

Artificial Intelligence for Automated Grading and Treatment Planning in Periodontitis

Yuanlong Li^{1*}

¹ School of Dentistry, Qilu medical University, Zibo, China

Abstract. Periodontitis is a widespread chronic inflammatory disease that continues to threaten oral health and contributes to systemic complications. Its diagnosis and grading largely depend on probing and radiographic assessment, yet these approaches vary across clinicians and lack precision in detecting early bone alterations. Deep learning has been introduced to address these shortcomings by automatically analysing dental images and extracting both global bone patterns and site-specific features relevant to disease severity. Encoder–decoder networks can delineate alveolar bone contours and periodontal pockets, while classification models combine these representations to generate reproducible grading outcomes. Compared with conventional methods, such systems offer more consistent evaluation of complex regions, reduce observer variability, and shorten the time required for clinical interpretation. Integration with structured reporting further facilitates incorporation into electronic health records, enabling routine use in follow-up and treatment planning. Remaining barriers include annotation inconsistency, equipment-related variability, and limited validation across centers. This review aims to synthesize current progress in deep learning–based grading of periodontitis, clarify unresolved challenges, and outline requirements for clinical adoption.

1 Introduction

Periodontitis is a chronic infectious disease that begins with gingival inflammation and progressively involves the alveolar bone, periodontal ligament, and cementum. It is one of the leading causes of tooth loss in adults. Epidemiological data indicate that the prevalence of periodontitis among adults aged 35–44 is as high as 20%–50%, and this proportion further increases among the elderly population [1]. In addition to posing a serious threat to oral health, periodontitis is closely associated with various systemic diseases, such as cardiovascular diseases, diabetes, and adverse pregnancy outcomes, making it a significant risk factor affecting overall health [2]. Currently, the diagnosis and staging of periodontitis primarily rely on clinical examinations (such as periodontal probing depth and bleeding indices) and imaging studies (such as periapical radiographs, panoramic radiographs, and cone-beam computed tomography, CBCT). The assessment of periodontal pocket depth, degree of alveolar bone loss, and extent of inflammation is crucial in determining whether non-surgical treatment or surgical intervention is required. Therefore, accurate staging of periodontitis not only directly influences treatment efficacy but also serves as a core component in evaluating patient prognosis and developing personalized treatment plans [3]. In complex cases or scenarios requiring precise staging, traditional methods alone have become insufficient to meet the clinical demands of modern periodontics. There is an

urgent need for more objective, efficient, and intelligent auxiliary tools.

In recent years, deep learning (DL), as an important branch of artificial intelligence (AI), has demonstrated significant potential in the field of medical image analysis. By constructing deep neural network models, particularly convolutional neural networks (CNNs), deep learning can automatically learn and extract multi-level, high-dimensional feature information from vast amounts of image data. This makes it especially suitable for processing complex and variable data types such as dental images [4]. In medical image segmentation and object recognition tasks, models such as U-Net and Mask R-CNN can accurately locate periodontal tissue boundaries, identify bone resorption areas, and quantify key indicators like periodontal pocket depth. By training on large datasets of annotated periodontal images, deep learning models can achieve automated judgment of different periodontitis stages, thereby improving diagnostic consistency and accuracy [5, 6]. Although deep learning has advanced periodontal image analysis, its clinical use is still limited. The shortage of large annotated datasets narrows model applicability, device and protocol variations introduce systematic bias, and inconsistent expert labels weaken training reliability. Focused methods, including automatic ROI cropping, tailored augmentation, and transfer learning from related medical domains, help address these issues but cannot replace multicentre datasets and external validation. This review aims to assess current deep learning approaches for automated grading of

* Corresponding author: 15911168688longna@gmail.com

periodontitis, comparing model accuracy, multimodal integration, and device robustness, while also examining challenges of small-sample learning, interpretability, and regulatory compliance.

2 Datasets and Image Preprocessing

The development of imaging resources for automated periodontitis grading has shifted from small single-center collections to larger multi-center standardized datasets. One study aggregated approximately 3,000 panoramic radiographs and over 1,000 CBCT sets, stored uniformly in DICOM format with de-identification, while recording key metadata such as voxel size and exposure parameters [7]. A common approach is to use panoramic radiographs for large-scale screening and preliminary grading, with CBCT reserved for three-dimensional quantification to capture buccal-lingual bone plates and furcation anatomy, providing a balance between accessibility and precision [8]. However, inconsistencies in imaging protocols and quality control thresholds across centers reduce the comparability of cross-domain evaluations.

