visualised. This is clinically relevant in periodontal diagnostics because suboptimal bitewing image quality is common and may compromise interpretation of the alveolar bone crest [6]. Although agreement with the GT was moderate, the model showed high sensitivity for detecting insufficient visualisation. The tendency to overestimate inadequate visualisation resulted in some false-positive classifications (in 17% of the images). This can be attributed to the GT-filtering algorithm, which is asymmetrical in that it punishes the model by removing potentially valid model predictions and hides any GT false negatives. However, it was a necessary step, since nearly all of the images had at least one partially visible tooth. In a clinical decision-support setting, where overlooking diagnostically limited images could have more serious consequences, this may however be preferable.

The study population largely reflected patients encountered in general dental practice, with most proximal surfaces exhibiting normal or mildly reduced marginal bone support. This distribution is relevant because early radiographic changes are often the most difficult to recognise clinically. However, the limited representation of advanced periodontal destruction restricts conclusions regarding model performance in more severe disease. Further validation in populations with a broader range of bone loss is therefore warranted.

The ability to compare radiographs longitudinally represents another potentially valuable application. Although the diagnosis and monitoring of periodontitis primarily rely on clinical parameters such as probing depths, attachment level measurements, and bleeding on probing, serial radiographs provide important information regarding marginal bone loss, a

defining feature of periodontitis and a key criterion distinguishing it from gingivitis [2]. The pilot framework enabled the systematic comparison of bone levels across examinations and may facilitate the identification of progressive radiographic bone loss, a task that is often time-consuming and inconsistently performed in routine practice. As this functionality was evaluated in a limited number of cases, further validation is required before conclusions regarding its clinical applicability can be drawn.

The longitudinal analyses also highlighted several limitations. In regions with complex bone morphology, particularly where double bone contours were present, the model preferentially identified the most coronal bone level and therefore underestimated deeper defects. In addition, differences in radiographic projection between examinations influenced the apparent location of anatomical landmarks and may have affected measurements of small changes over time. These findings emphasise that AI-generated measurements should be interpreted within their radiographic and clinical contexts. Furthermore, measurements were derived from pixel distances without correction for magnification and should therefore be regarded as relative rather than calibrated absolute values.

Several strengths of this study should be highlighted. The PerioAID framework was developed using a large dataset of 1000 bitewing radiographs with expert annotations provided by two calibrated experts: a periodontist and a maxillofacial radiologist. External validation demonstrated maintained performance despite differences in image acquisition technology, supporting generalisability across imaging systems. Although the external validation dataset was annotated by a single observer, the minimal observed differences between observers and model predictions suggest that this was unlikely to have substantially influenced the results. Furthermore, the exclusive use of bitewing radiographs enhances clinical relevance, as this modality is routinely used for assessing marginal bone levels in general dental practice. To our knowledge, this is the first study to combine the automated assessment of marginal bone levels on bitewing radiographs with external validation and image-quality assessment.

Future research should focus on prospective validation in general practice workflows, integrating radiographic findings with periodontal clinical parameters, and evaluating the framework's impact on diagnostic decision-making. Further methodological refinement is also needed to improve performance in anatomically complex regions and to strengthen robustness under varying image-quality conditions.

Conclusions

The PerioAID framework demonstrated high agreement with expert annotations for the assessment of marginal bone levels on bitewing radiographs and was able to flag radiographs unsuitable for reliable bone-level assessment. The framework also showed potential for the longitudinal monitoring of radiographic bone levels and may serve as a clinical decision-support tool by supporting quality assurance, facilitating the early detection of marginal bone loss, and improving consistency in periodontal assessment. However, clinical parameters remain essential for the diagnosis of periodontitis, and further validation is required before routine clinical implementation.

