npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02077-0

A novel AI-powered radiographic analysis surpasses specialists in stage II–IV periodontitis detection: a multicenter diagnostic study

Check for updates

Yuan Li[1,2,3,4,11], Zhiming Cui[5,11], Lanzhuju Mei[5,11], Yu Xie[1,2,4,11], Lorenzo Marini[4,6], George Pelekos[7], Wen Gu[1,2,3,4], Xiaoyu Yu[1,2,4], Xinyu Wu[1,2], Xindi Wei[2], Leran Tao[1,2], Ke Deng[7], Andrea Pilloni[4,6], Dinggang Shen[5,8,9] ✉ & Maurizio S. Tonetti[1,2,3,4,10] ✉

Missed periodontitis diagnoses are common, and AI dental radiography systems based on clinical standards can enhance reliable detection. We introduce and evaluate HC-Net+, a deep-learning model that mimics clinical pathways while integrating localized tooth lesion analyses with broad contextual understanding. This is the first AI model developed from orthopantomograms (OPGs) linked to clinical diagnoses, pre-trained and fine-tuned with 10,881 OPGs, and tested with dual benchmarking against 382 clinically labeled OPGs covering 10,198 teeth and 760 radiographically labeled OPGs from four diverse international centers. It outperformed periodontal specialists' diagnostic accuracy (AUROC: 94.2% vs. 85.6%, $p < 0.01$). The system significantly improved early periodontitis detection across training and experience levels, enabling junior dentists to match specialist performance with AI support. Performance remained consistent in the multicenter evaluation, achieving >92.4% accuracy across all locations. HC-Net+'s ability to surpass specialist accuracy while making diagnostic expertise more accessible positions it as a transformative tool for precision dentistry. The diagnostic trials were registered at ClinicalTrial.gov (NCT05513599) on 08/23/2022 and (NCT06306677) on 03/12/2024.

Periodontitis is a major public health problem. It is highly prevalent worldwide, with more than one billion adults living with severe periodontitis and a projected 44% increase by 2050[1]. Together with the resulting tooth loss, it contributes significantly to the loss of disability-adjusted life years[2]. As a chronic inflammatory disease, it increases the burden of inflammation, thereby affecting general health[3–5]. Lastly, it substantially increases health expenditures. Periodontitis can be prevented and treated. Affected subjects, however, do not recognize early symptoms and remain undiagnosed until the late stages of the disease, when treatment becomes more complex, costly, and generally less effective[6].

Diagnosis relies on an invasive, time-consuming, and technically demanding physical examination to detect periodontal bleeding upon probing and clinical attachment loss. This is performed with a periodontal probe, a thin ruler with millimeter markings, that is used to record site-specific tissue breakdown and inflammation between the gingiva and the

[1]Shanghai PerioImplant Innovation Center, Institute for Oral, Craniofacial and Sensory Research, Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China. [2]National Clinical Research Center of Oral Diseases, National Center of Stomatology, Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China. [3]Department of Oral Implantology, College of Stomatology, Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China. [4]Oral Health Biomedical Intelligence Technology (ORAL-BIT) Sino-Italian Laboratory, Shanghai, China. [5]School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai, China. [6]Section of Periodontics, Department of Oral and Maxillofacial Sciences, Sapienza University of Rome, Rome, Italy. [7]Division of Periodontology and Implant Dentistry, Faculty of Dentistry, The University of Hong Kong, SAR Hong Kong, China. [8]Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China. [9]Shanghai Clinical Research and Trial Center, Shanghai, China. [10]European Research Group on Periodontology, Brienz, Switzerland. [11]These authors contributed equally: Yuan Li, Zhiming Cui, Lanzhuju Mei, Yu Xie. ✉e-mail: Dinggang.Shen@gmail.com; tonetti@hku.hk

tooth. The definition of individual cases uses the 2018 classification system[7], which categorizes the disease into stages based on severity and complexity. Stage I cases are characterized as incipient periodontitis, which cannot be reliably distinguished from gingivitis and are therefore managed through preventive care. Stage II-III and IV cases denote initial, severe, and advanced forms of the disease, encompassing the spectrum of clinically detectable disease. In stages II-IV, comprehensive treatment procedures are needed to decrease the high risk of tooth loss. In many countries, insufficient healthcare resources and a lack of trained oral health workforce continue to be principal barriers, leading to widespread underdiagnosis, inadequate initial management of periodontitis, and inappropriate referrals to periodontal specialists[8]. This situation affects individual health outcomes and has broad socioeconomic implications.

In primary care settings (e.g., general dental practices), the detection of marginal alveolar bone loss on routine radiographs is frequently used to screen for suspected periodontitis cases. This approach is often preferred over a more comprehensive periodontal examination, which involves periodontal probing[9]. It employs the estimation of the distance between two fixed radiographic landmarks, the cementoenamel junction of the tooth (CEJ), and the perceived level of the position of the marginal alveolar bone (MAB) on the two-dimensional projection to estimate the radiographic bone loss (RBL). Due to the presence of non-mineralized supracrestal fibers extending 1–2 mm between the MAB and the CEJ, RBL exceeding this threshold is required to diagnose periodontitis. Detecting early periodontal tissue destruction on radiographs is challenging, as it is often masked, and only cases with more severe bone loss can be reliably identified. Compounding this issue, in dentistry, radiographs are not taken by radiologists but by general dentists, and image quality varies significantly, with only a minority of radiographs being of optimal quality[10]. Additionally, the interpretation of radiographs is performed by clinicians with varying levels of training and expertise. Consequently, the general reliability of identifying RBL is limited (Cohen's kappa values of 0.454–0.482), and even the intra-examiner reproducibility is moderate (kappa value of 0.739)[11]. Digital imaging or color enhancement led to only marginal improvements. Yet, panoramic dental radiographs (orthopantomograms, OPGs) are among the most frequently taken radiographs. It has been estimated that approximately 0.8% of the population in the Netherlands and 5.6% of the population in the UK undergo OPG radiographic imaging annually[12]. However, it is essential to note that OPGs are technique-sensitive and susceptible to image distortions. Consequently, the improved and standardized interpretation of RBL on these images presents a significant opportunity for the early detection of periodontitis.

Artificial intelligence agents have recently been used to assist clinicians in identifying measurement landmarks and providing automatic measurements of the RBL distance[13–19]. This can improve the accuracy of the measurement and, in some more advanced systems, can also flag a particular image for additional attention. However, a critical limitation of these systems is that they are typically trained and validated using radiographic readings by clinicians as the gold standard. Thus, their performance is inherently constrained by the same subjectivity, variability, and fundamental biological limitations (e.g., the difficulty in accurately identifying the position of the marginal alveolar bone, which only indirectly reflects the advancing front of periodontitis) that characterize conventional radiographic assessment.

To overcome this fundamental constraint, our recent work has focused on developing the HC-Net[20], a deep-learning network trained using clinical diagnoses from periodontal probing as the ground truth, rather than identifying radiographic landmarks. Clinical labels were assigned to each tooth. This end-to-end approach allows the model to learn directly from the clinical probing, which captures the true status of the periodontium. Although clinical probing is more time-consuming and resource-intensive, it identifies more subtle changes masked in conventional radiographic imaging and offers the possibility to overcome the fundamental limitation of landmark-based RBL assessment method. In addition, the HC-Net integrates tooth-level analysis and global level analysis with the clinical decision-making process for case definition[21] and provides a probability of an image being from a subject with stage II-IV periodontitis (the clinically detectable forms of periodontitis). Our network was tailored for an oral health workforce with relatively low periodontal expertise or working in a high-volume and low-resource medical setting, potentially improving accessibility and equity. HC-Net automatically reads the image and offers a tentative case definition. As demonstrated in our previous study, HC-Net performed better than other typical deep-learning image classification methods. However, HC-Net was developed based on the datasets from a single center. Further pre-training and fine-tuning of the network and external testing are needed.

We hypothesize that a fully automatic, AI-based OPG reading system can be generalized and achieve adequate diagnostic accuracy for detecting stage II-IV periodontitis, rivaling that of experienced specialists, and is superior to that of general dentists or dental students. Additionally, we propose that AI assistance can enhance the accuracy of clinicians with varying levels of expertise. The complete development and evaluation of the AI-based system is illustrated in Fig. 1. Specifically, we report sequential experiments to (i) pre-train and fine-tune HC-Net into HC-Net+, (ii) compare the performance of HC-Net+ with HC-Net, (iii) interpret the network results, iv) compare the performance of HC-Net+ with dentists of different training and expertise levels, (v) test whether HC-Net+ assistance improves the performance of dentists across different skill levels, (vi) assess the performance of HC-Net+ against a panel of experts in a multicenter, multinational, real-world dataset, and (vii) analyze misclassification.

Results

Characteristics of the datasets and populations

A total of 10,400 unlabeled radiographic images from a public dataset and four private dental clinics were collected for network pre-training. The internal development dataset consisted of 481 subjects involving 12,111 teeth, while the external dataset I comprised 382 subjects involving 10,198 teeth. All cases were labeled at the tooth level with a clinical diagnosis based on the ground truth from periodontal examinations and were collected from patients seeking dental care at Prince Philip Dental Hospital in Hong Kong (HKPPDH) and Shanghai Ninth People's Hospital Pudong Clinic (SH9HPDC). The external dataset I included 239 negative and 143 positive cases (stage II-IV periodontitis). Compared to the internal development dataset, the subjects in the external dataset I were younger and exhibited lower severity and extent of the disease, posing a challenge for validation. The external dataset II included OPG images from 760 patients across four centers. According to the periodontal specialist panel's consensus, this dataset contained 170 negative and 590 positive cases. Detailed information on subject demographics and characteristics is provided in Supplementary Tables 1–4. The experimental steps encompassing network development, pre-training, fine-tuning, and testing are illustrated in Fig. 1. A brief overview of the network architecture is presented in Fig. 2. The STARD diagrams for the two diagnostic trials are shown in Supplementary Figs. 1 and 2.

Network pre-training and fine-tuning

To enhance HC-Net's generalizability beyond single-center data, we employed a self-supervised contrastive learning strategy for pre-training, followed by supervised fine-tuning. During the pre-training phase, we utilized an extensive collection of unlabeled OPG images to learn robust and transferable feature representations. The contrastive learning framework was designed to maximize alignment between various augmented views of the same image while ensuring distinction between different images[22].

