JURNAL INFOTEL
VOL. 17, NO. 4, NOVEMBER 2025, PP. 892–919. DOI:10.20895/INFOTEL.V17I4.1401

RESEARCH ARTICLE

Deep Learning for Periodontitis Diagnosis on Two Dimensional Dental Radiograph: A Systematic Review

Jordan Valentino Lomanto[1] and Monica Widiasri[2,*]

[1,2] Universitas Surabaya, Indonesia

*Corresponding email: monica@staff.ubaya.ac.id

Received: June 30, 2025; Revised: November 13, 2025; Accepted: November 24, 2025.

Abstract: Periodontitis is an inflammatory disease that affects the supporting structures of the teeth and is a major contributor to tooth loss. Traditional diagnosis through clinical examination and manual interpretation of two-dimensional (2D) dental radiographs is prone to variability and subjectivity. The emergence of deep learning (DL) has improved the way medical images are analyzed, including dental radiography. This study systematically reviews the existing literature that uses DL approaches for the diagnosis of periodontitis using two-dimensional (2D) dental radiographic images and evaluates their diagnostic performance compared to clinical evaluations. A systematic literature review (SLR) was conducted following the PRISMA 2020 protocol and guided by the PICO (Populations, Interventions, Comparisons, Outcomes) framework. Five major databases (Scopus, PubMed, Semantic Scholar, Web of Science, and ScienceDirect) were searched for relevant studies published between 2016 and 2025. A total of 27 studies (in 29 reports) were included based on eligibility criteria, covering classification, segmentation, or detection tasks using panoramic, periapical, or bitewing radiographs. Most DL models achieved excellent performance with classification accuracies often exceeding 80% and segmentation Dice coefficients greater than 0.88. Although some models outperformed clinicians, external validation and real-world deployment remain limited. In conclusion, this review shows the feasibility of DL approaches in the diagnosis of automated periodontitis using 2D radiographs, although challenges and limitations remain in standardization, robust validation, and integration into clinical workflows.

Keywords: Systematic Review, Deep Learning, Periodontitis Diagnosis, 2D Dental Radiograph, Computer Vision

 ISSN: 2085-3688; e-ISSN: 2460-0997

1 Introduction

Periodontitis is a chronic inflammatory disease that affects the supporting structures of teeth and remains one of the most prevalent oral health problems worldwide [1, 2]. It is one of the main causes of tooth loss in adults and has been shown to be associated with other health conditions such as diabetes, cardiovascular disease, and adverse pregnancy outcomes [3]. Therefore, a timely and accurate diagnosis of periodontitis is critical to effective treatment and prevention of further complications [4]. Traditionally, diagnosis relies on clinical examination, including probing depth and bleeding on probing, combined with visual assessment of 2D dental radiographic images, such as panoramic and periapical radiographs. However, the interpretation of these images is highly subjective, prone to inter- and intra-examiner variability, and limited by human perceptual constraints [5].

The introduction of artificial intelligence (AI), particularly deep learning (DL), has brought a major transformation to medical image analysis. DL models, especially convolutional neural networks (CNN), have shown strong capabilities to detect and classify complex patterns in various imaging modalities. In the field of dentistry, these models offer the potential to automate radiograph interpretation, minimize diagnostic errors, and support decision-making. Recent studies have explored the use of DL for analyzing 2D dental images in detecting conditions such as caries, oral lesions, and more recently periodontal diseases [6, 7]. Among these applications, the diagnosis of periodontitis remains particularly challenging due to its subtle and often overlapping radiographic features [8].

Despite increasing research in this area, the effectiveness of DL models in diagnosing periodontitis remains inconsistent due to variations in data set quality, imaging modalities, model architectures, and evaluation methods. Furthermore, new DL architectures, such as Vision Transformers (ViT) [9, 10], EfficientNet [11], MobileNet [12], YOLO [13, 14], along with their hybrid variants [15, 16], have created the urgency of evaluating their suitability for this task. Therefore, a complete and comprehensive systematic review of current evidence is needed to determine whether DL-based approaches can offer reliable and accurate solutions to identify periodontitis from 2D radiographic images.

The purpose of this review is to address the gap by evaluating studies that use DL models to diagnose periodontitis using 2D dental radiographic images, with or without comparison to the' manual evaluation of clinicians. Studies involving any DL architecture, single, combined or hybrid, used to detect, classify, or segment periodontitis from 2D radiographic images, including panoramic, periapical, or bitewing radiographs, were included in the scope of this review regardless of the origin or evaluation protocol of the dataset.

This systematic literature review (SLR) follows the PICO framework (Populations, Interventions, Comparisons, Outcomes) and adheres to the PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) guidelines to ensure methodological transparency and reproducibility [17, 18]. The primary research question that guides this review is "Does a deep learning approach prove effective in diagnosing periodontitis from 2D dental radiographic images compared to manual evaluation by clinicians?". The findings of this review are expected to benefit both researchers and clinicians by providing insight into current capabilities, methodological trends, existing gaps, and directions for future development and implementation in clinical workflows.

2 Methodology

2.1 Research Design

The review was guided by the PICO framework to clearly define the scope and focus of the study. The Population (P) included 2D dental radiographic images, including panoramic, periapical, and bitewing radiographs, which are commonly used in periodontal diagnosis. Intervention (I) consisted of DL approaches, including CNN, hybrid architecture, transfer learning models, and other neural architecture designed for image recognition and classification tasks. Comparison (C) refers to traditional or manual diagnostic methods conducted by trained dentists or periodontists, included where available, to assess the relative performance of diagnostics performed using a DL framework. The Outcome (O) focused on the feasibility and diagnostic performance of DL models in detecting periodontitis, measured by relevant metrics such as sensitivity, specificity, precision, and overall classification accuracy.

2.2 Information Sources

A systematic search was conducted on five electronic databases: Scopus, PubMed, Semantic Scholar, Web of Science, and ScienceDirect. These databases are recognized as credible and trusted academic sources with a wide coverage of relevant and up-to-date peer-reviewed literature. PubMed was included in particular for its strong focus on medical research, as well as its open-access availability. The search was limited to peer-reviewed articles published in English between January 2016 and May 2025 to relevantly capture recent advances in deep learning (DL) for dental imaging.

2.3 Search Strategy

The search strategy was built around four core conceptual areas: deep learning, 2D dental radiography, periodontitis, and image recognition. Keywords were adapted and refined using terms from existing systematic reviews, with modifications to include the latest terminologies in DL-driven diagnostics (Table 1). Boolean operators and database-specific filters were applied to refine the results and eliminate irrelevant records (Table 2).

Table 1: Search keywords

Core Concept	**Keywords**
Deep Learning	“Deep Learning” OR “Deep Neural Network” OR “Convolutional Neural Network” OR “Recurrent Neural Network” OR “CNN” OR “RNN” OR “YOLO” [19]
2D Dental Radiography	“Dental Radiography” OR “Dental Image” OR “Dental X-Ray” OR “Panoramic” OR ""eriapical” OR “Bitewing” [20]
Periodontitis	“Periodontitis” OR “Periodontal Disease” OR “Periodontal Bone Loss” OR “Alveolar Bone Loss” [21,22]
Image Recognition	“Detection” OR “Classification” OR “Recognition” OR “Segmentation”

2.4 Eligibility Criteria

Inclusion criteria were defined to ensure the relevance and quality of the selected studies. Articles were included if they: (1) applied deep learning techniques to analyze 2D dental radiographic images; (2) aimed at detecting or diagnosing periodontitis; (3) reported on diagnostic performance using quantitative metrics; (4) focused on human subjects and clinical imaging; and (5) published in English in peer-reviewed journals or reputable conference proceedings. Studies were excluded if they: (1) employed 3D imaging modalities (e.g., CBCT or MRI); (2) did not utilize deep learning methods or provided insufficient model detail; (3) did not explicitly address the diagnosis of periodontitis (focused solely on other oral diseases); or (4) were reviews, editorials, or opinion articles.

2.5 Selection Process

The study selection process was conducted in four distinct phases according to the PRISMA 2020 guidelines: identification, screening, eligibility, and inclusion [18]. The PRISMA standardized checklist ensures a transparent and traceable flow of the literature selection process by specifying which databases and keywords were used, what time range was covered, how many studies were found, screened and excluded along with the reasons, thus ensuring reproducibility in future research.

Initially, records were identified by applying the predefined search strategy in five electronic databases, as mentioned in the Information Sources section. All retrieved citations were imported into a reference management system, and duplicates were systematically removed in the identification phase. In the selection phase, the titles and abstracts of the remaining records were reviewed to assess their relevance with the inclusion and exclusion criteria. Studies that did not meet the criteria were excluded. The Full-text versions of the articles that met the criteria were then retrieved for eligibility assessment. Each article was evaluated to determine whether it met all inclusion criteria, was relevant to the research question, applied DL approaches on 2D dental radiographs, and provided quantitative diagnostic performance outcomes. Finally, studies that met all all criteria were included in the qualitative synthesis. The entire selection process was documented and presented using a PRISMA flow diagram in the Results section.