For annotation and staging alignment, multiple radiologists assign labels (mild, moderate, severe) using unified standards and delineate structures such as periodontal pocket contours and bone resorption ranges with ITK-like tools. Some cohorts report intra-group correlation coefficients of around 0.87 and Kappa values near 0.82, suggesting moderate consistency [9]. Nonetheless, blurred cemento-enamel junctions and metallic artifacts often lead to variable tolerance for boundary errors across teams, highlighting the need for more precise operational definitions and curated example libraries.

To reduce device-related variability and artifact interference, preprocessing commonly includes grayscale normalization and N4 bias field correction [10]. Pretrained segmentation models are often employed to automatically locate regions of interest, which are subsequently cropped into 256×256 pixel views. Online data augmentation strategies such as small rotations, mirroring, noise injection, and contrast adjustment are used to improve robustness against variations in imaging angles and quality. Class imbalance between mild, moderate, and severe cases is partially addressed through oversampling and focal loss. However, general augmentation has limited capacity to correct cross-device domain shifts, making domain adaptation and style alignment essential for further progress [11].

3 Deep Learning Architectures

Automated grading is generally structured as a two-step process: segmentation followed by classification. Enhanced U-Net models integrate stronger encoders and attention gates to reduce background noise, while hybrid loss functions combining Dice and focal terms are applied to improve both boundary delineation and recognition of difficult cases [12]. Reported performance includes Dice scores of about 0.91 for

periodontal pockets and 0.88 for bone resorption, though improvements in pixel-level overlap do not always translate into higher grading accuracy, particularly when early lesions are small in extent.

Grading networks often employ a dual-branch design. The global branch captures dental arch curvature and overall bone density distribution through global pooling and shape descriptors, while the local branch extracts fixed-size patches from anatomical sites such as mesial, distal, buccal, and lingual surfaces, guided by segmentation masks. These features are fused through channel attention and linear transformations. To prevent overconfidence, training commonly incorporates label smoothing and temperature scaling. Cross-validation at the patient level has produced weighted F1 scores near 0.89, demonstrating close alignment with expert labels. An important practical finding is that allocating low-confidence predictions for manual review yields greater benefit than marginal gains in overall average performance [13].

End-to-end and stepwise strategies emphasize different priorities. End-to-end pipelines deliver faster throughput, suitable for batch screening and settings with limited resources, averaging about 1.8 seconds per case. Stepwise methods, with an average of 2.3 seconds, allow manual correction at the segmentation stage and provide stronger interpretability and clinical oversight. On the same dataset, end-to-end grading accuracy reached 84%, compared with 86% for stepwise approaches, largely due to the possibility of refining segmentation before classification. A more pragmatic solution involves adaptive routing, where high-confidence samples follow the rapid end-to-end path while low-confidence or complex furcation cases are redirected to the stepwise workflow with manual verification. Determining the superiority of any architecture should not rely solely on single-center closed-loop tests; external validation and performance audits in real-world settings remain essential.

4 Implementation and Validation in Periodontal Practice

Evaluations across several institutions demonstrated that interpretation time per case could be reduced from approximately 15 minutes to about 3 minutes, thereby alleviating outpatient congestion and reducing clinician workload. Yet, efficiency alone is not a sufficient endpoint. Follow-up over 6–12 months is needed to determine whether time savings lead to meaningful clinical benefits, such as improved tooth retention, reduced probing depth, stabilization of clinical attachment levels, and stronger patient compliance. For clinical integration, the system can be deployed as a plug-in within imaging workstations supporting DICOM and supplemented by a cross-platform desktop application accessible across hospital networks. Once images are uploaded, the software automatically detects the modality, applies tailored preprocessing, and runs the relevant models. Outputs include segmentation masks, lesion quantification, staging results, heatmaps, and overlay measurements. Reports follow structured

templates, incorporating periodontal pocket depth maps, bone resorption percentages, and calibrated confidence scores (0–100), which can be directly transferred into electronic health records and follow-up workflows. Subsequent studies should jointly evaluate process indicators and prognostic outcomes in the same patient cohort, using calibration and decision curve analysis to establish whether procedural gains translate into measurable health improvements [14].