This experiment aimed to evaluate the effects of network pretraining and fine-tuning. Both internal and external testing datasets were utilized to compare the diagnostic accuracy of HC-Net+ and HC-Net. There was no significant difference between the two networks in internal testing, as illustrated in Fig. 3a–c and Table 1. With a threshold of 0.7, the percentages of correctly predicted images by HC-Net and HC-Net+ were 91.7% and 94.8%, respectively (Fig. 3g). In the external dataset I, HC-Net+ outperformed HC-Net in all accuracy estimates (Fig. 3d–f and Table 1), showing that the network's generalizability improved through pre-training

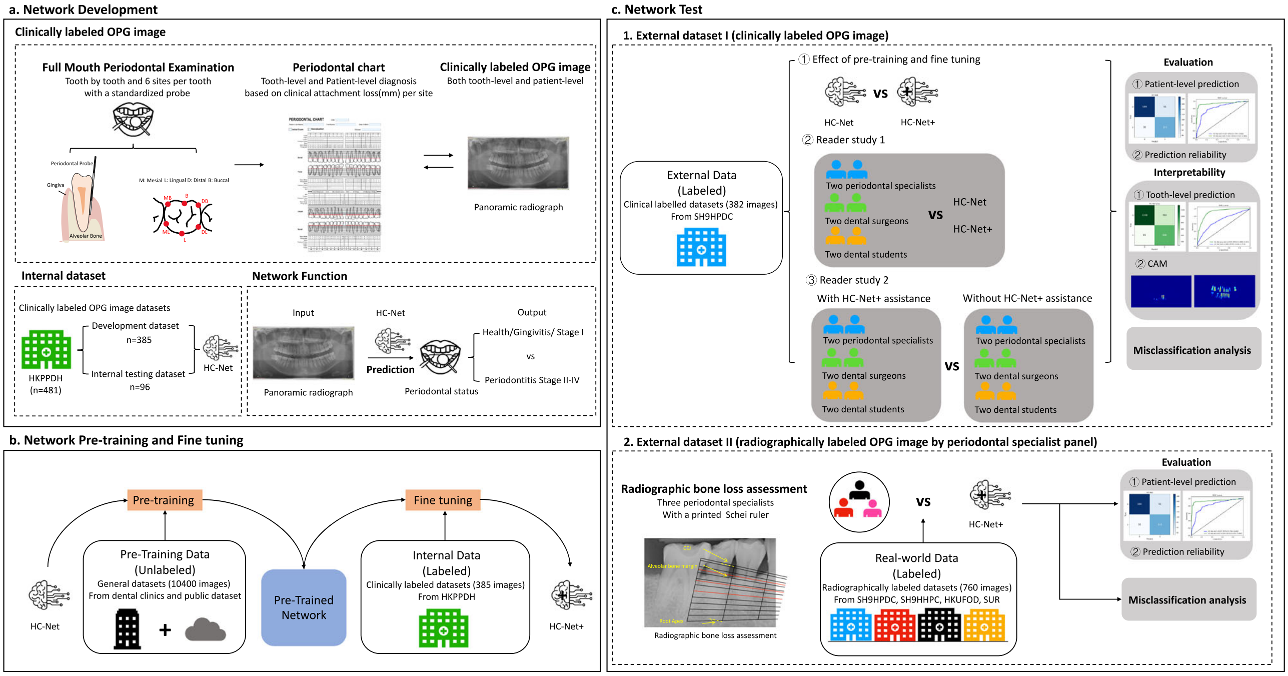

a. Network Development
Clinically labeled OPG image
Full Mouth Periodontal Examination
Tooth by tooth and 6 sites per tooth
with a standardized probe
Periodontal Probe
Gingiva
Alveolar Bone
M: Mesial L: Lingual D: Distal B: Buccal
MB
B
DB
ML
L
DL
Periodontal chart
Tooth-level and Patient-level diagnosis
based on clinical attachment loss(mm) per site
Clinically labeled OPG image
Both tooth-level and patient-level
Panoramic radiograph
Internal dataset
Clinically labeled OPG image datasets
Development dataset
n=385
Internal testing dataset
n=96
HKPPDH
(n=481)
HC-Net
Network Function
Input
HC-Net
Output
Panoramic radiograph
Prediction
Periodontal status
Health/Gingivitis/ Stage I
vs
Periodontitis Stage II-IV
b. Network Pre-training and Fine tuning
Pre-training
Fine tuning
HC-Net
Pre-Training Data
(Unlabeled)
General datasets (10400 images)
From dental clinics and public dataset
Pre-Trained
Network
Internal Data
(Labeled)
Clinically labeled datasets (385 images)
From HKPPDH
HC-Net+
c. Network Test
1. External dataset I (clinically labeled OPG image)
External Data
(Labeled)
Clinical labelled datasets (382 images)
From SH9HPDC
① Effect of pre-training and fine tuning
HC-Net
vs
HC-Net+
② Reader study 1
Two periodontal specialists
Two dental surgeons
Two dental students
vs
HC-Net
HC-Net+
③ Reader study 2
With HC-Net+ assistance
Without HC-Net+ assistance
Two periodontal specialists
Two dental surgeons
Two dental students
vs
Two periodontal specialists
Two dental surgeons
Two dental students
Evaluation
① Patient-level prediction
② Prediction reliability
Interpretability
① Tooth-level prediction
② CAM
Misclassification analysis
2. External dataset II (radiographically labeled OPG image by periodontal specialist panel)
Radiographic bone loss assessment
Three periodontal specialists
With a printed Schei ruler
CEJ
Alveolar bone margin
Root Apex
Radiographic bone loss assessment
vs
HC-Net+
Real-world Data
(Labeled)
Radiographically labeled datasets (760 images)
From SH9HPDC, SH9HHPC, HKUFOD, SUR
Evaluation
① Patient-level prediction
② Prediction reliability
Misclassification analysis

Fig. 1 | Development and validation of HC-Net+. Diagram illustrating the development and evaluation of the HC-Net+ system. **a** Network development: A calibrated examiner conducted a full-mouth periodontal examination using a standardized probe to measure probing pocket depth, bleeding on probing, and clinical attachment levels at six sites per tooth. The clinically labeled OPG images from Prince Philip Dental Hospital in Hong Kong (HKPPDH) were utilized to develop and internally test the HC-Net. HC-Net is a binary classifier that detects stage II-IV periodontitis from health/gingivitis/periodontitis stage I. **b** Network pre-training and fine-tuning: The HC-Net was pre-trained with 10,400 unlabeled OPG images from a public dataset and four dental clinics. This pre-trained network was subsequently fine-tuned using the internal dataset to produce HC-Net+. **c** Network test: The network test was conducted with two external datasets. The external dataset I was obtained as a reference through a full-mouth periodontal examination performed and interpreted by a calibrated examiner. The external dataset II was obtained from four medical centers, including Hong Kong University Faculty of Dentistry (HKUFOD), Shanghai Ninth People's Hospital Pudong Clinic (SH9HPDC), Shanghai Ninth People's Hospital Huangpu Clinic (SH9HHPC), and Sapienza University of Rome Dental Hospital, Italy (SUR), and utilized to assess the alignment of HC-Net+ with a panel of periodontal specialists. The reference for the external dataset II was the consensus decision made by a periodontal expert panel based on radiographic assessment with the assistance of a modified Schei ruler to enhance precision.

and fine-tuning. Compared to HC-Net, the number of missed diagnoses in positive cases for HC-Net+ decreased from 54 to 14, while the number of misdiagnoses in negative cases dropped from 27 to 11. With a threshold of 0.7, the percentages of correctly predicted images by HC-Net and HC-Net+ were 78.8% and 93.5%, respectively (Fig. 3h).

To systematically evaluate the impact of the pretraining data scale, we conducted ablation studies with progressively larger datasets (2000, 5000, and all 10,400 unlabeled OPGs). Performance was improved consistently with the increased pretraining data volume, as evidenced by AUROC gains of 1.1% (2000 images), 2.0% (5000 images), and 2.2% (full dataset, 10,400 images) on the internal dataset (Supplementary Table 5). This demonstrates that the self-supervised framework can effectively leverage larger datasets to enhance feature learning and generalization.

Network interpretation

To investigate the network's interpretability in detecting periodontitis stages II-IV, tooth-level prediction scores from the tooth-level analysis were analyzed, and also heatmaps from the global-level analysis were created to visualize the areas contributing most to the network's attention.

Tooth level. A tooth-level prediction score was obtained through analysis of segmented tooth patches. One-hundred stratified random samplings from the external datasets involving 2596 teeth were used to evaluate the performance of tooth-level predictions against the ground truth represented by the clinical examination of that tooth. The accuracy of HC-Net+ was 83.1%, while the accuracy of HC-Net was 68.5%. HC-Net+ achieved a sensitivity of 0.905 (95% CI 0.884–0.922) and a specificity of 0.792 (95% CI 0.772–0.810), while HC-Net attained a sensitivity of 0.571 (95% CI 0.538–0.603) and a specificity of 0.745 (95% CI 0.724–0.765, Fig. 4a–c).

Global level. The global-level analysis created the heatmaps through classification activation maps (CAMs). The confidence value was scaled between 0 and 1, where a higher number indicated greater confidence in classifying the image. Representative examples of the CAMs are visualized in Fig. 4d for the external dataset I. CAMs were tested for consistency with the periodontitis stage II-IV teeth identified by the clinical examination for each image. Consistency was high, indicating that single lesions in the dentition could be traced (Fig. 4e).