Table 2: Search queries (April 2025)

Database	**Queries (April 2025)**	**n**
Scopus	(“Deep Learning” OR “Deep Neural Network” OR “Convolutional Neural Network” OR “Recurrent Neural Network” OR “CNN” OR “RNN” OR “YOLO”) AND (“Dental Radiography” OR “Dental Image” OR “Dental X-Ray” OR “Panoramic” OR “Periapical” OR “Bitewing”) AND (“Periodontitis” OR “Periodontal Disease” OR “Periodontal Bone Loss” OR “Alveolar Bone Loss”) AND (“Detection” OR “Classification” OR “Recognition” OR “Segmentation”)	100

Database	Queries (April 2025)	n
PubMed	(“Deep Learning” OR “Deep Neural Network” OR “Convolutional Neural Network” OR “Recurrent Neural Network” OR “CNN” OR “RNN” OR “YOLO”) AND (“Dental Radiography” OR “Dental Image” OR “Dental X-Ray” OR “Panoramic” OR “Periapical” OR “Bitewing”) AND (“Periodontitis” OR “Periodontal Disease” OR “Periodontal Bone Loss” OR “Alveolar Bone Loss”) AND (“Detection” OR “Classification” OR “Recognition” OR “Segmentation”)	58
Semantic Scholar	(“Deep Learning” OR “Deep Neural Network” OR “Convolutional Neural Network” OR “Recurrent Neural Network” OR “CNN” OR “RNN” OR “YOLO”) AND (“Dental Radiography” OR “Dental Image” OR “Dental X-Ray” OR “Panoramic” OR “Periapical” OR “Bitewing”) AND (“Periodontitis” OR “Periodontal Disease” OR “Periodontal Bone Loss” OR “Alveolar Bone Loss”) AND (“Detection” OR “Classification” OR “Recognition” OR “Segmentation”)	73
Web of Science	(“Deep Learning” OR “Deep Neural Network” OR “Convolutional Neural Network” OR “Recurrent Neural Network” OR “CNN” OR “RNN” OR “YOLO”) AND (“Dental Radiography” OR “Dental Image” OR “Dental X-Ray” OR “Panoramic” OR “Periapical” OR “Bitewing”) AND (“Periodontitis” OR “Periodontal Disease” OR “Periodontal Bone Loss” OR “Alveolar Bone Loss”) AND (“Detection” OR “Classification” OR “Recognition” OR “Segmentation”)	92
Science Direct	(“Deep Learning” OR “Neural Network”) AND (“Panoramic” OR “Periapical” OR “Bitewing”) AND (“Periodontitis” OR “Bone Loss”) AND (“Detection” OR “Classification”)	108

2.6 Data Extraction

Data extraction was conducted systematically using a predefined extraction framework to ensure consistency and reproducibility in all included studies. The template was developed based on systematic reviews involving artificial intelligence (AI) applications in medical imaging.

Bibliographical data captured included the study ID (including citation and authorship), year of publication, and the country in which the research was conducted. To evaluate data provenance, data availability and data quality, data sources were noted and classified by origin, such as universities, hospitals, university hospitals, clinics, dental schools,

public datasets, or if no mention was made at all. For imaging modalities, the type of image was recorded as panoramic, periapical, or bitewing radiographs. The sample size of the data set was also extracted to assess the scale and robustness of the experiments, along with the data split ratio used to separate the datasets into training, validation, and testing subsets. Additionally, data augmentation techniques, such as image rotation, flipping, contrast adjustment, or noise injection, were also noted to understand each study's strategy in addressing data availability limitations.

Detailed information on each study's DL model architecture was also recorded, such as CNN, ResNet, U-Net, and other custom-designed frameworks. Associated computer vision tasks for diagnosis, whether classification, object detection, or segmentation, were identified to capture the purpose of the DL application. Comparisons were made between the output of the DL model and the evaluations of the clinicians, where applicable, to assess the clinical relevance of the real-world. In addition, information on model validation methods, such as k-fold cross-validation, hold-out split, or external validation, was also extracted to assess the reproducibility of the research. Finally, the performance metrics of each study were extracted, such as accuracy, sensitivity, specificity, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC), depending on the type of task.

The data collected were then analyzed and organized into summary tables. This allows for comparisons and identifications of trends, patterns, usages, and approaches across included studies.

2.7 Performance Evaluation Metrics

The models' performance in DL-related studies commonly evaluated using standard quantitative metrics such as accuracy (1), precision (or positive predictive value, PPV) (2), recall (or sensitivity) (3), specificity (4), negative predictive value (NPV) (5), and F1-Score (6). These metrics are calculated using the derivative form of the confusion matrix. Four components in performance metrics, true positive (TP), false positive (FP), true negative (TN), and false negative (FN), are used as a basis for more meaningful metrics. A prediction is considered a TP when the model correctly identifies a positive case as being positive, whereas FP occurs when the actual value is negative, resulting in a false positive prediction. Similarly, TN and FN follow the same logic as their positive counterparts but apply when the actual value is negative.

Accuracy measures how close the overall predictions of the model are to being correct. Precision measures the proportion of correctly identified positives, which is useful when the cost of false positive is high (e.g., falsely predicting a healthy tooth as having periodontitis introduces unnecessary cost of further treatments). While recall compares correctly predicted positives to all actual positives, specificity compares correctly predicted negatives to all actual negatives. Prioritizing recall in early screening tasks can be advantageous as missed positive cases could lead to more severe consequences (e.g. missing a decaying tooth may allow disease progression leading to other complications, so it is favorable to treat uncertain cases as potential positives). In confirmatory cases, where false alarms could be costly (e.g., false diagnosis of oral cancer causes unnecessary stress to the patient), specificity plays an important role. NPV calculates correctly predicted negatives in proportion to all predicted negatives, and high NPV indicates strong confidence in negative predictions. To help balance between maximizing true detection (recall) and minimizing

overcalling positives (precision) in cases where missed diagnosis and overtreatment can be expensive, the F1-Score is a valuable metric worth considering.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4}$$

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{5}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

For segmentation tasks, the quality of the delineation can also be measured using the dice similarity coefficient (DSC) (7), the Jaccard similarity index (or intersection over union, IoU) (8) and the pixel accuracy (PA) (9). The DSC can be used as a metric to indicate how strong the agreement between the predicted annotation and the ground truth is. With correctly detected areas receiving more weight, DSC can be sensitive to small irregular structures, which can be a valuable metric for cases such as subtle lesions detection. Unlike DSC, which favors only the correct predictions, IoU also applies strict penalty for extra areas that do not actually belong to the original object of interest. IoU is often preferred when the precise localization is crucial because both under- and over-segmentation contributes to lower value. However, PA, like the name itself, measures the proportion of correctly classified pixels to all pixels in an image. Although PA can be a reflection of the overall segmentation correctness of the model, it can be biased when the image contains large background areas, which are easy to classify. Therefore, PA is usually interpreted alongside other metrics like DSC or IoU.

$$\text{DSC} = 2 \times \frac{|\text{Prediction} \cap \text{Ground Truth}|}{|\text{Prediction}| + |\text{Ground Truth}|} \tag{7}$$

$$\text{IoU} = \frac{|\text{Prediction} \cap \text{Ground Truth}|}{|\text{Prediction} \cup \text{Ground Truth}|} \tag{8}$$

$$\text{PA} = \frac{\text{Correctly predicted pixels}}{\text{Total pixels}} \tag{9}$$

3 Results

3.1 Study Selection

A total of 431 records were initially identified through comprehensive searches in five academic databases: Scopus (100), PubMed (58), Semantic Scholar (73), Web of Science (92) and

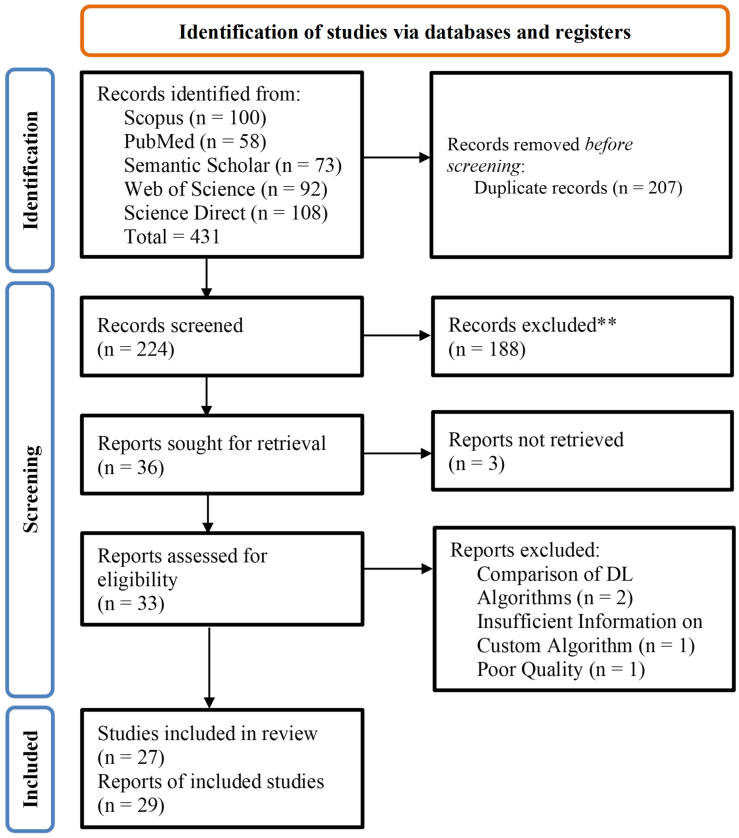

Figure 1: PRISMA flow diagram.