Declarations

Abbreviations

The following abbreviations are used throughout the manuscript:
AI: Artificial Intelligence
CEJ: Cemento-Enamel Junction
CMOS: Complementary Metal-Oxide Semiconductor
GT: Ground Truth
IQR: Interquartile Range
IoU: Intersection over Union
MBL: Marginal Bone Level
Obs1: Observer 1
Obs2: Observer 2
YOLO: You Only Look Once
κ: Cohen's Kappa

Ethics approval and consent to participate

This study was conducted in accordance with the ethical principles of the Declaration of Helsinki and all applicable national regulations governing research involving human participants. The study was approved by the Swedish Ethical Review Authority (approval numbers 2021-03456 and 2023-01750-02). All radiographic data were anonymised before analysis. The requirement for informed consent was waived by the Swedish Ethical Review Authority due to the retrospective study design.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analysed during the current study are not publicly available due to ethical and privacy considerations related to human radiographic data but are available from the corresponding author on reasonable request

Competing Interests

The authors declare that they have no competing interests

Funding

This study was funded by grants from the County Council of Västerbotten under grant numbers RV-991220 and RV-1011103 (PLU). The funding body did not influence the study design, data collection, analysis, interpretation, or manuscript writing.

Authors' contributions

PL conceived the study, generated the initial idea, and contributed to study design, data analysis, and manuscript drafting.
MG contributed to study conception and design, data acquisition, data analysis, and manuscript drafting.
HMC contributed to study conception and design, data acquisition, data analysis, and manuscript drafting.
MN contributed to study conception, data acquisition, data analysis, and manuscript drafting.
JM contributed to data acquisition, data analysis, and manuscript drafting.
FÖ contributed to study conception and design, and reviewed the manuscript.
KK contributed the external data set and description, and contributed to manuscript drafting and reviewing.

Acknowledgements

Not applicable

References

1. Pihlstrom, B.L., B.S. Michalowicz, and N.W. Johnson, *Periodontal diseases*. Lancet, 2005. **366**(9499): p. 1809-20.
2. Tonetti, M.S., H. Greenwell, and K.S. Kornman, *Staging and grading of periodontitis: Framework and proposal of a new classification and case definition*. J Periodontol, 2018. **89 Suppl 1**: p. S159-S172.
3. Hausmann, E., *Radiographic and digital imaging in periodontal practice*. J Periodontol, 2000. **71**(3): p. 497-503.
4. Mol, A., *Imaging methods in periodontology*. Periodontol 2000, 2004. **34**: p. 34-48.
5. Bragger, U., *Radiographic parameters: biological significance and clinical use*. Periodontol 2000, 2005. **39**: p. 73-90.
6. Olsson, D., E. Levring Jaghagen, and M. Garoff, *Operator-related low diagnostic quality in bitewing examinations performed with sensors*. Acta Odontol Scand, 2026. **85**: p. 245-253.
7. Eke, P.I., et al., *Update on Prevalence of Periodontitis in Adults in the United States: NHANES 2009 to 2012*. J Periodontol, 2015. **86**(5): p. 611-22.
8. Thorbert-Mros, S., B. Cassel, and T. Berglundh, *Age of onset of disease in subjects with severe periodontitis: A 9- to 34-year retrospective study*. J Clin Periodontol, 2017. **44**(8): p. 778-783.
9. Tonetti, M.S., et al., *Impact of the global burden of periodontal diseases on health, nutrition and wellbeing of mankind: A call for global action*. J Clin Periodontol, 2017. **44**(5): p. 456-462.
10. Schwendicke, F., W. Samek, and J. Krois, *Artificial Intelligence in Dentistry: Chances and Challenges*. J Dent Res, 2020. **99**(7): p. 769-774.
11. Lee, J.H., et al., *Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm*. J Dent, 2018. **77**: p. 106-111.
12. Chatzopoulos, G.S., et al., *Clinical Applications of Artificial Intelligence in Periodontology: A Scoping Review*. Medicina (Kaunas), 2025. **61**(6).
13. Chiang, H.M., et al., *How accurate is AI in detecting marginal jaw bone loss? A systematic review and meta-analysis*. J Dent, 2025. **163**: p. 106151.
14. Khubrani, Y.H., et al., *Detection of periodontal bone loss and periodontitis from 2D dental radiographs via machine learning and deep learning: systematic review employing APPRAISE-AI and meta-analysis*. Dentomaxillofac Radiol, 2025. **54**(2): p. 89-108.