Network performance in the testing datasets

Comparison between AI (HC-Net +) and dentists. The network's diagnostic accuracy and reliability were compared with clinicians who read these images in a fully randomized, masked, crossed trial. Six dental surgeons with different levels of training and experience (two periodontal specialists, two dental surgeons, and two dental students) independently reviewed the same image datasets to detect stages II-IV twice with a two-week wash-out period. The specialists had three years of clinical periodontal experience, and the dental surgeons had three years of professional clinical experience; the dental students were postgraduate students in their first year of study. Figure 5a shows that the intra-rater repeatability was substantial to almost perfect. The inter-rater reproducibility was moderate to substantial and aligned with previous experiments, highlighting clinicians' challenges when reading these images. The confusion matrices shown in Fig. 5b–d indicate that the false negative results increased as the level of training and experience decreased. Conversely, dental students were less likely to make false-positive predictions than more experienced clinicians. Periodontal specialists and dental surgeons were more likely to make false predictions while assessing the more subtle stage I and II disease images (Fig. 5e). Dental students, however,

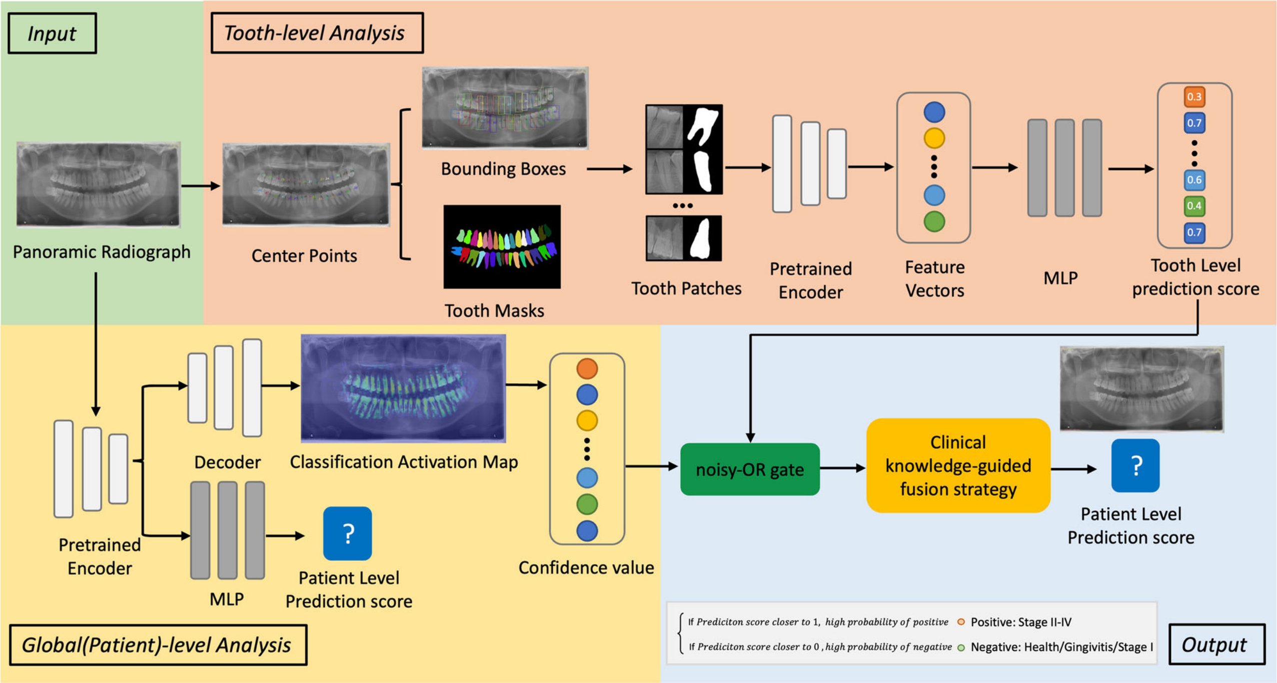

Input
Panoramic Radiograph
Tooth-level Analysis
Center Points
Bounding Boxes
Tooth Masks
Tooth Patches
Pretrained Encoder
Feature Vectors
MLP
0.3
0.7
0.6
0.4
0.7
Tooth Level prediction score
Pretrained Encoder
Decoder
Classification Activation Map
MLP
?
Patient Level Prediction score
Confidence value
Global(Patient)-level Analysis
noisy-OR gate
Clinical knowledge-guided fusion strategy
?
Patient Level Prediction score
If Prediciton score closer to 1, high probability of positive
If Prediciton score closer to 0 , high probability of negative
Positive: Stage II-IV
Negative: Health/Gingivitis/Stage I
Output

Fig. 2 | Network architecture. An illustration of the network architecture. The process begins with the input of an OPG image. The tooth-level analysis detects and segments each tooth, followed by a tooth-level classification network that determines the prediction score for each tooth. Meanwhile, the global(patient)-level analysis executes a full image prediction and employs the classification activation map (CAM) to enhance the network's attention to each tooth. The global (patient)-level prediction result is obtained through a fusion strategy that combines information from both sections. The clinical diagnostic principle has been integrated into the fusion strategy, where, if at least two teeth are likely classified as stage II-IV periodontitis, the case will be considered to have a high probability of being diagnosed with stage II-IV periodontitis. Please see "Methods" for a detailed description.

misclassified a significant proportion of stage III cases. These data confirm that clinicians had difficulties correctly classifying the initial stages of periodontitis.

Table 1 shows that HC-Net+ outperformed the dentists with different levels of training and experience in terms of diagnostic accuracy and reliability. HC-Net+ achieved the highest accuracy and was significantly better than the clinicians (McNemar's test, $P < 0.01$) in all tested parameters except for the sensitivity of the specialists. The specialists exhibited a slightly lower accuracy compared to HC-Net +. The dental surgeons and dental students demonstrated significantly lower accuracy. However, the specificity of dental students was nearly equivalent to that of the HC-Net +. Triplicate runs of HC-Net+ gave identical results.

Stage II periodontitis had the highest rate of missed diagnosis (Figs. 3i and 5e). The missed diagnosis rates for stage II were 20.6% (13/63) for HC-Net+, 25.4% (16/63) for specialists, 55.6% (35/63) for HC-Net, 44.4% (28/63) for dental surgeons, and 88.9% (56/63) for dental students.

Improvement of dentist's performance with AI (HC-Net +) assistance. A randomized masked experiment tested the effect of HC-Net+ assistance on clinician performance. After training on interpreting AI assistance, raters were sequentially asked to assess radiographic images alone and then images supplemented with the patient-level AI prediction value and an image of the CAM heatmap generated by HC-Net +. Supplementary Fig. 3 shows an example of the questionnaire, which included 382 cases. Once a question was answered, the previous response could not be modified. The same six dentists participated in this experiment.

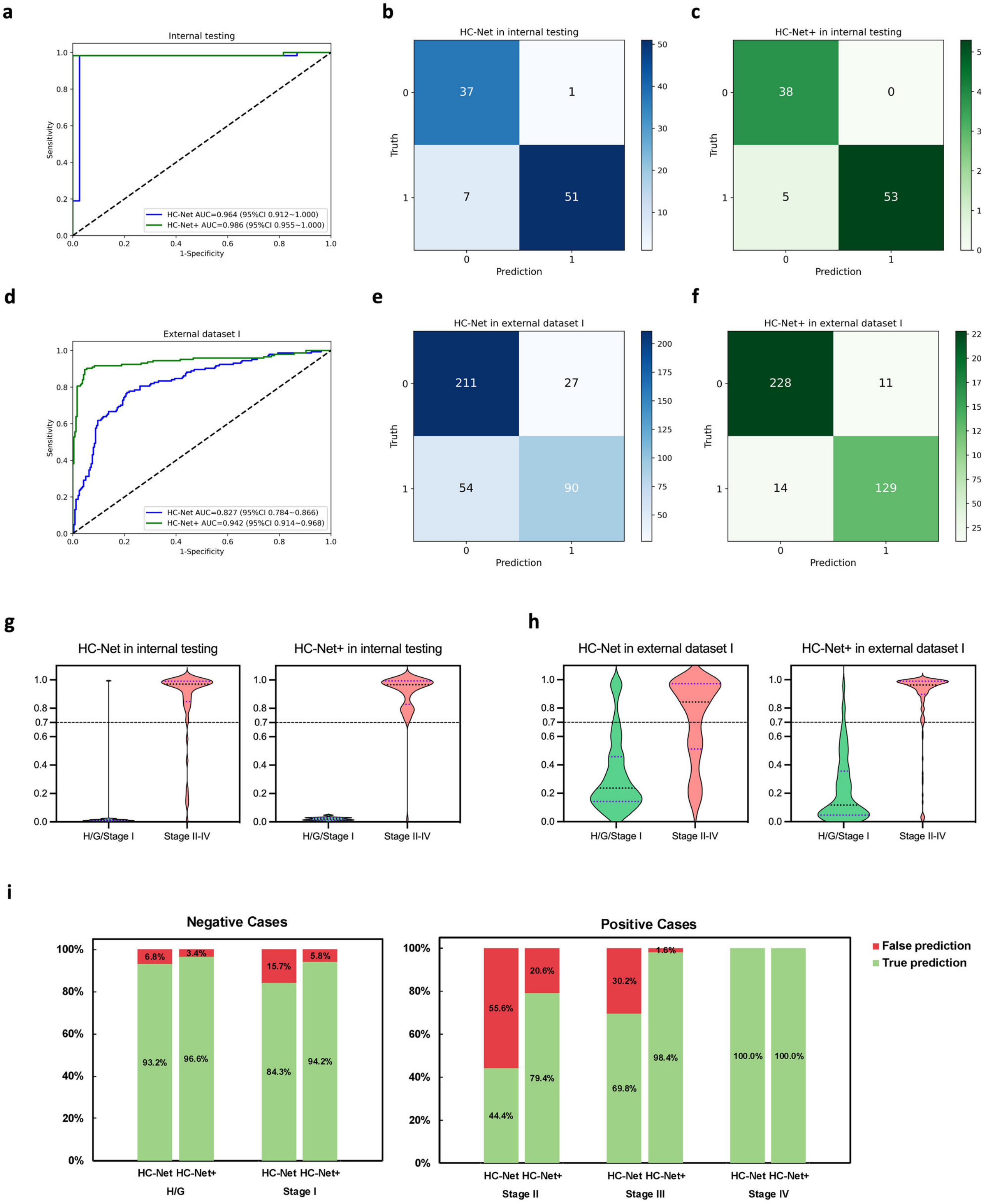
Table 1 shows significant improvement in AUROC, accuracy, sensitivity, and specificity across dentists with different levels of training and experience. The enhancements in diagnostic metrics suggest that HC-Net+ positively influenced the dentists' diagnostic decisions, outweighing any response changes it introduced.