ScienceDirect (108). After the removal of 207 duplicates, 224 unique records remained for screening. During the screening phase, 188 records were excluded based on the evaluation of the title and abstract because they did not meet the predefined inclusion criteria.

Subsequently, the full texts of 36 articles were assessed for eligibility. 3 full-text articles could not be retrieved despite attempts to contact the corresponding authors. Of the retrieved studies, 4 were excluded during eligibility assessment, 2 studies focused only on comparative analysis between deep learning (DL) algorithms without addressing diagnostic feasibility, 1 study was excluded due to poor methodological quality with the dataset consisting only of 40 images, and 1 study used a very customized hourglass network architecture without adequate generalizability.

As a result, 27 studies reported across 29 reports were included in the final synthesis. Two studies were reported in multiple publications and were excluded, hence the difference between the number of studies and reports (Table 4). The detailed process of identification, screening, eligibility, and inclusion is presented in the PRISMA 2020 flow diagram Figure 1.

3.2 General Characteristics

The selected studies were published between 2016 and 2025, with an increase in studies published after 2020, indicating a growing interest in combining medical imaging and artificial intelligence (AI). Geographically, studies were conducted in different countries. Most originated from East Asia, like South Korea and China, followed by the United States, European regions such as Spain and the United Kingdom, and Middle Eastern regions, including Saudi Arabia (Table 4). This distribution highlights a global recognition of the potential for AI integration in dental diagnostics, although variations in imaging standards, clinical practices, and data accessibility may affect study outcomes.

3.3 Imaging Modalities

The studies included in this review utilized a variety of 2D dental radiographic modalities, with panoramic radiographs the most widely used, followed by periapical and bitewing radiographs. Several studies focused only on one type of image, while a few combined multiple modalities (Table 4). The choice of radiograph type was based on the diagnostic objective. Each modality has different advantages in the visualization of periodontal structures and contributes differently to the development and performance of diagnostic algorithms.

Panoramic radiographs (Figure 2 (c)) provide a broad overview of the entire jaw in a single image. They are commonly used in early screening because of their wide field of view, allowing the detection of bone changes throughout the mouth. However, they are prone to certain limitations, such as image distortion, overlapping anatomical features, and lower resolution, thus compromising their precision. Periapical radiographs (Figure 2 (b)), on the other hand, offer focused, high-resolution views of individual teeth and surrounding bone structures. They are often used in studies focusing more on the precise detection of periodontal defects such as vertical bone loss and FI. The clarity and precision make them suitable for segmentation tasks in DL, where accurate identifications of small changes are essential. Bitewing radiographs (Figure 2(a)), although less frequently used in the reviewed studies, also serve as an important modality for early detection of bone loss and interproximal defects. They are commonly used in routine dental checks and are typically tasked with classifications focused on posterior regions. Although their coverage is limited compared to panoramic or periapical views, bitewings provide high-resolution images in interproximal spaces, making them valuable for focused assessments and, thus, could complement other modalities when integrated into multimodal diagnostic models.

3.4 Dataset Attributes

The sample sizes between studies varied greatly, ranging from fewer than 500 to over several thousand image samples, demonstrating the difference in data resources and accessibility. Many studies addressed limitations in dataset size by adding data to improve generalizability and prevent overfitting. Dataset sources were acquired primarily from university hospitals, dental schools, and clinical institutions (Table 4). Some studies used publicly available datasets, while some studies did not clearly specify their data source, which may affect the reproducibility and comparability of results found in the literature.

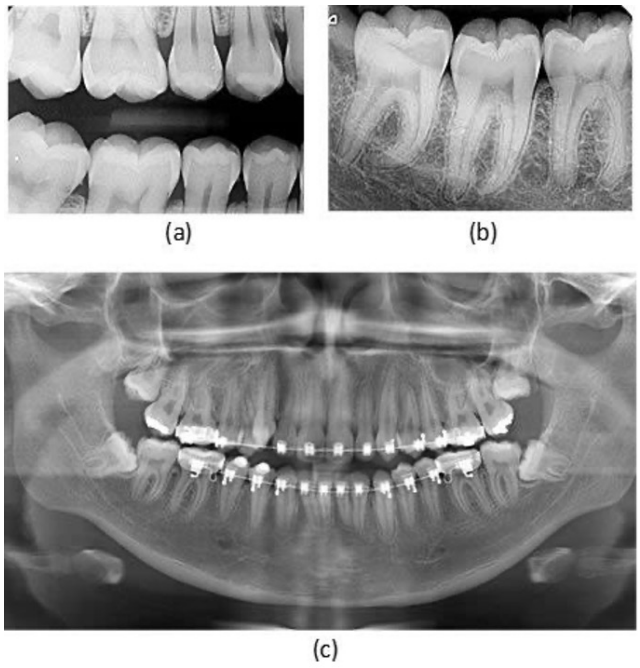

Figure 2: (a) Bitewing, (b) Periapical, (c) Panoramic Radiograph [23].

3.5 Deep Learning Models

The studies included in this review utilized a variety of DL architectures, as summarized in Table 3, to perform diagnostic tasks on 2D dental radiographs. Across the 27 studies, the most frequently applied model was from the CNN (Convolutional Neural Network) family, appearing in 8 out of 27 reviewed studies, either in the base form or in more specialized architectures such as VGG (Visual Geometry Group), ResNet (Residual Network), Inception, and customized CNN variants. With the appearance in also 8 studies, YOLO (You Only Look Once) architectures (v4-v9) were also equally popular in studies involving periodontitis cases because of their evolving capabilities from pure object detection to include also segmentation since YOLOv5. For segmentations, U-Net was the most used model, appearing in 7 studies. Another segmentation architecture, Mask R-CNN, belongs to the R-CNN (Region-based CNN) family, was reported in 5 studies, while Faster R-CNN and Keypoint R-CNN, both for object detection, were found in 3 studies. Emerging transformer-based models, such as SegFormer and Vision Transformer (ViT), were only used in 2 studies. Additionally, hybrid or ensemble approaches combining CNN with traditional machine learning (ML) classifiers or multi-model pipelines appeared in 4 studies.

CNN consists of components including convolutional layers, pooling layers, and fully connected layers (Figure 3). Convolutional layers apply filters to extract features, pooling layers reduce spatial dimensions, and fully connected layer interpret the extracted features. These models were used primarily for classification tasks, including binary and multiclass classifications. CNN-based classifiers have been reported to have reliable performance with accuracy scores around 80% [24, 25].

Table 3: Model appearance in studies

Model Category	Models	Number of Studies
CNN-based	CNN, VGG-16, ResNet-18, AlexNet, GoogLeNet, Inception v3, VGG19, HYNETS, PDCNN, PAR-CNN	8
YOLO	YOLOv4, YOLOv5, YOLOv8, YOLOv9	8
U-Net	U-Net	7
Object Segmentation	Mask R-CNN	5
Object Detection	Faster R-CNN, Keypoint R-CNN	3
Transformer-based	SegFormer, ViT	2
Hybrid / Ensemble	CNN + traditional classifiers (RF/SVM/NB/LR/KNN), Mask R-CNN + XGBoost, Mask R-CNN + U-Net	4

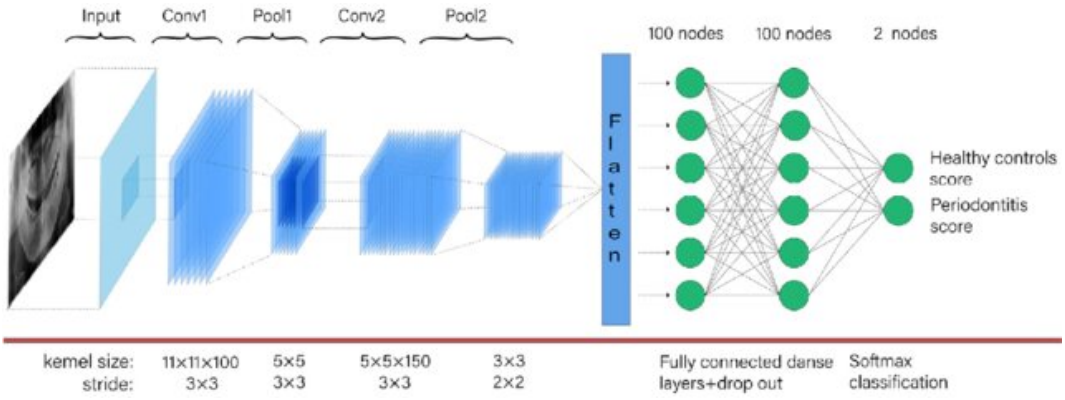

Figure 3: CNN architecture visualization [24].

VGG-16 (Visual Geometry Group with 16 layers), one of the variants of CNN, was frequently implemented for both classification and detection. It consists of 16 layers, including 13 convolutional layers and 3 fully connected layers, with small 3×3 filters to capture complex patterns while still maintaining a manageable number of parameters (Figure 4). In studies in which the main task was binary classification (differentiating between healthy and diseased teeth), VGG-16 was reported to have a satisfactory precision of approximately 73%, supported with moderate agreement with periodontists. However, performance declined to around 59% when dealing with multiclass classifications such as periodontitis severity classification (normal, mild, moderate, severe) [1].