15. Iacob, A.M., et al., *Automated Detection of Periodontal Bone Loss in Two-Dimensional (2D) Radiographs Using Artificial Intelligence: A Systematic Review*. Dent J (Basel), 2025. **13**(9).
16. Hellen-Halme, K., A. Lith, and X.Q. Shi, *Reliability of marginal bone level measurements on digital panoramic and digital intraoral radiographs*. Oral Radiol, 2020. **36**(2): p. 135-140.
17. SBU, *Chronic Periodontitis: Prevention, Diagnosis and Treatment*. 2004, Swedish Council on Technology Assessment in Health Care.
18. Lee, S.W., et al., *Evaluation by dental professionals of an artificial intelligence-based application to measure alveolar bone loss*. BMC Oral Health, 2025. **25**(1): p. 329.
19. AlGhaihab, A., et al., *Automatic Detection of Radiographic Alveolar Bone Loss in Bitewing and Periapical Intraoral Radiographs Using Deep Learning Technology: A Preliminary Evaluation*. Diagnostics (Basel), 2025. **15**(5).
20. Erturk, M., M.U. Ozic, and M. Tassoker, *Deep Convolutional Neural Network for Automated Staging of Periodontal Bone Loss Severity on Bite-wing Radiographs: An Eigen-CAM Explainability Mapping Approach*. J Imaging Inform Med, 2025. **38**(1): p. 556-575.
21. Cassiano, L.B.A., et al., *Evaluation of an artificial intelligence-based model in diagnosing periodontal radiographic bone loss*. Clin Oral Investig, 2025. **29**(4): p. 195.

Tables

Table 1. Model performance on the 200-image test set as well as the 100-image external dataset versus ground truth (GT).

Measure	GT (test set) vs. model	GT (external) vs. Model
Images	200	100
Teeth	1744	854
IoU* ≥ 0.3	95.0%	93.8%
Median IoU	0.95	0.94
Keypoint visibility agreement	94.5%	92.4%
Keypoint visibility kappa	0.85	0.79
Visible keypoint pairs (*n*)**	4774	2310
Keypoint median (mm)	0.21	0.28
Keypoint IQR (mm) ***	0.20	0.28
Keypoints within 0.3 mm	70.6%	53.9%
Keypoints within 0.5 mm	89.7%	79.3%
CEJ-MBL measurements (*n*)	2238	1092
CEJ-MBL median (mm)	0.25	0.33
CEJ-MBL IQR (mm)	0.32	0.41
CEJ-MBL within 0.3 mm	57.0%	45.2%
CEJ-MBL within 0.6 mm	86.0%	77.1%

* Intersection over union

** Cemento–enamel junction (CEJ) and marginal bone level (MBL) keypoints present on the same proximal tooth surface*** Interquartile range

Table 2. Ground truth (GT)-filtered image-level quality classification on the 200-image test set versus model performance.

Measure	GT vs. Model
Evaluable images	199
Eligible teeth	1067
Image-level agreement (accuracy)	78.4%
Image-level kappa	0.56
GT insufficient-quality proximal surfaces per image, mean	0.98
Model insufficient-quality proximal surfaces per image, mean	1.42
Sensitivity (recall)	91.5%
Specificity	63.4%
Positive predictive value (precision)	74.0%
Negative predictive value	86.8%
F1 score	81.8%