Multicentre randomized and blinded evaluations provide stronger evidence for robustness across devices and populations. One trial reported approximately 89% agreement between the system and senior clinicians ($\text{Kappa} \approx 0.82$), with junior clinicians achieving roughly 31% higher accuracy when using the tool, particularly in anatomically complex regions such as furcations. Nevertheless, many published studies omit detailed disclosure of institutional composition and device distribution, raising concerns about selection bias and limiting generalizability. To ensure actionable evidence, reports should present not only overall accuracy and timing but also subgroup confusion matrices, Brier scores, calibration slopes, and decision curves, enabling clinicians to define threshold-based rules for triggering manual review.

Quality assurance and safety oversight require adherence to full-lifecycle standards for medical software. Key measures include real-time monitoring of model health, drift detection, and scheduled regression testing of critical cases. Threshold adjustments and pipeline modifications should follow formal change-control procedures, with major updates deployed incrementally and rollback options preserved. All alterations must be logged for compliance and auditability. For data security, local inference is preferred, while cross-center collaboration may adopt federated learning combined with secure aggregation and differential privacy to limit exposure of raw data. Reporting only efficiency or agreement improvements without documenting adverse event rates or cost-effectiveness provides insufficient justification for adoption in regulatory and reimbursement contexts [15].

5 Challenges and Future Directions

Key obstacles to reliable deployment include imaging heterogeneity and the propagation of annotation errors. Variability introduced by different manufacturers and reconstruction kernels can restrict segmentation and grading performance, with low-dose CBCT scans and metal artifacts posing particular challenges. More robust strategies combine pre-inference activation of metal artifact suppression and denoising tuned to device fingerprints and acquisition metadata, alongside domain-adaptive normalization or style alignment to mitigate cross-domain disparities. At the population level, reduced sensitivity in patients over 80 years of age or those with type 2 diabetes likely reflects anatomical and bone metabolic variations. Federated learning with parameter servers offers a pathway to collaborative model updates without direct data sharing, striking a balance between accuracy and communication cost.

Annotation practices also require tighter governance. Version-controlled standards with explicit tolerances should be implemented; for example, allowing ± 0.3 mm deviations for the cemento-enamel junction and bone crest, and applying arbitration procedures for ambiguous cases. Following each annotation round, recalculating ICC and Kappa provides quantitative evidence of consistency gains. Publicly available manuals and automated quality-control algorithms already provide a foundation, but progress depends on the development of reusable sample libraries and reproducible experiment checklists that ensure comparability across institutions.

Emerging research emphasizes two directions. Multimodal fusion integrates bone resorption volume and density with clinical indices such as probing depth, bleeding index, tooth mobility, and microbial profiles including 16S rRNA, raising recognition accuracy for severe cases to over 90 percent. Lightweight deployment applies quantization-aware training and knowledge distillation to compress models to only a few megabytes, allowing real-time use on chairside terminals while enabling interactive 3D reconstruction for surgical planning. The broad adoption of such approaches, however, depends on economic evaluations sensitive to local population structures and reimbursement frameworks. For these systems to enter routine care, three criteria must be met: data resources must cover diverse devices and patient groups, models must retain calibration and net benefit in external datasets, and system integration must ensure seamless interoperability with PACS and electronic health records under auditable workflows. Only when multicentre real-world trials confirm both clinical endpoints and cost-effectiveness can deep learning-based grading of periodontitis operate as a reliable tool in daily practice.

6 Conclusion

Periodontitis affects a large proportion of adults and remains the main reason for tooth loss in older populations, with documented associations to cardiovascular disease and diabetes. Clinical management requires not only identifying the presence of disease but also distinguishing its severity, since treatment for early bone changes differs markedly from interventions in advanced loss. Current grading still relies on probing indices and radiographs, which often miss subtle defects at the alveolar crest or underestimate bone resorption in furcation areas. Deep learning methods have been introduced to fill these gaps. Segmentation networks can outline alveolar bone contours and periodontal pockets with pixel-level precision, and classification models that integrate overall bone density with site-specific markers have produced staging results that align more closely with expert evaluation in complex regions. Early clinical tests suggest measurable benefits, such as shorter interpretation times and more consistent staging in multi-rooted teeth where novices usually struggle.

The evidence, however, is not without caveats. Most available datasets originate from single centers and narrow patient groups, limiting the spectrum of variation that models can learn from. Differences in CBCT protocols, reconstruction kernels, or device vendors introduce systematic biases that hinder transferability between clinics. Even expert annotations are inconsistent when cemento-enamel junctions are blurred or when metallic artifacts obscure bone margins, which undermines the reliability of training labels. These challenges illustrate why promising accuracy on internal datasets cannot be equated with readiness for clinical adoption. Solutions need to go beyond incremental tuning, focusing instead on shared repositories, standardized acquisition protocols, and reproducible annotation guidelines that are tested across independent cohorts.