ResNet (Residual Network) is another specialized architecture of CNN that supports deeper network stacks through residual connections, some commonly used versions being ResNet-18, ResNet-50 (Figure 5), and ResNet-101. Introduce skip connections that allow the input to bypass layers and be added directly to the output (Figure 6), which helps preserve gradients during backpropagation, making it possible to train networks with hundreds of layers. These models performed very well in classification tasks, with trade-offs

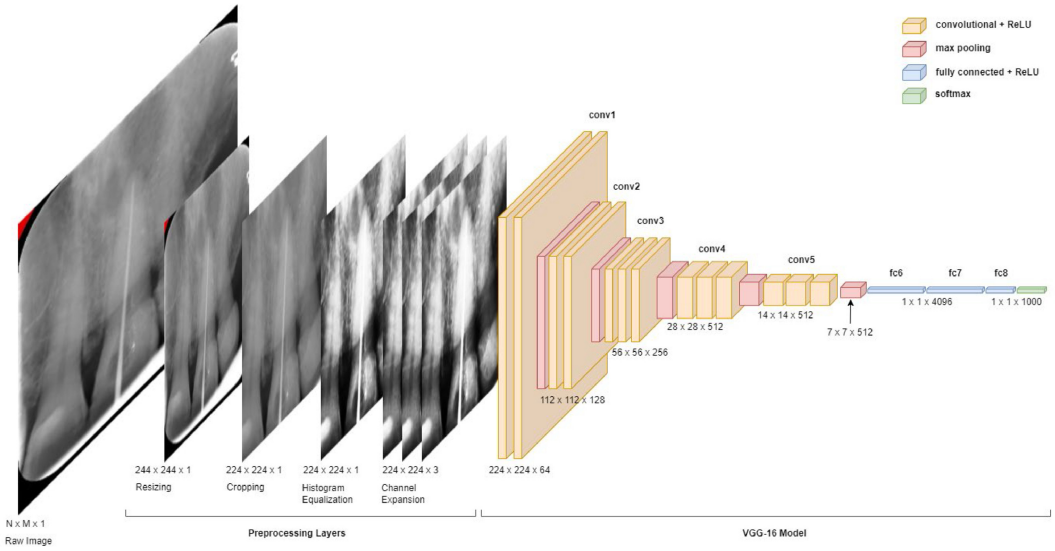

Figure 4: VGG-16 architecture visualization [26].

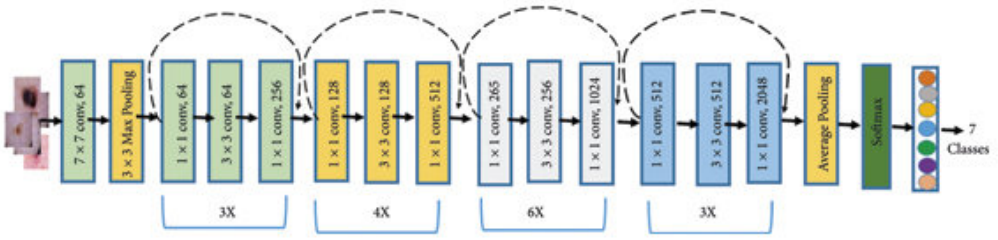

Figure 5: ResNet-50 architecture visualization [28].

between computational efficiency and accuracy. ResNet-18, pre-trained on ImageNet, has been reported to have yielded an accuracy, sensitivity, and specificity of more than 95% and achieved AUC values greater than 0.98 in binary classification tasks [27]. These numbers show that even the shallower ResNet model demonstrates high accuracy in detecting and staging periodontitis when combined and fine-tuned with transfer learning.

For segmentation tasks, U-Net has been featured in numerous studies that focus on the pixel-level localization of periodontal structures. It is made of a symmetric encoder-decoder structure (Figure 7), in which the encoder captures the context through series of convolution and pooling layers, while the decoder enables localization using up sampling and convolution layers. The U-Net-based models were able to outperform other architectures in segmentation tasks, achieving Dice similarity coefficients above 0.91 and the

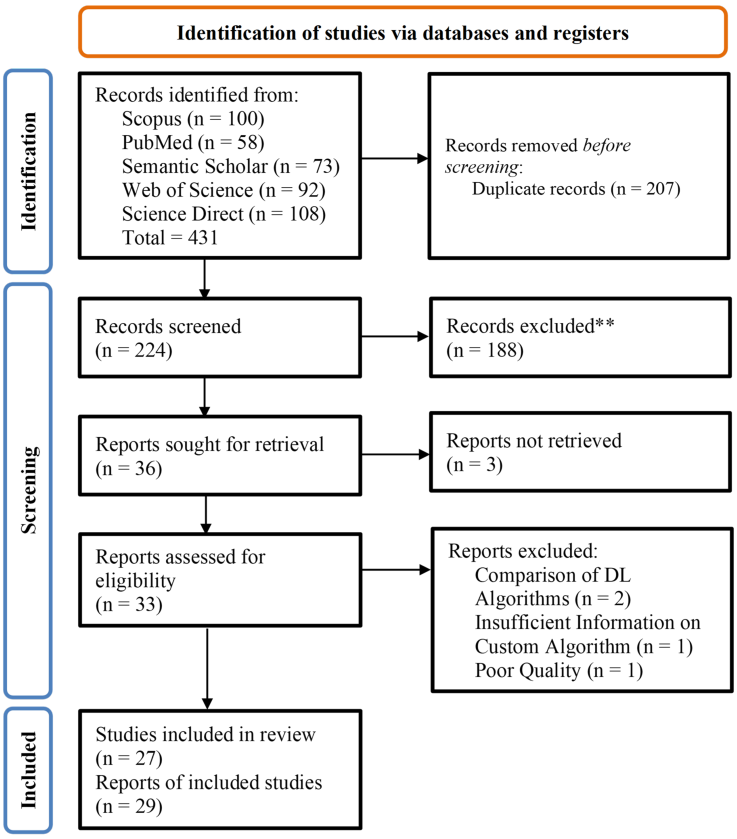

Figure 6: Skip connection in ResNet.

Jaccard index above 0.87 [29,30], making them ideal for highlighting bone contours at the pixel-level.

YOLO (You Only Look Once) was another commonly used architecture in studies with object detection as its primary task. YOLO differs from traditional two-stage architectures, like R-CNN (Region-based CNN) or Faster R-CNN, by treating detection as a single regression problem, which allows it to process an entire image in one pass. Divide the input image into grids and simultaneously predict multiple bounding boxes and class probabilities for each grid cell. Like ResNet, YOLO also has different versions developed over time, some commonly used versions found in the studies were YOLOv5 and YOLOv8 (Figure 8). With segmentation ability supported since YOLOv5, models that incorporate the architecture were able to demonstrate high localization precision, reflected by the mean mean average precision (mAP) values between 0.85 and 0.95 [5,30]. However, these models also have trade-offs between speed and accuracy, thus generally less effective in multiclass classification tasks compared to a two-stage detector like ResNet or VGG but suitable for studies that emphasize clinical applicability and speed.

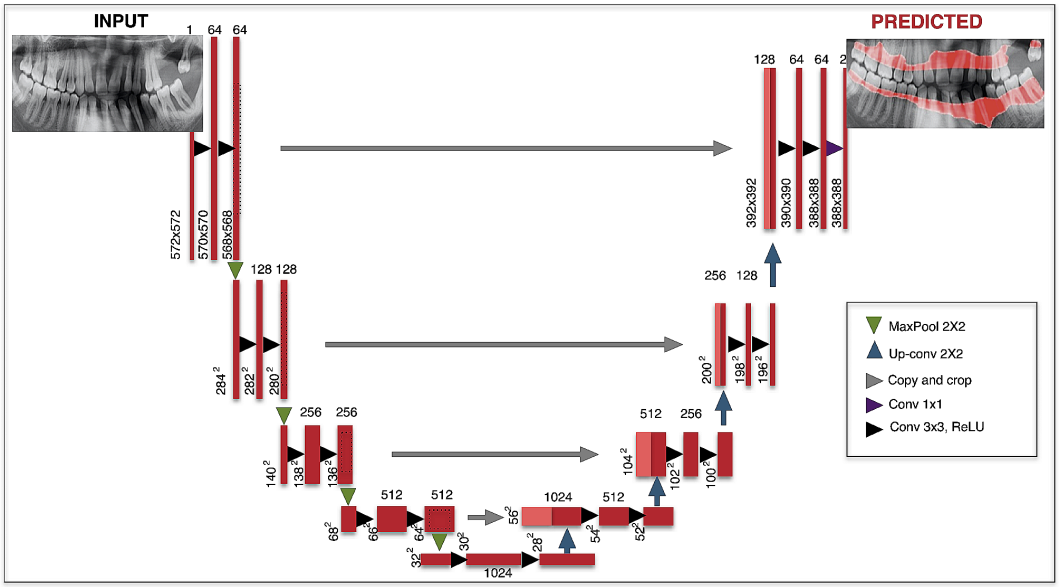

Figure 7: U-Net architecture visualization [31].

Some studies also implemented Mask R-CNN, another advanced two-stage framework that combines object detection with pixel-level segmentation. It builds on Faster R-CNN with the addition of a parallel branch to predict segmentation masks (Figure 9). The architecture uses a CNN backbone, such as ResNet, to extract features, a region proposal network (RPN) to generate object proposals, and a RoIAlign, which extracts features from each region of interest (RoI). In the reviewed studies, Mask R-CNN achieved strong segmentation performance, with reported Dice scores of up to 0.88 [32]. Its ability to perform instance segmentation makes it particularly well-suited for applications where multiple anatomical structures must be independently assessed.