Figures

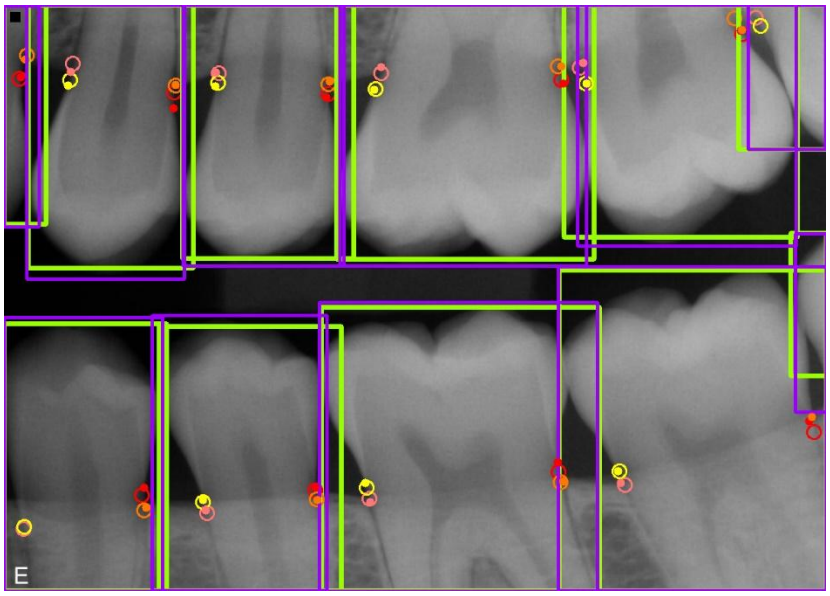


Figure 1. Example of a bitewing radiograph showing the combined annotations of two independent observers. Each tooth is enclosed within a bounding box. Green boxes represent observer 1 and purple boxes observer 2. Four keypoints are annotated per tooth when visible: the cemento–enamel junctions (CEJs) and the marginal bone crest at the mesial and distal tooth surfaces. Mesial CEJs and bone crests are shown in yellow and pink, whereas distal CEJs and bone crests are shown in red and orange. Circles represent annotations by observer 1 and dots represent annotations by observer 2.

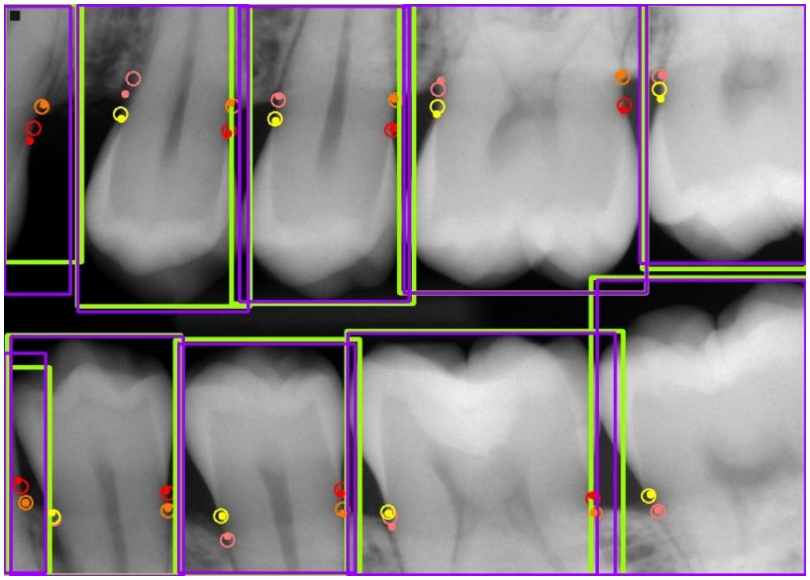


Figure 2. Example of a bitewing radiograph showing ground truth (GT) and model predictions. Each tooth is enclosed within a bounding box. Green boxes represent GT and purple boxes represent model predictions. Four keypoints are annotated per tooth when visible: the cemento–enamel junctions (CEJs) and the marginal bone crests at the mesial and distal surfaces. Distal CEJs and bone crests are indicated in red and orange, and mesial CEJs and bone crests are shown in yellow and pink. Circles represent GT annotations and dots represent model predictions.