Performance of HC-Net+ in external dataset II. External dataset II performance was assessed in a multicenter, multinational diagnostic trial including 760 cases. The reference test was the consensus of periodontal specialists reading the OPG images in a two-round, two-step fashion (Fig. 6a). The intra-rater consistency was generally almost perfect, with kappa values exceeding 0.8. However, the inter-rater consistency among the three experts was only substantial, with a Fleiss kappa value of only 0.65. The duplicate adjudication demonstrated a consistency of 0.64 for cases with inconsistent results. Regarding the time required for the expert diagnosis, 30 ± 28 seconds per case were necessary for healthy/gingivitis and stage I periodontitis, 62 ± 51 seconds for stage II periodontitis, and 29 ± 37 seconds for stage III-IV cases. The time required significantly differed by case type and rater ($P < 0.01$). This indicates that diagnostic results based on a single panoramic image may be unstable among experts and point to potential label mistakes, particularly for the challenging stage II cases.

Figure 6b–e and Supplementary Table 6 detail HC-Net +'s performance in external dataset II. HC-Net+ achieved an overall AUROC of 0.967 (95%CI 0.953–0.978), a sensitivity of 0.956 (95%CI 0.939–0.973), and a specificity of 0.826 (95%CI 0.771–0.879). With a threshold of 0.7, the percentage of correctly predicted images by HC-Net+ was 92.4% (Fig. 6c). HC-Net+ demonstrated robust generalization capability, maintaining consistent diagnostic performance across multicenter settings with substantial heterogeneity in disease prevalence and image quality, as detailed in Supplementary Tables 3 and 4. High sensitivity and lower specificity were also found in each center. These characteristics are consistent with the potential use of OPG images for screening purposes. Besides, the average time needed for HC-Net+ evaluation of each image was 0.02 ± 0.01 s.

Analysis of network misclassification

Disease severity and extent. Among the 382 images in the external dataset I, 25 were misclassified by the HC-Net+. For missed diagnoses in positive cases, we explored the effect of disease severity and extent on



network performance (Supplementary Fig. 4a, b). The number of teeth with 3 mm clinical attachment loss (CAL) (i.e., the amount of clinical attachment loss that can reliably be detected in routine clinical examination and is required for stage II case diagnosis) or greater and its corresponding percentage, were calculated for each positive case. These 14 missed diagnosed images were difficult to classify because they corresponded to localized disease in which less than 5 (or 20%) teeth had 3 mm CAL. For 11 misdiagnoses in negative cases, the most frequent error was associated with a foreign body, and 63.6% (7/11) of images contained multiple foreign bodies with a high radiographic density, such as dental implants, crowns, and fillings were misclassified as diseased.

As for the external dataset II, the distribution of prediction scores determined by HC-Net+ are shown in Fig. 6c. With a threshold of 0.7, the 58 inconsistent cases were observed in H/G/stage I (17.6%, 33/188), stage II

Fig. 3 | Effect of pre-training and fine-tuning. The network was pre-trained using self-supervised contrastive learning on 10,400 OPG images, followed by supervised fine-tuning on the internal dataset to enhance generalizability. **a**–**c** The ROC curves and confusion matrices for HC-Net and HC-Net+ in detecting stage II-IV periodontitis on the internal dataset are presented. **d**–**f** The ROC curves and confusion matrices for HC-Net and HC-Net+ in detecting stage II-IV periodontitis on external dataset I are shown. In the confusion matrices, the horizontal axis represents the predicted label, while the vertical axis displays the true label. **g** The distribution of patient-level prediction scores for negative cases (health/gingivitis/periodontitis stage I, denoted as H/G/Stage I, shown in green) and positive cases (stage II-IV periodontitis, denoted as stage II-IV, shown in red) as determined by HC-Net and HC-Net+ on the internal dataset is illustrated. The violin plot displays the complete range of prediction scores. The prediction scores range from 0 to 1, with scores closer to 1 indicating a higher probability of stage II-IV periodontitis. The cut-off value was consistently set at 0.7 for both networks based on the distribution characteristics of the prediction scores. **h** The distribution of patient-level prediction scores for negative and positive cases determined by HC-Net and HC-Net+ on the external dataset I is shown. The cut-off value remains consistent with that of the internal testing, which is 0.7. **i** The percentage of falsely (red) and correctly (green) predicted cases by HC-Net and HC-Net+ in periodontal health/gingivitis/stage I periodontitis as well as stage II, stage III, and stage IV periodontitis, respectively, is presented. The false prediction rates decreased with HC-Net+.

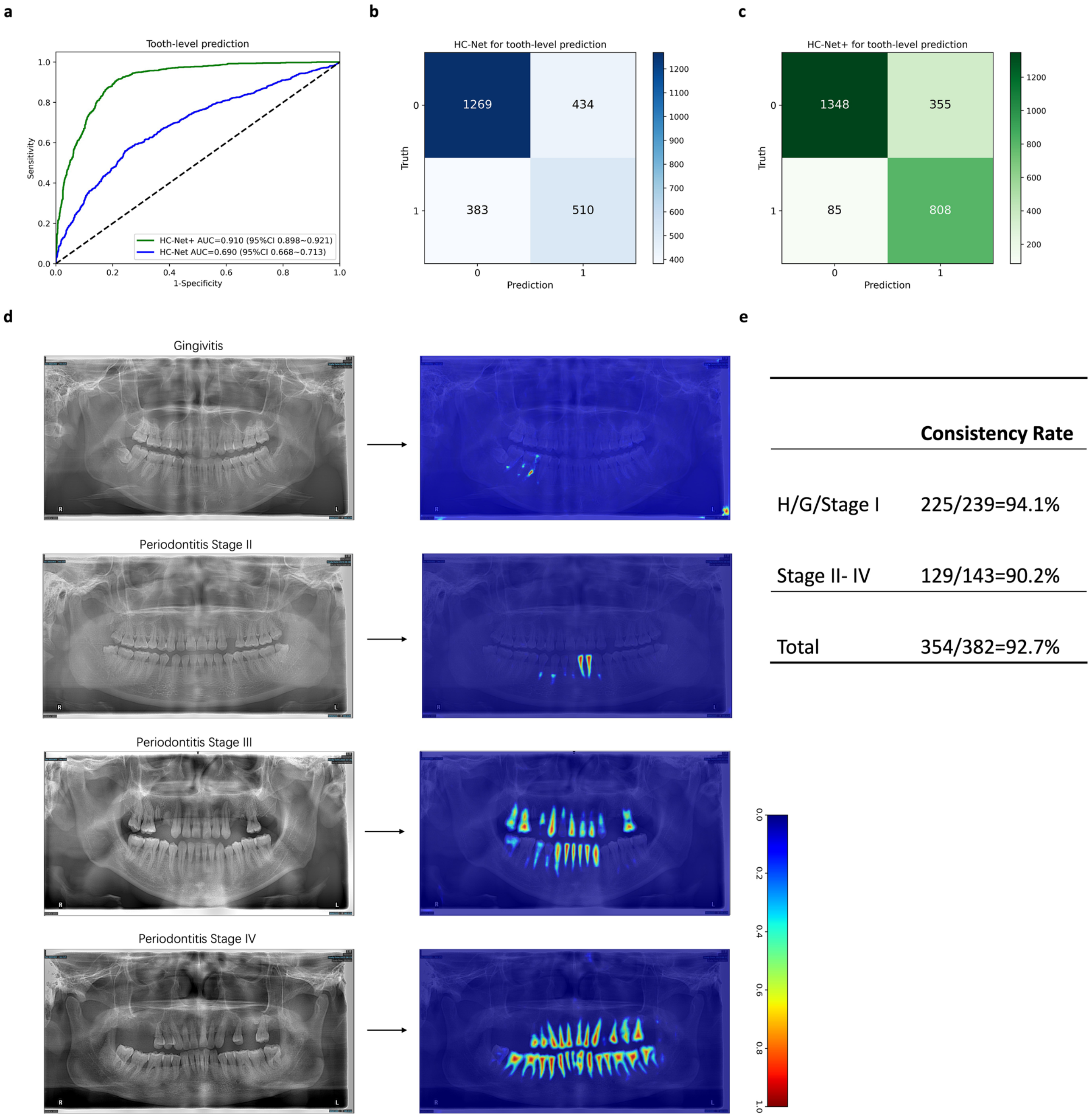
Table 1 | Comparisons of performances of networks and dentists with different training and expertise in predicting stage II-IV periodontitis

Networks and Dentists	Δ Change^	AUROC (95%CI)	Sensitivity (95%CI)	Specificity (95%CI)	Accuracy (95%CI)
① Effect of pre-training and fine tuning					
Internal testing dataset					
HC-Net		0.964 (0.956–0.972)	0.879 (0.795–0.963)	0.974 (0.862–0.999)	91.7 (84.6–96.3)
HC-Net+		0.986 (0.98–0.99)	0.914 (0.842–0.986)	1.000 (0.907–1.000)	94.8 (88.3–97.8)
External dataset I					
HC-Net		0.827 (0.784–0.866) **	0.625 (0.544–0.70) **	0.887 (0.84–0.921) **	78.8 (74.7–82.9) **
HC-Net+		0.942 (0.914–0.968)	0.902 (0.842–0.94)	0.954 (0.919–0.974)	(90.9–95.9)
② Reader study 1: AI vs Dentist					
External dataset I					
HC-Net+		0.942 (0.914–0.968)	0.902 (0.842–0.94)	0.954 (0.919–0.974)	93.5 (90.9–95.9)
Periodontal specialist		0.856 (0.784–0.927) **	0.874 (0.814–0.921)	0.836 (0.783–0.877) **	85.1 (81.6–88.7) **
Dental surgeon		0.799 (0.716–0.884) **	0.762 (0.692–0.814) **	0.836 (0.787–0.876) **	80.9 (77–84.9) **
Dental student		0.679 (0.595–0.765) **	0.405 (0.331–0.489) **	0.953 (0.921–0.971) **	(70.5–79.2) **
③ Reader study 2: AI-assisted vs unassisted					
External dataset I					
Periodontal specialist		0.857(0.831–0.883) **	0.851(0.803–0.889) **	0.863(0.829–0.892) **	85.9(83.2–88.2) **
Periodontal specialist with HC-Net+	7.9%	0.911(0.891–0.932)	0.924(0.885–0.95)	0.899(0.868–0.924)	90.8(88.9–92.9)
Δ Change^^		+5.4%	+7.3%	+3.6%	+4.9%
Dental surgeon		0.83(0.802–0.857) **	0.813(0.762–0.855) **	0.847(0.81–0.877) **	83.4(80.7–86) *
Dental surgeon with HC-Net+	6.2%	0.868(0.843–0.893)	0.854(0.807–0.892)	0.882(0.849–0.909)	87.2(84.9–89.6)
Δ Change^^		+3.8%	+4.1%	+3.5%	+3.8%
Dental student		0.786(0.755–0.816) **	0.712(0.655–0.763) **	0.859(0.824–0.889) **	80.4(77.5–83.1) **
Dental student with HC-Net+	8.2%	0.844(0.817–0.872)	0.792(0.739–0.836)	0.897(0.865–0.922)	85.7(83.3–88.2)
Δ Change^^		+5.8%	+8%	+3.8%	+5.3%

AUROC area under the receiver operating characteristic curve,; *CI* confidence interval. Accuracy was presented as percentage. DeLong test was used to compare AUROC, and McNemar's test was used to compare sensitivity, specificity and accuracy. *$p < 0.05$; **$p < 0.01$.
In experiment ① and ② comparisons, HC-Net+ was set as the reference. In experiment ③ comparisons, the group with AI assistance was set as reference.
Δ Change^: The rate response change between AI AI-assisted and AI unassisted group.
Δ Change^^: The change between AI AI-assisted and AI-unassisted groups in each accuracy metric.