Several studies developed custom or hybrid architectures, sometimes integrating multiple subnets or combining CNN backbones with attention mechanisms or custom heads. Despite the high performance, exceeding 80% in accuracy [33], they frequently lacked transparent implementation details, which makes it harder to reproduce.

Transfer learning was a strategy commonly used across various models, especially when working with relatively small datasets. This approach is also often found in studies that leveraged pre-trained models such as ResNet and VGG. By using pre-trained weights, such as ImageNet, studies were able to enhance performance, particularly in cases involving limited training data or class imbalance.

Which model to be used should consider diagnostic task requirements: CNN and ResNet for classification, U-Net and Mask R-CNN for segmentation, YOLO and Faster R-CNN for detection, and VGG-16 for hybrid use cases. Despite the capabilities and growing applications of DL in dental radiology, standardized benchmarks and open-source validations are still needed to ensure comparability between studies.

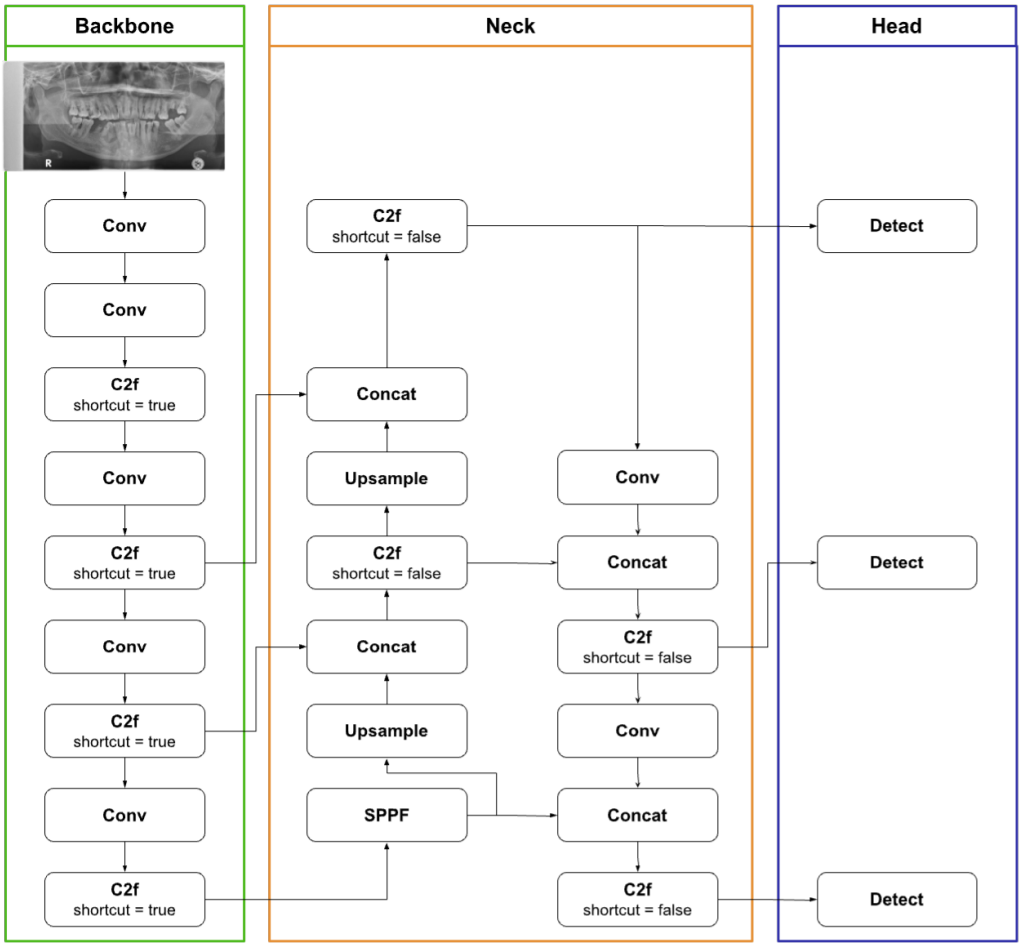

Figure 8: YOLOv8 architecture visualization.

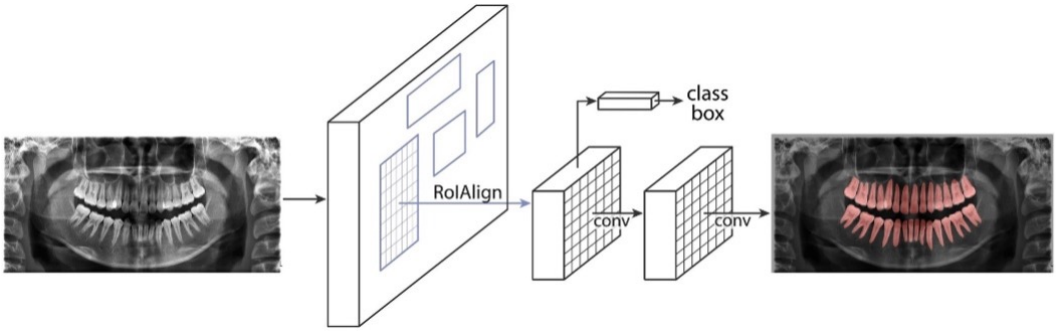

Figure 9: Mask R-CNN architecture visualization [33].

3.6 Diagnostic Tasks

Diagnostic objectives in all studies were categorized into three computer vision tasks, namely classification, segmentation, and object detection (Table 4). Classification tasks, which are the most frequently observed, focused on identifying the presence or severity of periodontitis. These models typically reported performance using metrics such as accuracy, precision (or positive predictive value, PPV), recall (or sensitivity), specificity, negative predictive value (NPV), F1-score, area under the receiver operating characteristic curve (AUC) and occasionally cross-entropy loss. Segmentation tasks focused mainly on defining pathological features, such as loss of alveolar bone. Studies performing segmentation frequently reported the Dice similarity coefficient (DSC), Jaccard index (or intersection over union, IoU), and pixel accuracy (PA) as key performance indicators. A smaller subset of studies implemented object detection models, such as YOLO variants, to localize specific regions of interest, such as periodontal pockets or vertical bone defects, within radiographic images. These models often used metrics such as average precision (AP), mean average precision (mAP), average recall (AR), mean average recall (mAR) and frame per second (FPS) for evaluation. Although several studies focused on a single vision task, others combined multiple approaches, such as segmentation followed by classification, to improve diagnostic performance.

3.7 Clinician Comparisons

Some studies compared the performance of DL models with the clinician-based diagnosis to assess the clinical applicability of the suggested systems. The diagnostic outputs of the models were evaluated against interpretations provided by general dentists or periodontists using the same radiographic datasets. The results generally indicated that the DL models were comparable and even superior to human experts in some cases related to the detection of periodontitis. Some studies used accuracy, sensitivity, and specificity comparisons, while others applied inter-agreement analyzes. However, comparative evaluations were not universally conducted, which therefore limits the generalizability of the findings.

Table 4: Selected studies

First Author (Country, Year)	**Datasets**	**DL Model**	**Task**	**Clinician Compare**	**Validation Method**	**Performance Metrics**
Alotaibi (Saudi Arabia, 2022) [1]	University Dataset (1724 periapical images; augmented; 70-20-10)	VGG-16	Detection	Yes	Hold-out validation	Accuracy, Precision / PPV, Recall / Sensitivity, F1 Score

First Author (Country, Year)	Datasets	DL Model	Task	Compare	Validation Method	Performance Metrics
Ameli (Italy, 2024) [34]	Private Clinical Dataset & Dental School (1000 periapical images; augmented; 80-10-10 for BL segmentation & 70-20-10 for Apex detection; additional 1582 images for test)	U-Net & YOLOv9	Segmentation & Classification	Yes	Hold-out validation	Accuracy, Precision / PPV, Recall / Sensitivity, F1 Score, Jaccard, mAP
Chang (Korea, 2020) [2]	University Hospital Dataset (340 panoramic images; augmented; 90-10)	Mask R-CNN & conventional CAD	Detection	Yes	Train-test split	Jaccard, Dice, PA
Chen (Taiwan, 2024) [35]	University Hospital Dataset (336 periapical images; augmented)	Mask R-CNN & U-Net	Detection	Yes	Hold-out validation	Accuracy, Precision / PPV, Recall / Sensitivity, F1 Score, Specificity, NPV, AUC
Dai (China, 2024) [36]	University Hospital Dataset (11120 periapical images; 60-15-25)	Alexnet / VGG16 / ResNet18 with RF / SVM / NB / LR / KNN	Classification	Yes	Hold-out validation	Accuracy, Recall / Sensitivity, Specificity, AUC
Erturk (Turkey, 2025) [13]	University Dataset (1752 bitewing images; augmented; 80-20)	YOLOv8	Classification	Yes	5-fold cross-validation	Accuracy, Precision / PPV, Recall / Sensitivity, F1 Score