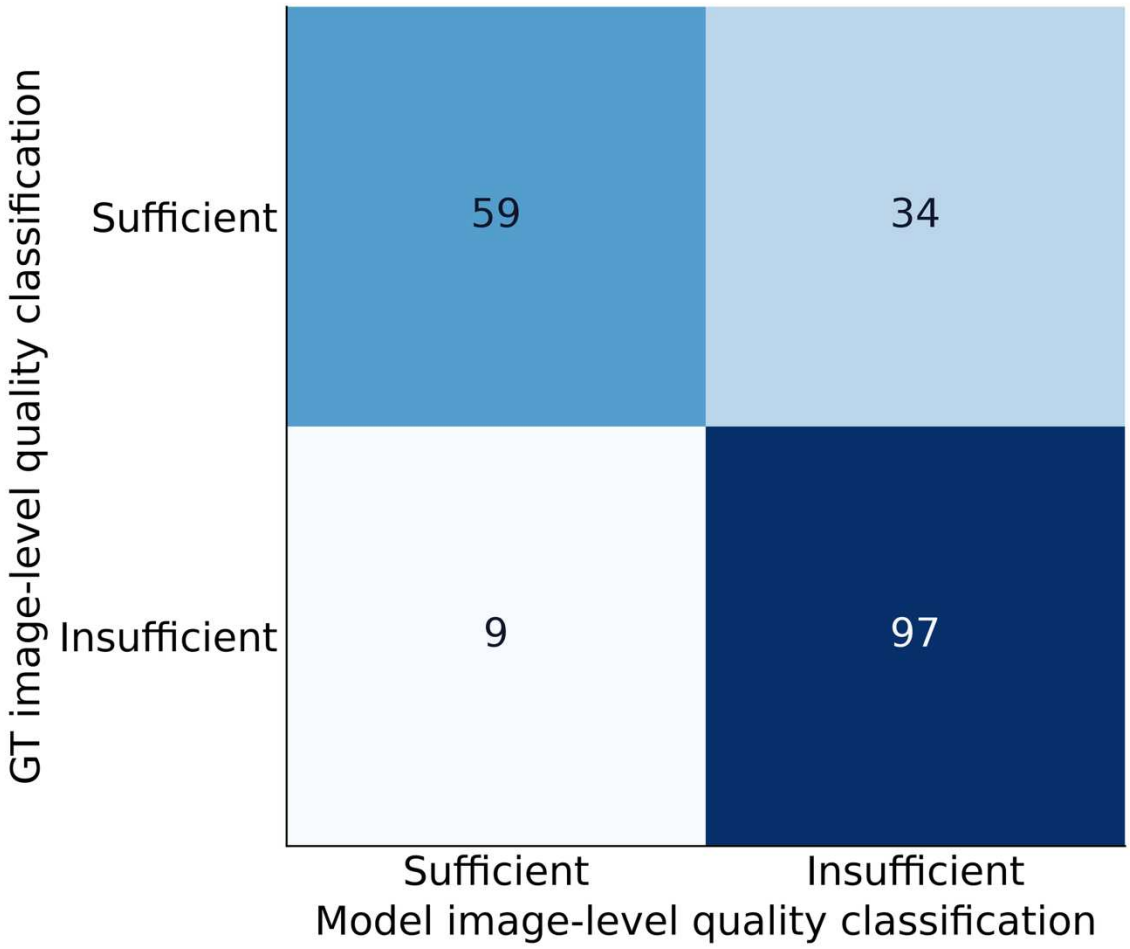


Figure 3. Ground truth (GT)-filtered image-level confusion matrix for sufficient versus insufficient image quality compared with model predictions on the 200-image test set. One image was excluded because the tooth bounding boxes failed to meet the inclusion criterion of at least one keypoint being visible at both proximal surfaces.