(14.7%, 22/150) and stage III/IV (0.7%, 3/422), as shown in Fig. 6d. Misclassifications predominantly occurred in early-stage periodontitis (H/G/stage I and stage II; 94.8% of errors, 55/58), reflecting inherent challenges in differentiating incipient lesions on single OPG image and potential specialist labeling inconsistencies in borderline cases.

OPG image quality. Image quality is a critical determinant of diagnostic accuracy. This experiment tested the impact of OPG image quality on the diagnostic performance of HC-Net+. Evaluation criteria, adapted from a validated tool[23], assessed OPG quality based on image blurring, periodontal ligament visibility, alveolar bone crest definition, and root apex localization. Each OPG was categorized as diagnostically optimal, diagnostically adequate, or poor but diagnosable. Two dentists independently assessed all images under standardized viewing conditions, resolving discrepancies through consensus review. Supplementary Table 7 shows that HC-Net+ maintained robust diagnostic performance across different quality images in the external dataset I. In the external dataset II, however, performance was significantly higher in better-quality images. This points to the ability of HC-Net+ to work well in images of different diagnostic quality and perhaps to the difficulty of the panel of specialists in correctly rating lower-quality ones. The effect of image quality on false predictions is shown in Supplementary Fig. 4c and d. False-positive



	Consistency Rate
H/G/Stage I	225/239=94.1%
Stage II- IV	129/143=90.2%
Total	354/382=92.7%

Fig. 4 | Network Interpretation. Experiments illustrate network interpretability at the tooth and global (patient) levels. The accuracy of tooth-level network predictions was assessed on a random stratified sample from the external dataset I comprising 100 patients and 2596 teeth. ROC curves (**a**) and confusion matrices (**b** and **c**) show the HC-Net (blue line and matrix) and HC-Net + (green line and matrix) for detecting stage II-IV periodontitis teeth. Predictions are based only on the tooth-level analysis section of the network and are evaluated against the tooth-specific ground truth based on clinical attachment level. HC-Net+ shows better performance than HC-Net. Individual tooth predictions result in better global(patient)-based predictions as multiple teeth are evaluated in the same case. The left column of figures in (**d**) shows examples of OPG images of negative (health/gingivitis/stage I periodontitis) and positive cases (stage II-IV periodontitis). The right column displays the heatmaps generated by the classification activation map of the global(patient)-level analysis section of the network. The heatmap illustrates the probability of the tooth within the image being classified as stage II-IV periodontitis, and the confidence values are shown in the bar chart (high probability in red, low probability in blue, and a probability close to 0.7 threshold in orange). **e** shows the rate of consistency between the teeth identified as stage II-IV periodontitis (positive) by classification activation map and the actual presence of clinical attachment loss (CAL) ≥ 3 mm of each tooth in an external dataset. The consistency rate was calculated by dividing the number of consistent cases by the total number of cases.

predictions were significantly more frequent in difficult-to-diagnose stage II periodontitis with poor-quality images.

2D imaging constraints. A fundamental limitation arises from the nature of panoramic radiography itself. As a 2D projection modality, OPGs inherently lack detailed crown texture information and 3D structural context, which are critical for identifying early periodontal changes such as gingival inflammation or incipient alveolar bone loss. While HC-Net+ demonstrates strong performance in detecting established bone loss patterns (stages II–IV), its ability to discriminate subtle early-stage lesions (e.g., gingivitis or stage I periodontitis) remains constrained by the resolution and dimensionality of imaging modality.

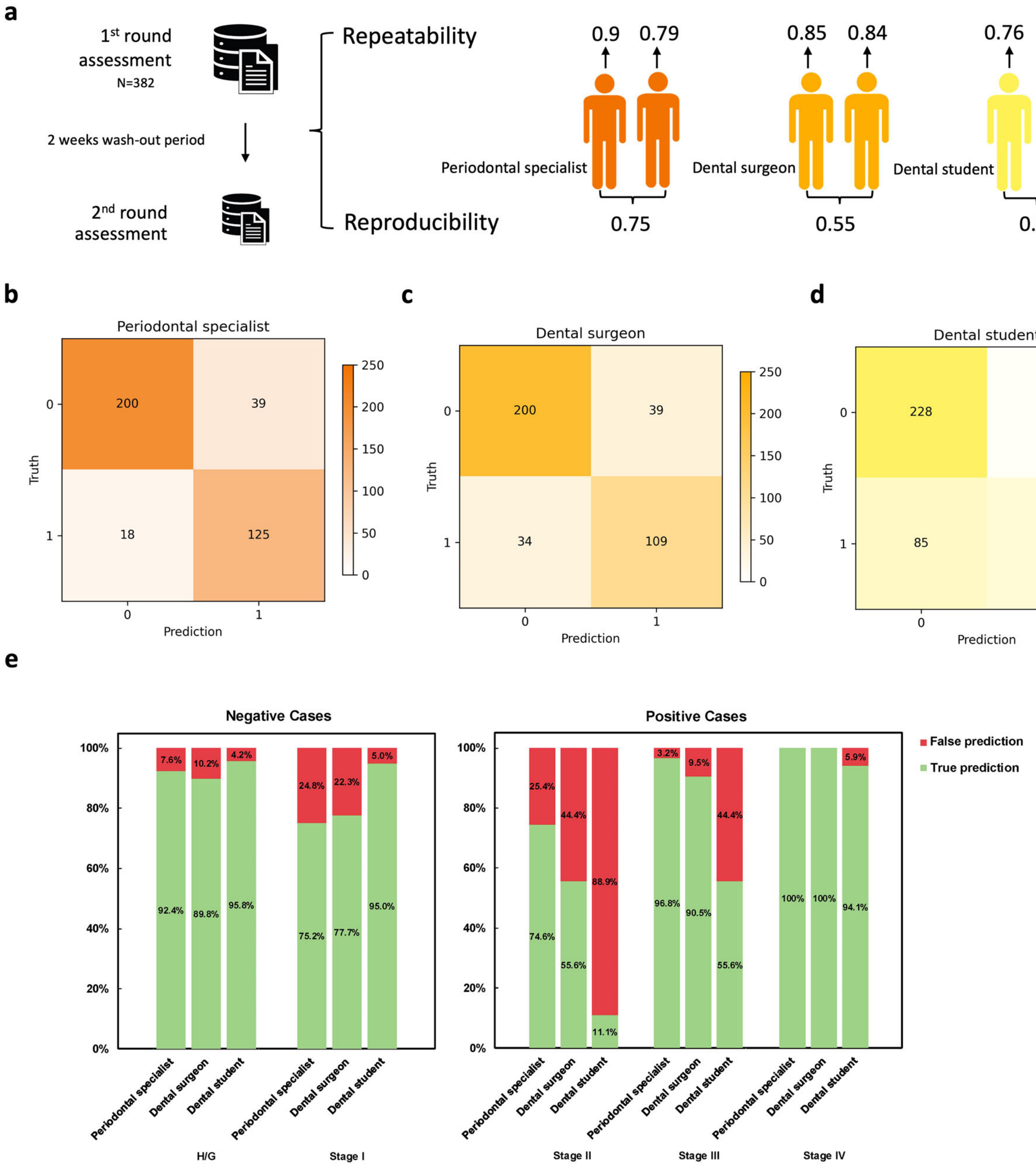


Fig. 5 | Diagnostic accuracy of dentists with different training and expertise. This experiment showcases the reliability and diagnostic accuracy of dental surgeons across three levels of training and expertise (i.e., periodontal specialist, dental surgeon and dental student) in identifying stages II-IV periodontitis using external dataset I. **a** This task requires dentists to review OPG images from external dataset I, assessing the consistency of each dentist's evaluation and the reproducibility among dentists with the same experience level. **b**–**d** The confusion matrices illustrate how periodontal specialists, dental surgeons and dental students perform in identifying stage II-IV periodontitis on the external dataset I. In each confusion matrix, the horizontal axis indicates the predicted label, while the vertical axis indicates the true label. **e** The rates of false and accurate predictions are provided for dentists with different levels of training and experience in identifying periodontal health/gingivitis (H/G) as well as stages I, stage II, stage III, and stage IV periodontitis, respectively. The true and false prediction rates differ significantly among dentists with varying training and experience, and also for different diagnosis ($p < 0.01$).

Future work will focus on multimodal fusion strategies, integrating OPGs with intra-oral photographs or 3D CBCT scans to leverage complementary visual features of soft tissue inflammation and early bone changes.

Discussion

Our experiments and diagnostic trials showed that our AI network detects stage II-IV periodontitis in OPG images more effectively than clinicians, thereby enhancing their diagnostic accuracy, regardless of their training and expertise levels. Additionally, HC-Net+ enabled the accurate identification of stage II-IV periodontitis and performed well with images of suboptimal quality, providing decisive improvements to conventional, landmark-based AI assistance. Our network was designed for an oral health workforce with limited periodontal expertise in high-volume, low-resource settings by providing specialist-level, automated radiographic analyses, potentially enhancing diagnostic accessibility and equity. OPGs are often taken in primary care across various health systems worldwide, and introducing our network could significantly improve early detection of periodontitis, thereby contributing to better treatment efficacy and cost containment.

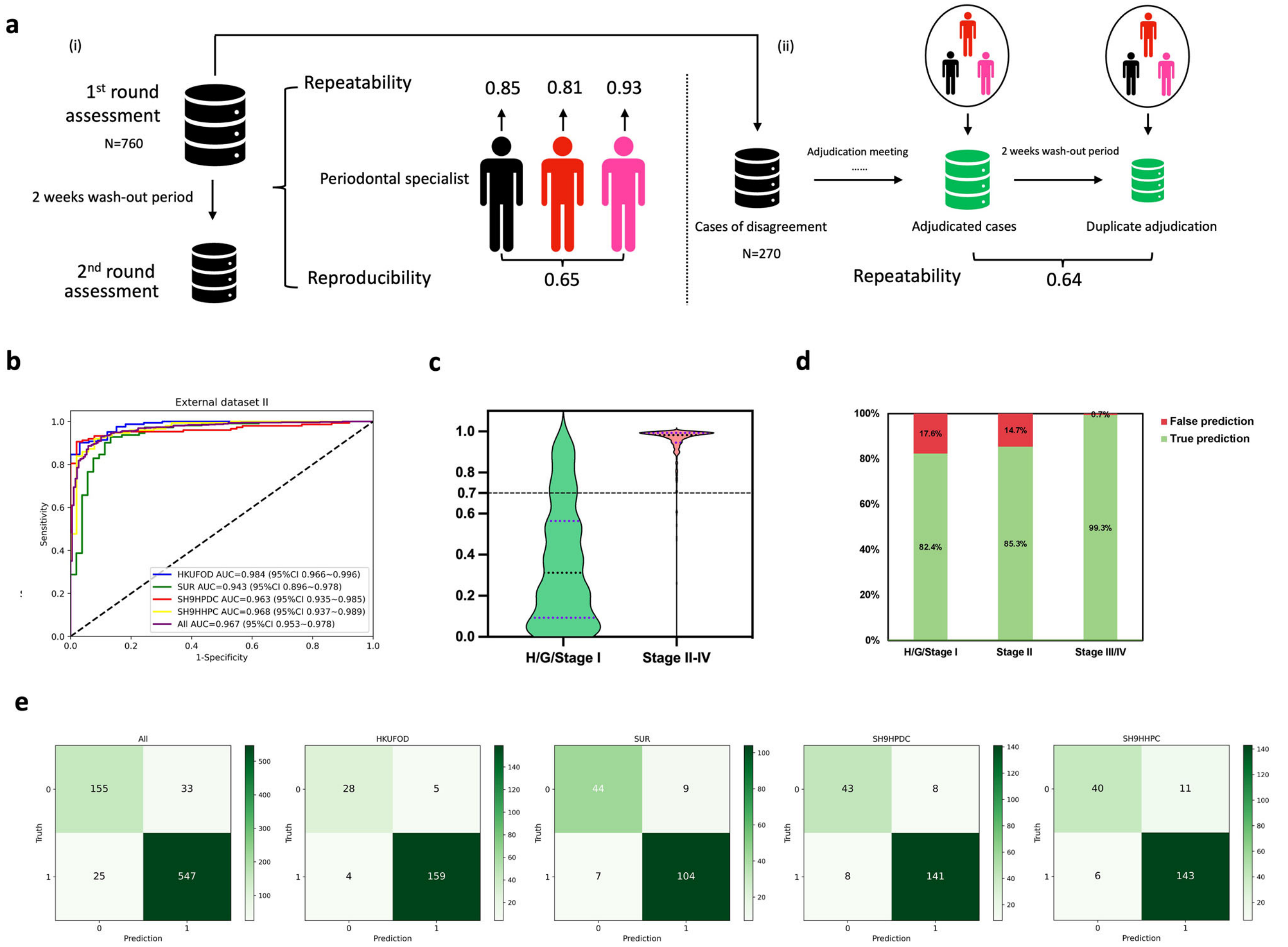


Fig. 6 | The performance of HC-Net+ in external dataset II. The panel of periodontal experts categorized the OPG images into three groups: (1) healthy, gingivitis, and stage I periodontitis, (2) stage II periodontitis, and (3) stage III-IV periodontitis, through a two-round, two-step adjudication process. Each periodontal specialist read all 760 cases in duplicate, allowing for a two-week washout period (36 ± 41 seconds per case). Specialists disagreed on 270 cases, which were jointly assessed twice by the panel of specialists, with another two-week washout period (121 ± 94 seconds per case). For analysis, stage II cases were combined with stage III-IV cases. **a**-i The intra-rater repeatability was nearly perfect, and the inter-rater reproducibility was substantial. **a**-ii The repeatability of the panel was also substantial. However, the results highlight the raters' difficulty in identifying the periodontitis stage diagnosis based on the radiographs. **b** The ROC curves of HC-Net+ for detecting stage II-IV periodontitis in OPG images were generated for each center and across all external dataset II. **c** The distribution of global(patient)-level prediction scores for healthy/gingivitis/stage I and stage II-IV as determined by HC-Net+ on the external dataset II was analyzed. A cut-off threshold of 0.7 was used to generate the confusion matrices. **d** The percentage of cases correctly and falsely predicted by HC-Net+ with the case definitions by the panel of specialists for periodontal health/gingivitis/stage I, stage II periodontitis, and stage III-IV periodontitis were reported, respectively. **e** The confusion matrices for HC-Net+ in detecting stage II-IV periodontitis in OPG images from each center and all external dataset II were created (the complete accuracy analysis is shown in Supplementary Table 6).

Our results originate from a few key technical choices designed to overcome the fundamental limitations of interpreting dental radiographic images: (i) labeling individual teeth and cases based on labor-intensive but highly sensitive periodontal probing; (ii) extensively pre-training on a large unlabeled dataset, capturing comprehensive features from original OPGs and improving the robustness of the network; and (iii) constructing our network based on the fusion of site-specific, global and clinical knowledge-guided strategies. HC-Net+ performance consistently outperforms dentists of varying training and expertise levels in terms of accuracy and stability. More importantly, this model has the potential to serve as an effective diagnostic aid, significantly enhancing diagnostic accuracy and showing promising prospects for use as a stand-alone tool or adjunct in clinical practice.

HC-Net outperformed other competitive algorithms on the internal test set in our previous publication[20]. However, HC-Net+ demonstrated improved generalization capability and robustness across different datasets compared to the original HC-Net. This suggests that this two-stage training approach effectively combines the advantages of large-scale self-supervised learning with task-specific supervised fine-tuning, enabling the network to learn more generalizable features while maintaining high diagnostic accuracy. The enhanced model showed particular improvement in handling variations in image acquisition parameters and center-specific characteristics.

The high accuracy was also attributed to the network architecture. The diagnostic information contained in large-scale OPG images is complex and frequently overlooked, which can result in unstable and inaccurate outcomes when relying on intuitive, whole-image analysis. Therefore, a well-designed tooth mapping from the OPG image was added to align the clinical examination procedure, which needs probing around each tooth. The hybrid approach designed in our network effectively leverages both local tooth-level and global image-level features. By integrating clinical diagnostic principles into the fusion strategy, this network can perform prediction in accordance with the clinical diagnostic logic of dentists. Additionally, it is worth noting that the external validation of this study found the performance of a periodontal specialist, based on a single OPG image, to be consistent with that reported in other studies[24,25]. This suggests that the level of the dentists involved in this study is reliable. Therefore, the results of the comparative experiments are valid, indicating that the superior performance of the model is not due to the low competence of the dentists.

Another essential element of this network is the interpretability. The results of predictions at the tooth level and the teeth identified with stage II-IV periodontitis on the heatmap demonstrated that both the tooth-level and global(patient)-level analyses within the architecture provide strong interpretability. This interpretability is crucial as it allows users to understand how the model achieves the final result. For this reason, the model can effectively assist clinicians in improving their diagnostic accuracy.

Due to the lack of matched full-mouth periodontal examination data for a large number of OPG images in the real world, this study used the bone loss assessments by a panel of periodontal specialists as a reference standard. In the external dataset II from four dental centers, HC-Net+ exhibited robust multicenter generalizability, sustaining diagnostic consistency across heterogeneous populations. However, it should be noted that the moderate inter-rater agreement (Fleiss' $\kappa = 0.64$) among specialists indicates that screening results based on a single OPG image may be unstable and suggests potential mistakes in the labeling of borderline cases. This may inflate false-negative rates in early-stage periodontitis (i.e., Stage II) and modestly bias AUROC estimates.

To the best of the authors' knowledge, our approach is the first work using clinical periodontal probing results as the ground truth for network learning, rather than relying on distance measurements of bone loss from radiographs. This fundamental difference in ground truth limits direct comparison with previous works[13–19]. Periodontal probing depth and clinical attachment loss measurements obtained through comprehensive clinical examination remain the gold standard for periodontitis diagnosis in clinical practice. By directly training our model on these clinical outcomes rather than intermediate radiographic measurements, HC-Net+ learns to associate radiographic features with actual disease states as determined by clinical examination. This approach bridges the gap between radiographic findings and clinical reality, making the AI system more clinically relevant than those trained solely on image-derived measurements.

There are limitations to this study. First, there is still a lack of a large number of OPG images labeled with clinical examinations to train and validate the network. Second, the accuracy and reliability of HC-Net+ may be significantly influenced by the OPG image quality, particularly for early-stage lesions. This limitation stems from the current absence of an integrated image quality assessment and adjustment module. It should be incorporated and validated to ensure diagnostic reliability across diverse image qualities for future deployment. Third, the model's diagnostic performance for localized stage II periodontitis is better than that of clinicians but not entirely stable, with multiple instances of false negatives. This is related to the number of affected teeth and the model needs additional OPG images of patients diagnosed with localized stage II periodontitis for targeted training. Fourth, due to the objective nature of disease progression and inherent limitation of OPG imaging modality, stage I periodontitis was grouped with periodontal health and gingivitis as negative cases, and cannot be detected by the HC-Net + . Future work may require multimodal data.

In conclusion, HC-Net+ represents a significant advancement in periodontitis screening, offering a highly accurate, reliable, and interpretable tool for dentists. The novel framework and the integration of pre-training and fine-tuning processes greatly enhanced the model's performance across various diagnostic metrics. Moreover, HC-Net+ not only surpassed the performance of dentists, but also demonstrated the ability to improve their diagnostic accuracy when used as an assistive tool. However, its performance in early-stage localized disease remains a challenge and may limit screening sensitivity. Future research should focus on fine-tuning the model's sensitivity in early-stage localized cases to facilitate clinical adoption.

Methods

Study design and datasets

This cross-sectional diagnostic study aims to improve the generalization of our previously developed HC-Net[20] and validate its performance in multicenter datasets. The model was developed to detect stage II-IV periodontitis in OPG images (Negative cases: Periodontal health, gingivitis, or stage I periodontitis; Positive cases: Stage II–IV periodontitis). The OPG datasets for network pre-training were collected from a public dataset (DENTEX MICCAI 2023 CHALLENGE) and four private dental clinics in Shanghai. The OPG datasets for network tests involve two external datasets. The inclusion criteria were: (1) subjects aged 18 or above who received panoramic radiographic examination. The exclusion criteria were: (1) edentulous adults, (2) pregnant females. The external dataset I was collected from a convenience sample of consecutive adult patients (>18 years old) seeking dental care between July 2023 and February 2024 at SH9HPDC. The external dataset II was collected from adult patients seeking dental care at four 4 different oral healthcare centers involving Hong Kong University Faculty of Dentistry (HKUFOD), The Sapienza University of Rome in Italy (SUR), Shanghai Ninth People's Hospital Huangpu Clinic (SH9HHPC), and SH9HPDC. This study is reported according to the checklist of artificial intelligence in dental research[26] and the STARD guidelines[27].

The reported diagnostic trials were conducted in accordance with the Declaration of Helsinki and were approved by the Institutional Review Board of the Shanghai Ninth People's Hospital (SH9H-2021-T408-3 and SH9H-2023-T369-1). The trial using external dataset I was obtained from a previous study and registered at ClinicalTrials.gov (NCT05513599) on 08/23/2022. The multicenter trial comprising external dataset II was registered at ClinicalTrials.gov (NCT06306677) on 03/12/2024. For the latter study, the IRB waived the need for informed consent for this retrospective analysis of anonymized radiographic images, as the research presented no risk to participants. For the multi-center radiographic images used in this study, the local ethics committees of all participating institutions granted approval for the use of their de-identified data under the protocol approved by the lead center (Shanghai Ninth People's Hospital).

The OPG images from the internal dataset had a resolution of 2903 × 1536, while the external dataset I images had a resolution of 1935 × 1024. The resolution of OPG images in the external dataset II are: HKUFOD (resolution: 2903 × 1536), SUR (resolution: 2864 × 1504), SH9HHPC (resolution: 1935 × 1024), and SH9HPDC (resolution: 1935 × 1024). For model testing, all images were resized to 1024 × 512.

Sample size calculation

There are no comparable screening and diagnostic studies for evaluating an AI-based imaging system for detecting periodontitis stages II-IV in OPG images. Based on the study hypothesis, the sample size was estimated using the confidence interval method, assuming an actual Area Under the Receiver Operating Characteristics Curve (AUROC) of 0.9 and a 2:1 positive-to-negative test result ratio. This results in a required sample size of approximately 369 subjects in external datasets. The sample size calculation was performed using PASS 15.0.5 using the confidence intervals for the AUROC module (NCSS, LLC, USA).

Reference standards

External dataset I (Full-mouth periodontal examination). The OPG images in the external datasets were linked to the ground truth, which consisted of a full-mouth periodontal examination performed and interpreted by a calibrated examiner who obtained the standard case definition. A trained examiner (XYW) conducted a thorough periodontal assessment using a standardized probe (UNC-15, Hu Friedy, Chicago, USA), measuring probing pocket depth, bleeding on probing, and clinical attachment level (CAL) at six sites per tooth, excluding third molars. The definitions for periodontal health, gingivitis, and periodontitis were based on the 2017 World Workshop on the Classification of Periodontal and Peri-implant Diseases[28]. Periodontal health was defined by the absence of gingival inflammation (BOP < 10%) and no attachment loss due to periodontitis, while gingivitis was identified by gingival inflammation (BOP ≥ 10%) without attachment loss. Periodontitis was diagnosed when inter-dental CAL was present at two or more non-adjacent teeth, with the disease stage determined by the most affected tooth. Stage I was defined as having 1–2 mm of CAL, stage II as 3–4 mm, and stage III/IV as 5 mm or greater. For some analyses, periodontal health status was recoded into a binary outcome: periodontal health/gingivitis/stage I periodontitis (Negative), or stage II-IV periodontitis (Positive).

External dataset II (Radiographic evaluation). A substantial volume of OPG images in real-world settings lacks corresponding clinical examination information. To extend study validity, radiographic bone loss assessed on OPG image by a panel of international periodontal specialists was used as the reference standard for case classification in the external dataset II. The presence of marginal alveolar bone loss (RBL) was assessed with a modified Schei ruler[29] with red lines corresponding to 15% and 33% thresholds to assist in staging periodontitis based on the 2018 Classification[7]. Images were categorized into 3 groups: (i) Absence of marginal bone loss or RBL less than 15% of the root length (periodontal health, gingivitis, or stage I periodontitis); (ii) stage II periodontitis based on 15% ≤ RBL ≤ 33% of root length; (iii) stage III or IV periodontitis based on RBL > 33% of root length. Detectable RBL in at least two non-adjacent teeth was required to diagnose periodontitis stage II-IV. For analysis, stage II cases were combined with stage III/IV cases. Three professorial-level experts (LM, GP, MST) independently assessed OPG images using a Google Form that automatically recorded the time for the evaluation of each image. To evaluate the reproducibility of the individual raters, a second-round assessment by each expert was conducted 2 weeks later. In case of disagreement among panel members, the case was discussed in an online adjudication meeting among the experts. A duplicate adjudication was conducted 2 weeks after achieving agreement to assess panel reproducibility.

Network architecture

HC-Net is a hybrid classification network designed as a binary classifier for detecting stages II-IV of periodontitis in OPG images, as previously described by our team[20]. The framework consists of three main interconnected components (Tooth-level analysis branch, Global-level analysis branch, and Clinical knowledge-guided fusion strategy) that capture both local and global features while integrating clinical diagnostic principles (Fig. 2).

Tooth-level analysis branch. For tooth detection, we employ a heatmap-based approach, where tooth center points are located using a tooth detection network (1) :

$$H = f_d(I; \theta_d) \tag{1}$$

where I is the input OPG image, H is the resulting heatmap, and θ_d represents the detection network parameters. Then, we convert the detected heatmap to each tooth center c_i, and use a fixed size of bounding box (i.e., 64×64) to crop the target tooth, denoted as P_i. For tooth-level classification, each cropped tooth patch P_i is processed by a classification network (2) :

$$s_i = f_c(P_i; \theta_c) \tag{2}$$

where $s_i \in [0, 1]$ is the probability score indicating whether the tooth exhibits stage II-IV periodontitis, and θ_c represents the classification network parameters.

Global analysis branch. This branch processes the entire OPG image I through a DenseNet-161 backbone for feature extraction. Note that this backbone is pre-trained on a large unlabeled dataset. Then, these features are used for patient-level classification $P = f_b(I; \theta_b)$ (3), where $P \in [0, 1]$ represents global features and θ_b are the network parameters. Simultaneously, we generate a Classification Activation Map (CAM) M, highlighting regions associated with periodontitis. To supervise the CAM, we generate a distance map upon the panoramic X-ray image, based on Euclidean Distance Transform with areas of positive tooth masks.

Clinically knowledge-guided fusion strategy. This fusion strategy integrates tooth-level and global(patient)-level predictions following clinical diagnostic principles: (1) If all teeth have scores $s_i < 0.3$, the patient is classified as healthy; (2) If two or more non-adjacent teeth have scores $s_i > 0.7$, the patient is diagnosed with periodontitis; (3) For uncertain cases, we employ an adaptive noisy-OR gate[30] (4) :

$$q = 1 - \prod_{i=1}^{N} (1 - \alpha_i s_i) \tag{4}$$

Where q represents the fused probability, N is the total number of teeth, and α_i is a learnable attention weight derived from the global CAM. The final prediction $\hat{y}$ at the patient level is obtained by (5) :

$$\hat{y} = \lambda q + (1 - \lambda) p \tag{5}$$

Where, λ is a weighting parameter used to balance the contributions from the tooth-level and global(patient)-level predictions.

Network Training

The network is trained end-to-end using a comprehensive loss function (6) :

$$L = L_{dec} + L_{cls} + L_{pat} + L_{cam} + L_{gate} \tag{6}$$

where L_{dec} is the tooth detection loss (i.e., focus loss), L_{cls} and L_{pat} are the tooth-level and global(patient)-level classification losses (i.e., binary cross entropy), L_{cam} is the CAM regression loss (i.e., mean squared error), and L_{gate} is the adaptive noisy-OR loss (i.e., binary cross entropy).

The framework is trained end-to-end using a comprehensive loss function that combines tooth-level detection and classification losses, patient-level classification and CAM regression losses, and adaptive noisy-OR gate loss. Regarding implementation details, the network is trained with an Adam optimizer and a learning rate of 0.0003 through 300 training epochs. The framework was implemented using the PyTorch platform on an NVIDIA A100 GPU.

Network pre-training and fine-tuning

We generated positive pairs for each OPG image (anchor) through various data augmentation techniques, including intensity adjustments, contrast manipulation, random noise injection, and geometric transformations. These augmentations were carefully chosen to represent realistic variations in radiographic imaging while maintaining the essential diagnostic features. The network was trained to bring the representations of augmented views from the same image closer together in the feature space while separating representations from different images. In practice, we first applied random intensity scaling by uniformly sampling a factor between 0.8 and 1.2, which mimics varying exposure levels. Next, we introduced contrast manipulation by adjusting gamma values in a range of 0.8 to 1.5, effectively simulating subtle discrepancies. We also performed geometric transformations such as random rotation within ±10, and random scaling (up to 10% enlargement or reduction). These transformations were designed to maintain the global structure of the jaw and teeth while slightly altering local appearances, thereby requiring the model to learn invariant features pivotal for periodontal disease classification. For implementing the training, each input image was passed twice through the same augmentation pipeline to create the pair of views. The contrastive objective then used these positively matched image pairs to pull their latent representations closer together in the feature space. Meanwhile, images sourced from different anchors in the same training batch were treated as negatives, compelling the network to push their representations further apart. This combination of positively and negatively paired examples, embedded in a stochastic gradient descent routine, formed the foundation of the self-supervised contrastive learning phase. The contrastive learning function can be defined as (7) :

$$L_{contrastive} = -\sum_{i=1}^{N} \log \frac{exp(z_i \cdot z_i^{+}/\tau)}{\sum_{j=1}^{2N} 1_{[j\neq i]} exp(z_i \cdot z_j^{+}/\tau)} \tag{7}$$

Where z_i and z_i^+ are the normalized embeddings of two augmented views of the same image, τ is a temperature parameter used to control the sharpness of the distribution, N is the batch size, and $1_{[j\neq i]}$ is an indicator function.

After the self-supervised pre-training phase, we fine-tuned the network using labeled data from our internal dataset. This supervised fine-tuning phase allowed the network to adapt its learned representations specifically for periodontal disease classification while preserving the robust features learned during pre-training. The HC-Net+ was obtained after the parameters were frozen. HC-Net+ retained the underlying convolutional backbone of HC-Net but incorporated refinements inspired by the self-supervised pre-training process. The initial layers benefit from the weights learned during contrastive learning on large, unlabeled OPG image, capturing general radiographic features such as edges, textures, and common structures. Deeper in the network, additional context-aware layers are introduced. These can include dilated convolutions or attention modules that aggregate information from wider spatial regions, crucial for detecting subtle periodontal indicators that appear across broad areas of the panoramic image. In the final stage, the classification head departs from a simple global average pooling by combining both global average and global max pooling features, concatenating these representations before passing them to the fully connected layers for classification. This dual-pooling approach captures fine-grained local details alongside holistic structural patterns, resulting in improved discrimination between healthy and diseased cases.

Statistical analysis

The networks' performance was evaluated based on patient-level predictions. The AUROC, confusion matrix, sensitivity, specificity, accuracy, and corresponding 95% confidence intervals (CIs) were calculated to assess the network's performance. The cut-off value for a positive test was determined based on the distribution of the HC-Net prediction scores during internal testing, with 0.7 established as the cut-off value for all analyses. A prediction score above 0.7 was considered a positive test; otherwise, it was deemed negative. The 95% CIs were calculated using the bootstrap method with 1000 resamples[31]. The DeLong test was utilized to compare the AUROC differences, and the differences in sensitivity, specificity, and accuracy were analyzed using McNemar's or Chi-square tests. A p-value of less than 0.05 was considered statistically significant. All statistical analyses were performed using Python (version 3.8.19; Python Software Foundation, Wilmington, DE, USA) with the following key libraries: scikit-learn, NumPy, pandas, Matplotlib, and StatsModels (version 0.13.5).

Data availability

Original data are partly available for scientific collaboration from the authors upon reasonable request. The source code for HC-Net+ is available upon request for non-commercial purposes.

Received: 9 May 2025; Accepted: 9 October 2025;
Published online: 18 November 2025

References

1. Nascimento, G. G., Alves-Costa, S. & Romandini, M. Burden of severe periodontitis and edentulism in 2021, with projections up to 2050: The Global Burden of Disease 2021 study. *J. Periodontal Res.* **59**, 823–867 (2024).
2. Kassebaum, N. J. et al. Global, regional, and national prevalence, incidence, and disability-adjusted life years for oral conditions for 195 countries, 1990–2015: a systematic analysis for the global burden of diseases, injuries, and risk factors. *J. Dent. Res.* **96**, 380–387 (2017).
3. Ezzati, M. & Riboli, E. Can noncommunicable diseases be prevented? Lessons from studies of populations and individuals. *Science* **337**, 1482–1487 (2012).
4. Tonetti, M. S. et al. Treatment of periodontitis and endothelial function. *N. Engl. J. Med.* **356**, 911–920 (2007).
5. D'Aiuto, F. et al. Systemic effects of periodontitis treatment in patients with type 2 diabetes: a 12 month, single-centre, investigator-masked, randomised trial. *Lancet Diab. Endocrinol.* **6**, 954–965 (2018).
6. Tonetti, M. S., Jepsen, S., Jin, L. & Otomo-Corgel, J. Impact of the global burden of periodontal diseases on health, nutrition and wellbeing of mankind: A call for global action. *J. Clin. Periodontol.* **44**, 456–462 (2017).
7. Tonetti, M. S., Greenwell, H. & Kornman, K. S. Staging and grading of periodontitis: Framework and proposal of a new classification and case definition. *J. Clin. Periodontol.* **45**, S149–s161 (2018).
8. Organization, W. H. *Global oral health status report: towards universal health coverage for oral health by 2030*, (World Health Organization, 2022).
9. Jacobs, R., Fontenele, R. C., Lahoud, P., Shujaat, S. & Bornstein, M. M. Radiographic diagnosis of periodontal diseases - Current evidence versus innovations. *Periodontol 2000* **95**, 51–69 (2024).
10. Rushton, V. E., Horner, K. & Worthington, H. V. The quality of panoramic radiographs in a sample of general dental practices. *Br. Dent. J.* **186**, 630–633 (1999).
11. Meusburger, T., Wülk, A., Kessler, A., Heck, K., Hickel, R., Dujic, H. & Kühnisch, J. The Detection of Dental Pathologies on Periapical Radiographs-Results from a Reliability Study. *J Clin. Med* **12**, 2224 (2023).
12. Zhang, A., Critchley, S. & Monsour, P. A. Comparative adoption of cone beam computed tomography and panoramic radiography machines across Australia. *Aust. Dent. J.* **61**, 489–496 (2016).
13. Kim, J., Lee, H. S., Song, I. S. & Jung, K. H. DeNTNet: Deep Neural Transfer Network for the detection of periodontal bone loss using panoramic dental radiographs. *Sci. Rep.* **9**, 17615 (2019).
14. Guler Ayyildiz, B., Karakis, R., Terzioglu, B. & Ozdemir, D. Comparison of deep learning methods for the radiographic detection of patients with different periodontitis stages. *Dentomaxillofac. Radio.* **53**, 32–42 (2024).
15. Chang, H. J. et al. Deep learning hybrid method to automatically diagnose periodontal bone loss and stage Periodontitis. *Sci. Rep.* **10**, 7531 (2020).
16. Lee, C. T. et al. Use of the deep learning approach to measure alveolar bone level. *J. Clin. Periodontol.* **49**, 260–269 (2022).
17. Ertaş, K., Pence, I., Cesmeli, M. S. & Ay, Z. Y. Determination of the stage and grade of periodontitis according to the current classification of periodontal and peri-implant diseases and conditions (2018) using machine learning algorithms. *J. Periodontal Implant Sci.* **53**, 38–53 (2023).
18. Yu, H. et al. A cascading learning method with SegFormer for radiographic measurement of periodontal bone loss. *BMC Oral. Health* **24**, 325 (2024).
19. Jiang, L. et al. A two-stage deep learning architecture for radiographic staging of periodontal bone loss. *BMC Oral. Health* **22**, 106 (2022).
20. Mei, L. et al. Clinical knowledge-guided hybrid classification network for automatic periodontal disease diagnosis in X-ray image. *Med. Image Anal.* **99**, 103376 (2025).
21. Tonetti, M. S. & Sanz, M. Implementation of the new classification of periodontal diseases: Decision-making algorithms for clinical practice and education. *J. Clin. Periodontol.* **46**, 398–405 (2019).
22. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729-9738 (2020).
23. Kim, Y. et al. Comparative analysis of clinical image evaluation charts for panoramic radiography. *Oral. Radiol.* **40**, 520–529 (2024).

24. Machado, V., Proença, L., Morgado, M., Mendes, JJ. & Botelho, J. Accuracy of Panoramic Radiograph for Diagnosing Periodontitis Comparing to Clinical Examination. *J Clin Med.* **9**, 2313 (2020).
25. Douglass, C. W. et al. Clinical efficacy of dental radiography in the detection of dental caries and periodontal diseases. *Oral. Surg. Oral. Med Oral. Pathol.* **62**, 330–339 (1986).
26. Schwendicke, F. et al. Artificial intelligence in dental research: Checklist for authors, reviewers, readers. *J. Dent.* **107**, 103610 (2021).
27. Cohen, J. F. et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* **6**, e012799 (2016).
28. Papapanou, P. N. et al. Periodontitis: Consensus report of workgroup 2 of the 2017 World Workshop on the Classification of Periodontal and Peri-Implant Diseases and Conditions. *J. Periodontol.* **89**, S173–s182 (2018).
29. Bassiouny, M. A. & Grant, A. A. The accuracy of the Schei ruler: a laboratory investigation. *J. Periodontol.* **46**, 748–752 (1975).
30. Srinivas, S. A generalization of the noisy-or model. In *Uncertainty in Artificial Intelligence* 208–215 (Elsevier, 1993).
31. Efron, B. & Tibshirani, R. An introduction to the bootstrap. (Chapman and Hall, New York, 1993).

Acknowledgements

This work has been supported by the National Natural Science Foundation of China (NSFC grant No 82301157, 82441023, U23A20295, 62131015, 82394432, 6230012077), the China Ministry of Science and Technology (S20240085, STI2030-Major Projects-2022ZD0209000, STI2030-Major Projects-2022ZD0213100), the Key R&D Program of Guangdong Province, China (grant number 2023B0303040001), Capital's Funds for Health Improvement and Research (CFH2024-1-4101), and HPC Platform of ShanghaiTech University, and the European Research Group on Periodontology.

Author contributions

Yuan Li and Xie Yu contributed to the design and led the clinical trial implementation, clinical data curation, and analyses. Zhiming Cui and Lanzhuju Mei contributed to the design, led the implementation of the bioengineering work, and performed the AI work. Lorenzo Marini, George Pelekos, Wen Gu, Chunan Zhang, Xiaoyu Yu, Xinyu Wu, Xindi Wei, Leran Tao, Ke Deng, Andrea Pilloni, and Maurizio Tonetti contributed to data collection in the clinical trials. Dinggang Shen and Maurizio Tonetti conceived the work, led the design, supervised it, interpreted the data, and contributed to manuscript drafting. All authors revised the manuscript and took responsibility for it.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-02077-0.

Correspondence and requests for materials should be addressed to Dinggang Shen or Maurizio S. Tonetti.

Reprints and permissions information is available at http://www.nature.com/reprints

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025