Looking forward, several directions appear particularly pressing. The first is building large image resources with transparent annotation quality metrics, enabling researchers to benchmark models under comparable conditions. The second is integrating imaging analysis with clinical records, systemic risk factors, and microbial profiles, which could move grading from static classification toward personalized disease trajectories. The third is ensuring outputs are not black-box labels but interpretable evidence, such as visual overlays, calibrated confidence ranges, and quantified error margins that clinicians can interrogate. Only by demonstrating that these tools can improve patient retention of teeth, reduce unnecessary interventions, and function reliably across diverse healthcare settings will deep learning-based grading of periodontitis move from experimental reports to routine practice.

References

1. Tonetti MS, Jepsen S, Jin L, 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*. **44**, 456 (2017). <https://doi.org/10.1111/jcpe.12732>
2. Hajishengallis G, Chavakis T, et al. Local and systemic mechanisms linking periodontal disease and inflammatory comorbidities. *Nat Rev Immunol*. **21**, 426 (2021). <https://pubmed.ncbi.nlm.nih.gov/33510490/>
3. Caton JG, Armitage G, Berglundh T, et al. A new classification scheme for periodontal and peri-implant diseases and conditions - Introduction and key changes from the 1999 classification. *J Periodontol*. **70**, 89 (2018). <https://doi.org/10.1002/JPER.18-0157>
4. Herrera D, Sanz M, Shapira L, et al. Association between periodontal diseases and cardiovascular diseases, diabetes and respiratory diseases. *J Clin Periodontol*. **50**, 819 (2023). <https://doi.org/10.1111/jcpe.13807>
5. Heitz-Mayfield LJA. Conventional diagnostic criteria for periodontal diseases. *Periodontol* 2000. **95**, 10 (2024). <https://doi.org/10.1111/prd.12579>
6. Suh B, Yu H, Cha JK, et al. Explainable deep learning approaches for risk screening of periodontitis. *J Dent Res*. **104**, 45 (2025). <https://doi.org/10.1177/00220345241286488>
7. Chang HJ, Lee SJ, Yong TH, et al. Deep learning hybrid method to automatically diagnose periodontal bone loss and stage periodontitis. *Sci Rep*. **10**, 64509 (2020). <https://doi.org/10.1038/s41598-020-64509-z>
8. Krois J, Ekert T, Meinhold L, et al. Deep learning for the radiographic detection of periodontal bone loss. *Sci Rep*. **9**, 44839 (2019). <https://doi.org/10.1038/s41598-019-44839-3>
9. Li X, Zhao D, Xie J, et al. Deep learning for classifying the stages of periodontitis on dental images: a systematic review and meta-analysis. *BMC Oral Health*. **23**, 1017 (2023). <https://doi.org/10.1186/s12903-023-03751-z>
10. Xue T, Chen L, Sun Q. Deep learning method to automatically diagnose periodontal bone loss and periodontitis stage in dental panoramic radiograph. *J Dent*. **150**, 105373 (2024). <https://doi.org/10.1016/j.jdent.2024.105373>
11. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data*. **6**, 60 (2019). <https://doi.org/10.1186/s40537-019-0197-0>
12. Azad R, Aghdam EK, Rauland A, et al. Medical image segmentation review: The success of U-Net. *IEEE Trans Pattern Anal Mach Intell*. **46**, 10076 (2024). <https://doi.org/10.1109/TPAMI.2024.3435571>
13. Zhang J, Deng S, Zou T, et al. Artificial intelligence models for periodontitis classification: A systematic review. *J Dent*. **156**, 105690 (2025). <https://doi.org/10.1016/j.jdent.2025.105690>
14. Chang J, Chang MF, Angelov N, et al. Application of deep machine learning for the radiographic diagnosis of periodontitis. *Clin Oral Investig*. **26**, 6629 (2022). <https://doi.org/10.1007/s00784-022-04617-4>
15. Roy R, Chopra A, Karmakar S, et al. Applications of artificial intelligence for diagnosis of periodontal and peri-implant diseases: a narrative review. *J Oral Rehabil*. **52**, 1193 (2025). <https://doi.org/10.1111/joor.14045>