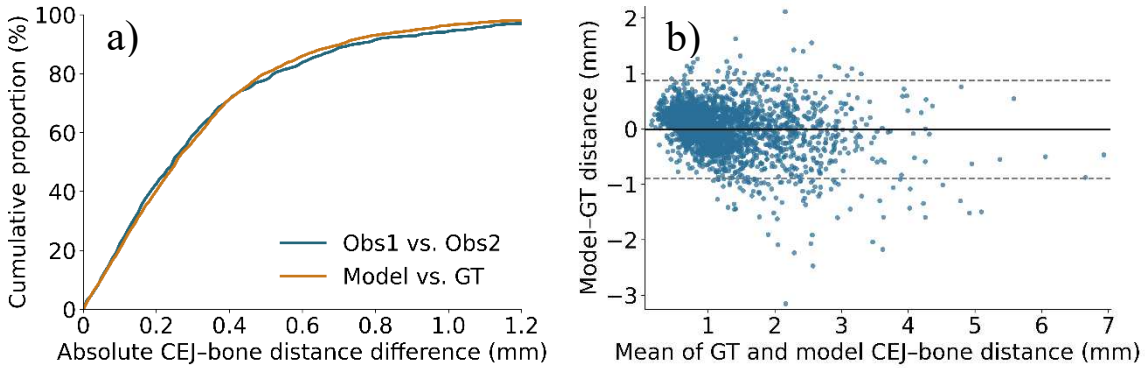
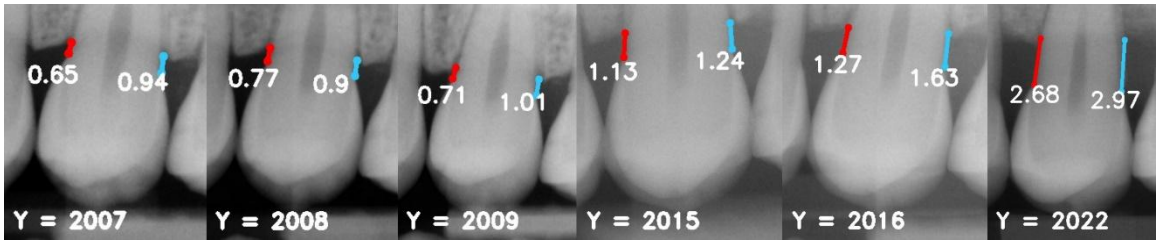
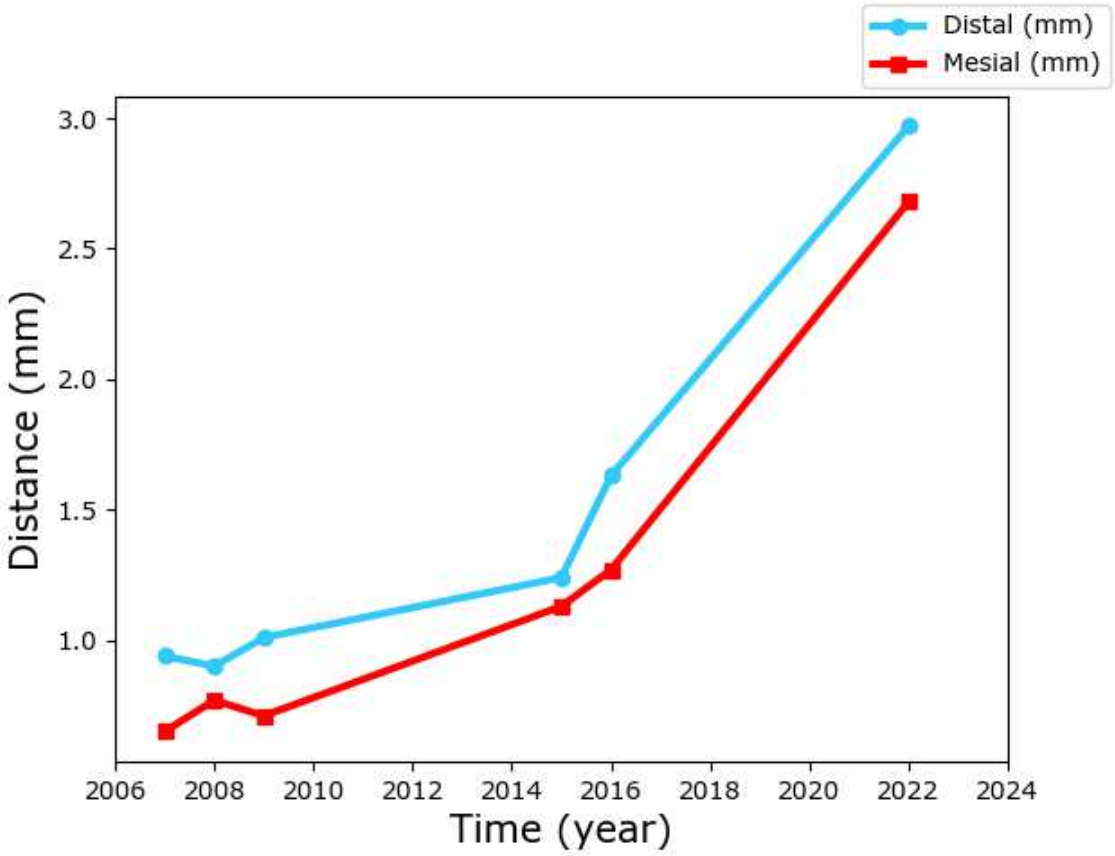


Figure 4. a) Cumulative threshold curves comparing absolute cemento–enamel junction (CEJ)–marginal bone distance differences for observer 1 (Obs1) versus Obs2 (102-image calibration subset) and model versus ground truth (GT) on 200-image test set. b) Bland–Altman plot for model versus ground truth (GT) and cemento–enamel junction (CEJ)–marginal bone level distance on the 200-image test set.