First Author (Country, Year)	Datasets	DL Model	Task	Compare	Validation Method	Performance Metrics
Jiang (China, 2022) [4]	University Hospital Dataset (640 panoramic images; augmented; 80-20)	UNet & YOLO-v4	Classification	Yes	Train-test split	Accuracy, Precision / PPV, Recall / Sensitivity, F1 Score, Specificity, AP
Jundaeng, J. (Thailand, 2024) [5]	Hospital Dataset (2000 panoramic images; 70-10-20)	YOLOv8	Classification	Yes	Hold-out validation	Accuracy, Precision / PPV, Recall / Sensitivity, F1 Score, Specificity, mAP
Kabir (USA, 2021) [29]	700 periapical images; augmented; 70-10-20	HYNETS	Classification	Yes	Hold-out validation	AUC, Jaccard, Dice, PA
Kong (China, 2023) [8]	University Hospital Dataset (1747 panoramic images; augmented; 70-10-20)	PDCNN	Detection	No	Hold-out validation	Accuracy, mAP, FPS
Kurt-Bayrakdar (Turkey, 2024) [31]	University Dataset (1121 panoramic images; 80-10-10)	U-Net	Detection	Yes	Hold-out validation	Accuracy, Precision / PPV, Recall / Sensitivity, F1 Score, AUC
Lee (USA, 2022) [30]	693 periapical images; 70-10-20; additional 644 images for evaluation	U-Net	Classification	Yes	External dataset	Accuracy, Recall / Sensitivity, Specificity, AUC, Jaccard, Dice, PA

First Author (Country, Year)	Datasets	DL Model	Task	Compare	Validation Method	Performance Metrics
Li (China, 2020) [32]	Hospital Dataset & University Hospital Dataset (298 panoramic images; additional 62 images for test)	Mask R-CNN & XGBoost	Classification	Yes	3-Fold random cross-validation	Accuracy, F1 Score, Dice, mAP
Lin (Taiwan, 2024) [37]	Hospital Dataset (281 periapical images for YOLOv8 & 194 periapical images for Mask R-CNN; augmented)	YOLOv8 & Mask R-CNN	Detection	Yes	Hold-out validation	Accuracy, Precision / PPV, Recall / Sensitivity, mAP
Q. Liu (China, 2023) [24]	University Hospital Dataset (1924 panoramic images; 66-20-14)	PAR-CNN	Detection	Yes	Hold-out validation	Accuracy, Recall / Sensitivity, Specificity
Y. Liu (China, 2025) [33]	University Dataset (238 panoramic images; Not specified)	Mask R-CNN & U-Net	Segmentation & Classification	No	External dataset	Accuracy, AP, AR
Mao (Taiwan, 2023) [38]	368 periapical images; augmented; 70-30	CNN with GoogLeNet/AlexNet/Inception v3/VGG19	Detection	No	Train-test split	Accuracy, Precision / PPV, Recall / Sensitivity, F1 Score
Ryu (Korea, 2023) [39]	University Hospital Dataset (4083 panoramic images; augmented)	Faster R-CNN	Detection	Yes	5-fold cross-validation	Precision / PPV, Recall / Sensitivity, F1 Score, AUC

First Author (Country, Year)	Datasets	DL Model	Task	Compare	Validation Method	Performance Metrics
Shon (Korea, 2022) [40]	University Hospital Dataset (4097 panoramic images; augmented)	U-Net & YOLOv5	Classification	Yes	Not specified	Accuracy, Precision / PPV, Recall / Sensitivity, F1 Score
Thanathornwong (Thailand, 2020) [41]	Hospital Dataset (100 panoramic images; 70-10-20)	Faster R-CNN	Detection	Yes	5-fold cross-validation	Recall / Sensitivity, F1 Score, Specificity
Tsoromokos (The Netherlands, 2022) [25]	University Dataset (446 periapical images; augmented)	CNN	Detection	No	Hold-out validation	Accuracy, Recall / Sensitivity, Specificity
Uzun Saylan (Turkey, 2023) [3]	University Dataset (685 panoramic images; 80-10-10)	YOLO-v5	Detection	Yes	Cross validation	Precision / PPV, Recall / Sensitivity, F1 Score
Vilkomir (USA, 2024) [27]	University Dataset (1078 periapical images; augmented)	ResNet-18	Classification	Yes	Hold-out validation	Accuracy, Precision / PPV, Recall / Sensitivity, Specificity, NPV, AUC
Vollmer (Germany, 2023) [42]	Public Dataset & Hospital Dataset (1414 panoramic images; augmented)	Keypoint R-CNN	Detection	No	External dataset	mAP, mAR
Yavuz (Turkey, 2024) [14]	University Dataset (1120 periapical & 1498 bitewing images; 80-10-10)	YOLOv8-cls	Classification	Yes	Hold-out validation	Accuracy, Precision / PPV, Recall / Sensitivity, Specificity
Yu (China, 2024) [43]	705 panoramic images; augmented; 80-10-10	SegFormer	Segmentation & Classification	No	Hold-out validation	F1 Score, Jaccard

First Author (Country, Year)	Datasets	DL Model	Task	Compare	Validation Method	Performance Metrics
Zhang (China, 2025) [9]	Hospital Dataset (506 panoramic images; augmented)	Vision Transformer (ViT)	Classification	Yes	Cross validation	Accuracy, Precision / PPV, Recall / Sensitivity, F1 Score, Cross Entropy

3.8 Validation Approaches

Validation strategies also varied among the reviewed studies. Most used internal validation techniques, with common approaches including train-test splits (e.g. 70:20:10 or 80:10:10) and k-fold cross-validation to ensure more reliable performance on limited datasets. Studies conducted validation using independent datasets from different sources are rare, even though they are more relevant to real-world settings. Additionally, no studies actually implemented real-time deployment or workflow integration (Table 4). This highlights a gap between algorithm development and practical implementation. External validation and real-time integration are essential for transitioning the system using the DL approach from experimental settings to everyday clinical practice.

3.9 Quality Assessment

The quality of the reviewed studies was evaluated using the QUADAS-AI tool (Quality Assessment of Diagnostic Accuracy Studies-Artificial Intelligence), which evaluates four domains, including Patient Selection, Index Test (AI model), Reference Standard, and Flow and Timing. Each domain was rated as Low (L), Moderate (M), High (H), or Unclear (U) risk of bias according to the following criteria mentioned in Table 5. The QUADAS-AI is adapted from the widely used QUADAS-2 framework. It is designed to assess the risk of bias and methodological quality with additional considerations for studies that involve AI-based systems.

Based on these criteria, 11 studies demonstrated an overall low risk of bias, 15 studies showed a moderate risk, 1 study was rated as unclear, and no study had a high risk of bias (Table 6). Common sources of bias were caused by incomplete reporting of the characteristics of the dataset, small sample sizes, and limited details of model validation. The single unclear case also further reflects the importance of using a standardized procedure for the annotation process.

Table 5: Domain risk criteria

Domain	Rate	Criteria
Patient Selection	Low (L)	Clear dataset, appropriate inclusion/exclusion, no data leakage.
	Moderate (M)	Minor sampling bias, such as limited dataset size.
	High (H)	Clear selection bias, train-test overlap.
	Unclear (U)	Dataset source or split not described.
Index Test (AI Model)	Low (L)	Clear explanation of model architecture, preprocessing, and evaluation method with independent test set.
	Moderate (M)	Minor details missing.
	High (H)	Model tuned or re-trained on test data, unclear evaluation.
	Unclear (U)	Model or methods not explained clearly.
Reference Standard	Low (L)	Expert labelling, clear and standardized criteria, independent of AI results.
	Moderate (M)	Experts involved but with unclear criteria or consistency.
	High (H)	Non-expert labelling, reference affected by AI.
	Unclear (U)	Expertise or labeling processes not reported.
Flow & Timing	Low (L)	Justified exclusions, proper validation.
	Moderate (M)	Minor issues, such as small test set or limited validation.
	High (H)	Major issues, such as reused patients, unexplained exclusions, and no validation.
	Unclear (U)	Data flow or validation process not described.
Overall Risk	Low (L)	Most domains rated Low.
	Moderate (M)	Mix of Low and Moderate rates.
	High (H)	More than 1 domain rated High.
	Unclear (U)	More than 1 unclear domain.

4 Discussion

This systematic review synthesized recent research on the application of deep learning (DL) models to diagnose periodontitis using 2D dental radiographic images. The analysis revealed that DL models, particularly CNN-based architectures, demonstrate strong diagnostic potential in tasks such as classification, segmentation, and object detection. Most models reported high performance in binary classification tasks, with accuracy frequently exceeding 85%, and segmentation models, particularly U-Net and Mask R-CNN, achieved Dice coefficients above 0.85. These findings suggest that DL systems can effectively support the assessment of periodontal disease.

A key strength of the current research is the diversity of the DL architecture used. The Convolutional Neural Network (CNN) served as a foundational model in many studies, while more advanced networks such as ResNet captured more detailed image features for severity classification. U-Net and Mask R-CNN proved effective in segmentation tasks, and YOLO and Faster R-CNN were advantageous for real-time detection, which is practical for

Table 6: Quality assessment

First Author (Country, Year)	Patient Selection	Index Test (AI Model)	Reference Standard	Flow& Timing	Overall Risk
Alotaibi (Saudi Arabia, 2022) [1]	L	L	L	L	L
Ameli (Italy, 2024) [34]	L	L	L	U	L
Chang (Korea, 2020) [2]	M	L	M	M	M
Chen (Taiwan, 2024) [35]	M	L	L	M	M
Dai (China, 2024) [36]	L	M	L	L	M
Erturk (Turkey, 2025) [13]	L	M	L	L	L
Jiang (China, 2022) [4]	L	L	H	U	L
Jundaeng (Thailand, 2024) [5]	L	L	L	L	L
Kabir (USA, 2021) [29]	L	L	U	M	M
Kong (China, 2023) [8]	M	M	L	M	M
Kurt-Bayrakdar (Turkey, 2024) [31]	L	L	L	M	L
Lee (USA, 2022) [30]	U	L	L	M	M
Li (China, 2020) [32]	M	L	L	M	M
Lin (Taiwan, 2024) [37]	M	M	L	M	M
Q. Liu (China, 2023) [24]	L	L	L	L	L
Y. Liu (China, 2025) [33]	M	L	L	M	M
Mao (Taiwan, 2023) [38]	U	L	L	H	M
Ryu (Korea, 2023) [39]	H	M	L	M	M
Shon (Korea, 2022) [40]	H	M	L	M	M
Thanathornwong (Thailand, 2020) [41]	M	L	L	M	M
Tsoromokos (The Netherlands, 2022) [25]	M	L	L	M	M
Uzun Saylan (Turkey, 2023) [3]	L	L	L	M	L
Vilkomir (USA, 2024) [27]	L	L	L	M	L
Vollmer (Germany, 2023) [42]	L	L	L	M	L
Yavuz (Turkey, 2024) [14]	L	M	L	M	M
Yu (China, 2024) [43]	L	L	L	M	M
Zhang (China, 2025) [9]	L	L	L	M	M

clinical workflow integrations. However, due to variations in the size, types of images, evaluation metrics, and validation protocols of the data set, comparisons between studies remain a very challenging issue.

Although many studies reported internal validation through train-test splits or k-fold cross-validation, only a small number conducted external validation using independent datasets. This limits the generalizability of the findings. Additionally, although some studies compared model output with clinician performance, they were inconsistent and lacked standardization. The limited use of prospective validation and real-time deployment as-

sessments further highlights the early experimental stage of DL integration in dental diagnostics.

Although transfer learning was widely utilized to overcome dataset limitations, many studies did not clearly report model implementation details, training parameters, and annotation quality, therefore hindering reproducibility. Only a few studies discussed the use of explainability or interpretability tools, a key component in building trust in healthcare settings.

In Indonesia, the deployment of AI-based diagnostic systems must comply with the national medical device framework governed by the Ministry of Health. Under "Undang-undang Nomor 17 Tahun 2023 tentang Kesehatan" (Law No. 17 of 2023 on Health), software with a medical purpose is recognized as a medical device, or Perangkat Lunak sebagai Alat Kesehatan (Software as a Medical Device, SaMD). Such tools must undergo product registration through the Regalkes system in accordance with "Peraturan Menteri Kesehatan Nomor 62 Tahun 2017 tentang Izin Edar Alat Kesehatan" (Minister of Health Regulation No. 62 of 2017 on Product Licensing of Medical Devices), which classifies devices by risk (Class A-D). In addition to that, any diagnostic software that claims clinical utility is subject to "Peraturan Menteri Kesehatan Nomor 63 Tahun 2017 tentang Cara Uji Klinik Alat Kesehatan yang Baik" (Minister of Health Regulation No. 63 of 2017 concerning Good Clinical Evaluation Methods for Medical Devices), which require evidence of safety, effectiveness, and clinical validation prior to approval. At present, Indonesia has no AI-specific regulatory framework, but existing laws for AI in healthcare have the ability to provide complementary regulation.

To advance this field, especially in Indonesia, future research should prioritize standardized benchmarking, robust multicenter external validation, transparent reporting of model architectures, training methods, and annotation processes, as well as post-market performance monitoring consistent with the regulations. Integration of explainable AI techniques and deployment-focused studies will also be essential to transition from technical feasibility to clinical utility.

5 Conclusion

This review confirms the growing capability of deep learning (DL) models in diagnosing periodontitis from 2D dental radiographs. In 27 studies, various architectures demonstrated consistency in performance for classification, segmentation, and detection tasks. CNN, ResNet, U-Net, YOLO, and Mask R-CNN are some of the most widely used and effective models, each suited for specific diagnostic purposes. The use of transfer learning further improved model performance in data-limited settings. Despite favorable outcomes, the quality and diversity of the datasets, the bias of subjective labeling annotations, and the variation of the evaluation methods hinder reproducibility and fair performance comparison. These factors limit the generalizability of the findings and their integration into real-world practice. In addition, insufficient model explainability caused by lack of validation under real patient conditions can reduce clinician' trust and pose challenges to regulatory approval.

For dentists, these findings indicate that AI-assisted (Artificial Intelligence) periodontal assessment can improve efficiency and consistency, while also emphasizing the need for careful validation before clinical application. For researchers and developers, the creation

of standardized datasets, unbiased labeling, and interpretable computer-aided diagnostic (CAD) systems are essential for robust and clinically relevant implementation. In conclusion, while DL can effectively support periodontal diagnosis from radiographic images, its successful adoption will depend on clear reporting standards, cross-disciplinary collaboration, and a greater focus on clinical validation and interpretability. Future studies should employ standardized and reproducible methods, as well as clinically relevant evaluation frameworks.

References

[1] G. Alotaibi, M. Awawdeh, F. F. Farook, M. Aljohani, R. M. Aldhafiri, and M. Aldhoayan, "Artificial intelligence (ai) diagnostic tools: utilizing a convolutional neural network (cnn) to assess periodontal bone level radiographically—a retrospective study," *BMC Oral Health*, vol. 22, no. 1, p. 399, 2022.

[2] H.-J. Chang, S.-J. Lee, T.-H. Yong, N.-Y. Shin, B.-G. Jang, J.-E. Kim, K.-H. Huh, S.-S. Lee, M.-S. Heo, S.-C. Choi, *et al.*, "Deep learning hybrid method to automatically diagnose periodontal bone loss and stage periodontitis," *Scientific reports*, vol. 10, no. 1, p. 7531, 2020.

[3] B. C. Uzun Saylan, O. Baydar, E. Yeşilova, S. Kurt Bayrakdar, E. Bilgir, İ. Ş. Bayrakdar, Ö. Çelik, and K. Orhan, "Assessing the effectiveness of artificial intelligence models for detecting alveolar bone loss in periodontal disease: a panoramic radiograph study," *Diagnostics*, vol. 13, no. 10, p. 1800, 2023.

[4] L. Jiang, D. Chen, Z. Cao, F. Wu, H. Zhu, and F. Zhu, "A two-stage deep learning architecture for radiographic staging of periodontal bone loss," *BMC Oral Health*, vol. 22, no. 1, p. 106, 2022.

[5] J. Jundaeng, R. Chamchong, and C. Nithikathkul, "Artificial intelligence-powered innovations in periodontal diagnosis: a new era in dental healthcare," *Frontiers in Medical Technology*, vol. 6, p. 1469852, 2025.

[6] A. G. Cantu, S. Gehrung, J. Krois, A. Chaurasia, J. G. Rossi, R. Gaudin, K. Elhennawy, and F. Schwendicke, "Detecting caries lesions of different radiographic extension on bitewings using deep learning," *Journal of dentistry*, vol. 100, p. 103425, 2020.

[7] R. Pauwels, D. M. Brasil, M. C. Yamasaki, R. Jacobs, H. Bosmans, D. Q. Freitas, and F. Haiter-Neto, "Artificial intelligence for detection of periapical lesions on intraoral radiographs: Comparison between convolutional neural networks and human observers," *Oral surgery, oral medicine, oral pathology and oral radiology*, vol. 131, no. 5, pp. 610–616, 2021.

[8] Z. Kong, H. Ouyang, Y. Cao, T. Huang, E. Ahn, M. Zhang, and H. Liu, "Automated periodontitis bone loss diagnosis in panoramic radiographs using a bespoke two-stage detector," *Computers in Biology and Medicine*, vol. 152, p. 106374, 2023.

[9] X. Zhang, E. Guo, X. Liu, H. Zhao, J. Yang, W. Li, W. Wu, and W. Sun, "Enhancing furcation involvement classification on panoramic radiographs with vision transformers," *BMC Oral Health*, vol. 25, no. 1, p. 153, 2025.

[10] H. Dujic, O. Meyer, P. Hoss, U. C. Wölfle, A. Wülk, T. Meusburger, L. Meier, V. Gruhn, M. Hesenius, R. Hickel, *et al.*, "Automatized detection of periodontal bone loss on periapical radiographs by vision transformer networks," *Diagnostics*, vol. 13, no. 23, p. 3562, 2023.

[11] M. A. Hasnain, Z. Ali, M. S. Maqbool, and M. Aziz, "X-ray image analysis for dental disease: A deep learning approach using efficientnets," *VFAST Transactions on Software Engineering*, vol. 12, no. 3, pp. 147–165, 2024.

[12] A. F. Boy, A. Akhyar, T. Y. Arif, and S. Syahrial, "Development of an artificial intelligence model based on mobilenetv3 for early detection of dental caries using smartphone images: A preliminary study," *Advances in Science and Technology. Research Journal*, vol. 19, no. 4, 2025.

[13] M. Erturk, M. Ü. Öziç, and M. Tassoker, "Deep convolutional neural network for automated staging of periodontal bone loss severity on bite-wing radiographs: An eigen-cam explainability mapping approach," *Journal of imaging informatics in medicine*, vol. 38, no. 1, pp. 556–575, 2025.

[14] M. B. Yavuz, N. Sali, S. K. Bayrakdar, C. Ekşi, B. S. İmamoğlu, İ. Ş. Bayrakdar, Ö. Çelik, K. Orhan, M. B. YAVUZ, N. SALİ, *et al.*, "Classification of periapical and bitewing radiographs as periodontally healthy or diseased by deep learning algorithms," *Cureus*, vol. 16, no. 5, 2024.

[15] Y. M. Alsakar, N. Elazab, N. Nader, W. Mohamed, M. Ezzat, and M. Elmogy, "Multi-label dental disorder diagnosis based on mobilenetv2 and swin transformer using bagging ensemble classifier," *Scientific Reports*, vol. 14, no. 1, p. 25193, 2024.

[16] A. Çelebi, A. Imak, H. Üzen, Ü. Budak, M. Türkoğlu, D. Hanbay, and A. Şengür, "Maxillary sinus detection on cone beam computed tomography images using resnet and swin transformer-based unet," *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, vol. 138, no. 1, pp. 149–161, 2024.

[17] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, *et al.*, "The prisma 2020 statement: an updated guideline for reporting systematic reviews," *bmj*, vol. 372, 2021.

[18] M. J. Page, D. Moher, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, *et al.*, "Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews," *bmj*, vol. 372, 2021.

[19] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Deep learning for diabetes: a systematic review," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2744–2757, 2020.

[20] R. Franciotti, M. Moharrami, A. Quaranta, M. Bizzoca, A. Piattelli, G. Aprile, and V. Perrotti, "Use of fractal analysis in dental images for osteoporosis detection: a systematic review and meta-analysis," *Osteoporosis International*, vol. 32, no. 6, pp. 1041–1052, 2021.

[21] E. Ferrara, B. Rapone, and A. D'Albenzio, "Applications of deep learning in periodontal disease diagnosis and management: a systematic review and critical appraisal," *Journal of Medical Artificial Intelligence*, vol. 8, 2025.

[22] Y. H. Khubrani, D. Thomas, P. J. Slator, R. D. White, and D. J. Farnell, "Detection of periodontal bone loss and periodontitis from 2d dental radiographs via machine learning and deep learning: systematic review employing appraise-ai and meta-analysis," *Dentomaxillofacial Radiology*, vol. 54, no. 2, pp. 89–108, 2025.

[23] G. Jader, J. Fontineli, M. Ruiz, K. Abdalla, M. Pithon, and L. Oliveira, "Deep instance segmentation of teeth in panoramic x-ray images," in *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 400–407, IEEE, 2018.

[24] Q. Liu, F. Dai, H. Zhu, H. Yang, Y. Huang, L. Jiang, X. Tang, L. Deng, and L. Song, "Deep learning for the early identification of periodontitis: a retrospective, multicentre study," *Clinical Radiology*, vol. 78, no. 12, pp. e985–e992, 2023.

[25] N. Tsoromokos, S. Parinussa, F. Claessen, D. A. Moin, and B. G. Loos, "Estimation of alveolar bone loss in periodontitis using machine learning," *international dental journal*, vol. 72, no. 5, pp. 621–627, 2022.

[26] M. A. Talib, M. A. Moufti, Q. Nasir, Y. Kabbani, D. Aljaghber, and Y. Afadar, "Transfer learning-based classifier to automate the extraction of false x-ray images from hospital's database," *International Dental Journal*, vol. 74, no. 6, pp. 1471–1482, 2024.

[27] K. Vilkomir, C. Phen, F. Baldwin, J. Cole, N. Herndon, and W. Zhang, "Classification of mandibular molar furcation involvement in periapical radiographs by deep learning," *Imaging science in dentistry*, vol. 54, no. 3, p. 257, 2024.

[28] M. R. Hasan, M. I. Fatemi, M. Monirujjaman Khan, M. Kaur, and A. Zaguia, "Comparative analysis of skin cancer (benign vs. malignant) detection using convolutional neural networks," *Journal of Healthcare Engineering*, vol. 2021, no. 1, p. 5895156, 2021.

[29] T. Kabir, C.-T. Lee, J. Nelson, S. Sheng, H.-W. Meng, L. Chen, M. F. Walji, X. Jiang, and S. Shams, "An end-to-end entangled segmentation and classification convolutional neural network for periodontitis stage grading from periapical radiographic images," in *2021 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pp. 1370–1375, IEEE, 2021.

[30] C.-T. Lee, T. Kabir, J. Nelson, S. Sheng, H.-W. Meng, T. E. Van Dyke, M. F. Walji, X. Jiang, and S. Shams, "Use of the deep learning approach to measure alveolar bone level," *Journal of clinical periodontology*, vol. 49, no. 3, pp. 260–269, 2022.

[31] S. Kurt-Bayrakdar, İ. Ş. Bayrakdar, M. B. Yavuz, N. Sali, Ö. Çelik, O. Köse, B. C. Uzun Saylan, B. Kuleli, R. Jagtap, and K. Orhan, "Detection of periodontal bone loss patterns and furcation defects from panoramic radiographs using deep learning algorithm: a retrospective study," *BMC Oral Health*, vol. 24, no. 1, p. 155, 2024.

[32] H. Li, J. Zhou, Y. Zhou, J. Chen, F. Gao, Y. Xu, and X. Gao, "Automatic and interpretable model for periodontitis diagnosis in panoramic radiographs," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 454–463, Springer, 2020.

[33] Y. Liu, L. Gao, Y. Jiang, T. Xu, L. Peng, X. Zhao, M. Yang, J. Li, and S. Liang, "Ai-aided diagnosis of periodontitis in oral x-ray images," *Displays*, vol. 86, p. 102895, 2025.

[34] N. Ameli, M. P. Gibson, I. Kornerup, M. Lagravere, M. Gierl, and H. Lai, "Automating bone loss measurement on periapical radiographs for predicting the periodontitis stage and grade," *Frontiers in dental medicine*, vol. 5, p. 1479380, 2024.

[35] I.-H. Chen, C.-H. Lin, M.-K. Lee, T.-E. Chen, T.-H. Lan, C.-M. Chang, T.-Y. Tseng, T. Wang, and J.-K. Du, "Convolutional-neural-network-based radiographs evaluation assisting in early diagnosis of the periodontal bone loss via periapical radiograph," *Journal of dental sciences*, vol. 19, no. 1, pp. 550–559, 2024.

[36] F. Dai, Q. Liu, Y. Guo, R. Xie, J. Wu, T. Deng, H. Zhu, L. Deng, and L. Song, "Convolutional neural networks combined with classification algorithms for the diagnosis of periodontitis," *Oral radiology*, vol. 40, no. 3, pp. 357–366, 2024.

[37] T.-J. Lin, Y.-C. Mao, Y.-J. Lin, C.-H. Liang, Y.-Q. He, Y.-C. Hsu, S.-L. Chen, T.-Y. Chen, C.-A. Chen, K.-C. Li, *et al.*, "Evaluation of the alveolar crest and cemento-enamel junction in periodontitis using object detection on periapical radiographs," *Diagnostics*, vol. 14, no. 15, p. 1687, 2024.

[38] Y.-C. Mao, Y.-C. Huang, T.-Y. Chen, K.-C. Li, Y.-J. Lin, Y.-L. Liu, H.-R. Yan, Y.-J. Yang, C.-A. Chen, S.-L. Chen, *et al.*, "Deep learning for dental diagnosis: a novel approach to furcation involvement detection on periapical radiographs," *Bioengineering*, vol. 10, no. 7, p. 802, 2023.

[39] J. Ryu, D.-M. Lee, Y.-H. Jung, O. Kwon, S. Park, J. Hwang, and J.-Y. Lee, "Automated detection of periodontal bone loss using deep learning and panoramic radiographs: a convolutional neural network approach," *Applied Sciences*, vol. 13, no. 9, p. 5261, 2023.

[40] H. S. Shon, V. Kong, J. S. Park, W. Jang, E. J. Cha, S.-Y. Kim, E.-Y. Lee, T.-G. Kang, and K. A. Kim, "Deep learning model for classifying periodontitis stages on dental panoramic radiography," *Applied Sciences*, vol. 12, no. 17, p. 8500, 2022.

[41] B. Thanathornwong and S. Suebnukarn, "Automatic detection of periodontal compromised teeth in digital panoramic radiographs using faster regional convolutional neural networks," *Imaging Science in Dentistry*, vol. 50, no. 2, p. 169, 2020.

[42] A. Vollmer, M. Vollmer, G. Lang, A. Straub, A. Kübler, S. Gubik, R. C. Brands, S. Hartmann, and B. Saravi, "Automated assessment of radiographic bone loss in the posterior maxilla utilizing a multi-object detection artificial intelligence algorithm," *Applied Sciences*, vol. 13, no. 3, p. 1858, 2023.

[43] H. Yu, X. Ye, W. Hong, R. Shi, Y. Ding, and C. Liu, "A cascading learning method with segformer for radiographic measurement of periodontal bone loss," *BMC Oral Health*, vol. 24, no. 1, p. 325, 2024.