Figure 5

a)


0.65
0.94
Y = 2007
0.77
0.9
Y = 2008
0.71
1.01
Y = 2009
1.13
1.24
Y = 2015
1.27
1.63
Y = 2016
2.68
2.97
Y = 2022


Distal (mm)
Mesial (mm)
Distance (mm)
3.0
2.5
2.0
1.5
1.0
2006
2008
2010
2012
2014
2016
2018
2020
2022
2024
Time (year)

b)

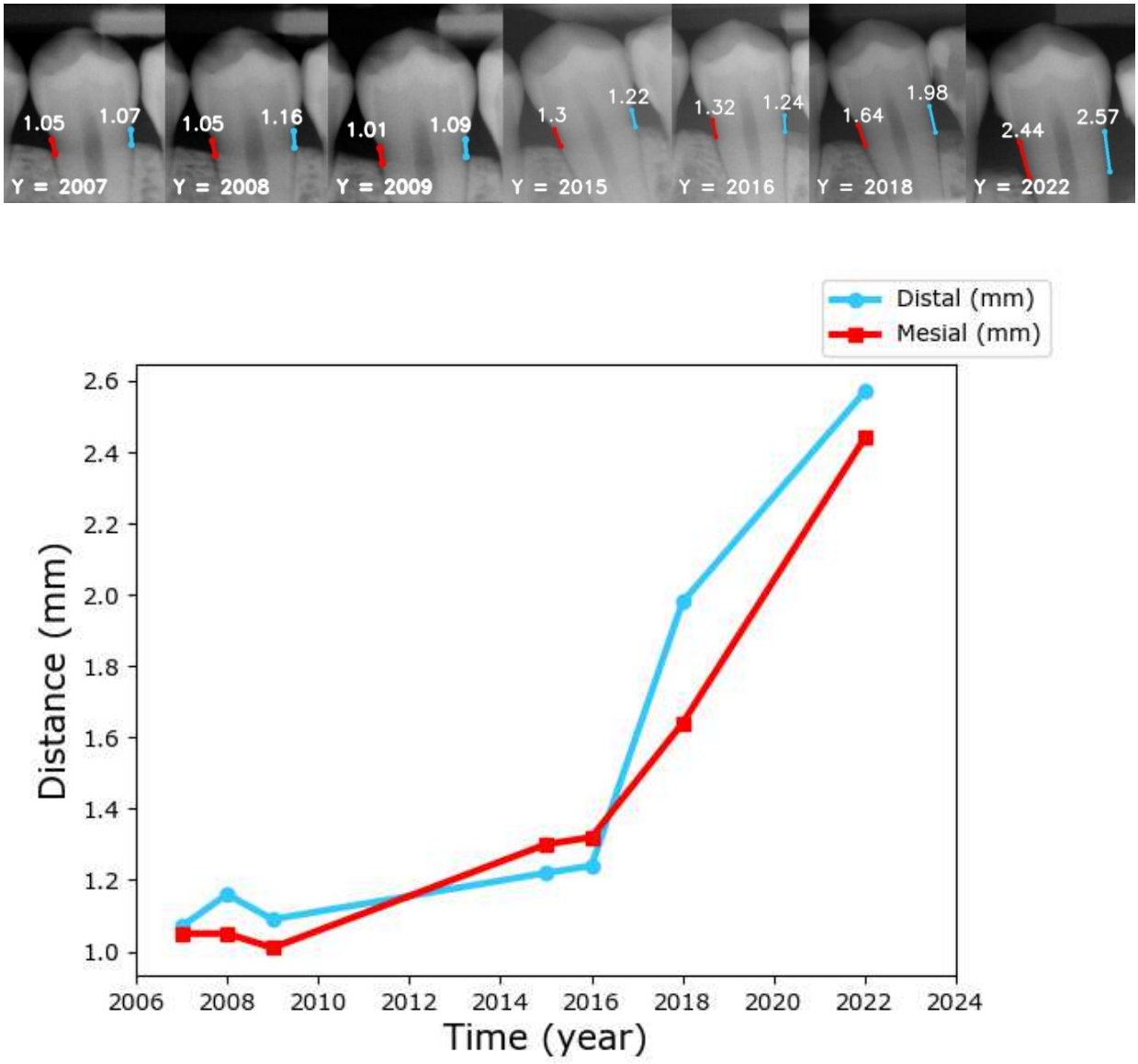


Figure 5. Example of the model's detection and measurement of Euclidean distances between the cemento–enamel junction and the marginal bone crest in bitewing radiographs obtained from the same individual (a and b, respectively) at different examination time points, with time expressed in years (Y).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SuppementaryTablesPerioAID.docx