BMC Oral Health

SYSTEMATIC REVIEW Open Access

Evaluating vision transformers and convolutional neural networks in the context of dental image processing: a systematic review

Turgut Felek[1,2], Hümeyra Tercanlı[3*] and Rümeysa Şendişçi Gök[4]

Abstract

Background The aim of this systematic review is to compare the efficacy of convolutional neural networks (CNN) and Vision Transformers (ViT) in the field of dental imaging, in order to examine in depth the potential, advantages, and limitations of both models in this domain.

Methods The search strings used in the study were "(("Vision Transformer" OR ViT OR "Transformer architecture") AND ("Convolutional Neural Network" OR CNN OR ConvNet) AND (Dental OR Dentistry OR "Maxillofacial" OR "Oral Radiology") AND (Image OR Imaging OR Radiograph))". The search was conducted in January 2025. Two investigators independently evaluated the full texts of all eligible articles and excluded those that did not meet the inclusion/exclusion criteria.

Results Of 2596 articles, 21 met the inclusion criteria. Depending on the task category, of the 21 studies that were reviewed, 14 (66.7%) utilized classification, while 7 (33.3%) utilized segmentation. Panoramic radiography is the most commonly used imaging modality (52.3%) and the ViT-based model was observed to have the highest performance (58%).

Conclusion ViT-based deep learning models tend to exhibit higher performance in many dental image analysis scenarios compared to traditional convolutional neural networks. However, in practice CNN and ViT approaches can be used in a complementary manner.

Keywords Vision transformers, Convolutional neural networks, Dental imaging, Image classification, Image segmentation

*Correspondence:
Hümeyra Tercanlı
ysl_hmyr25@hotmail.com
[1]Institute of Natural and Applied Sciences, Akdeniz University, Antalya, Türkiye
[2]Department of Information Technologies, Faculty of Dentistry, Akdeniz University, Antalya, Turkey
[3]Departmant of Oral and Maxillofacial Radiology, Faculty of Dentistry, Akdeniz University, Antalya, Türkiye
[4]Departmant of Oral and Maxillofacial Radiology, Faculty of Dentistry, Antalya Bilim University, Antalya, Türkiye

© The Author(s) 2025. **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/.

Introduction

In particular, dental images provide great convenience for dentists in the diagnostic evaluation of diseases, in the formulation of treatment plans, and in monitoring the treatment process [1]. The information density and complexity of dental images require robust computational techniques for accurate analysis and interpretation. In this context, deep learning-based models have achieved significant success in the processing and analysis of dental images in recent years [2]. In the field of dental radiology, deep learning techniques are widely used, especially for tasks such as classification and object detection. These techniques are effectively employed in numerous clinical applications such as identification and numbering of teeth, detection of caries, analysis of root morphology, detection of root fractures, assessment of alveolar bone loss, identification of periapical lesions, detection of impacted teeth, identification of bone lesions, diagnosis of maxillary sinusitis, detection of temporomandibular joint osteoarthritis, osteoporosis screening, assessing the proximity of the inferior alveolar nerve canal to the roots of mandibular molars, and classification of dental implant systems [3, 4].

Convolutional neural networks (CNNs), a class of deep learning architecture, are widely used in various medical image analysis domains such as segmentation, detection, anomaly classification, and computer-aided diagnosis [5]. CNNs learn a given image's intrinsic features and structural patterns in a manner similar to the way humans observe objects [5]. These networks extract meaningful features by directly analysing visual patterns from the raw image pixels [6]. In the field of dental imaging, the CNNs can provide rapid and accurate results for the detection of dental caries, periodontal disease and other abnormalities. Furthermore, CNNs are a preferred tool in dental imaging applications due to their high success rate in feature extraction and classification [7]. CNN's ability to use convolutional kernels that move across various parts of the screen to reduce the number of parameters is one of its primary benefits. As a result, the model becomes simpler and more useful for training [8].

In dentistry, there are four main application areas of CNNs: segmentation, detection, classification, and image quality enhancement. These applications can be performed on dental images acquired intraoral (periapical radiographs and bite-wing images) and extraoral (panoramic radiographs and cone-beam computed tomography (CBCT)) [9, 10]. An example of the application of CNNs in panoramic radiography for dental imaging is shown in Fig. 1.

In recent years, Vision Transformers (ViTs) have emerged as an alternative to traditional CNN models [11]. In computer vision tasks, dividing images into small patches and processing them with a transformer architecture is known as a ViT. It can learn the relationships between these patches and capture long-range dependencies [12]. The core principle of this model is based on the self-attention mechanism, which determines a sequence's representation by discovering relationships between its elements at various positions [13]. Looking at the advantages of ViTs, they use a flexible weighting strategy that can dynamically assign different weights to each component of the input, allowing the model to focus more on relevant features. In addition, ViTs are also widely used for image processing tasks such as image creation and captioning. An example of the ViT processing of a dental panoramic radiograph is shown in Fig. 2. ViTs can learn more about long-term dependencies and can handle data across dimensions more flexibly than traditional CNNs [13]. In summary, using ViTs in dental imaging can increase the flexibility of deep learning models by allowing for a more comprehensive contextual analysis.

Comparisons between CNNs and ViTs in the field of dental image processing are extremely important for understanding the advantages and limitations of each model. CNNs achieve high success especially in analyzing local patterns and structures, whereas the strength of ViTs in learning long-range dependencies allows for

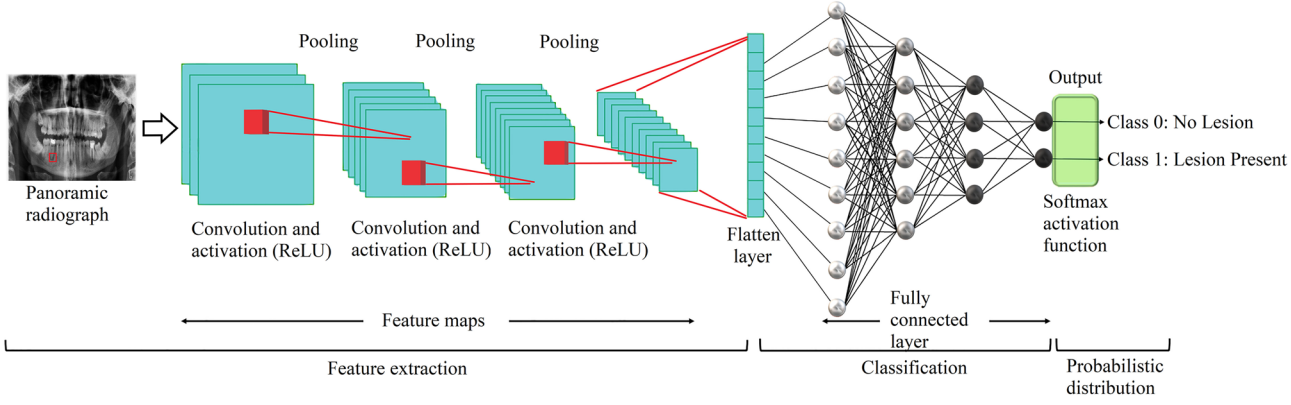


Fig. 1 An example of dental image analysis using CNN on panoramic radiograph

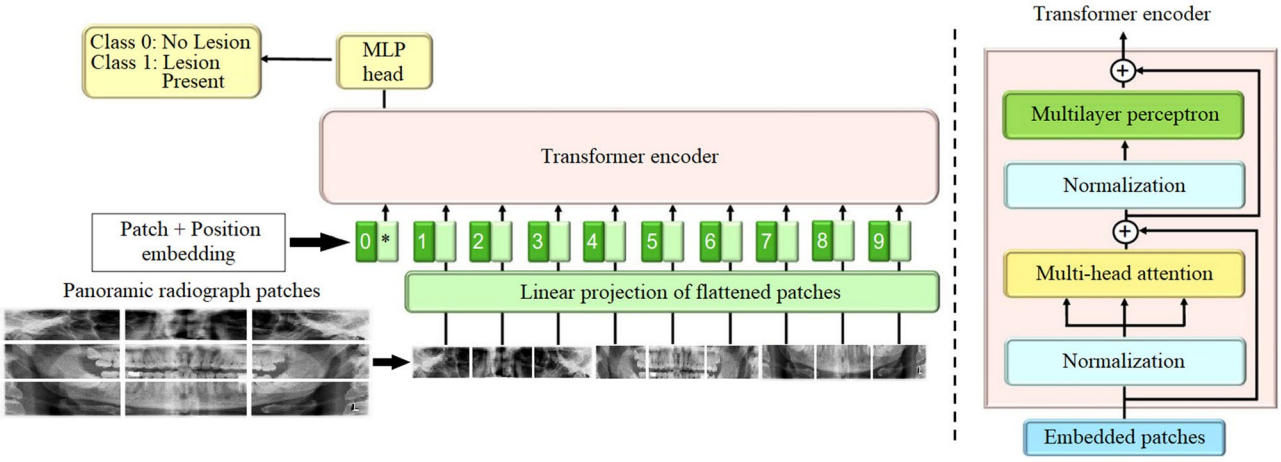


Fig. 2 An example of dental image analysis using ViT on panoramic radiograph

more comprehensive analysis [14]. However, each model has certain limitations. For example, training CNNs on very large datasets increases computational costs, whereas ViTs require more data and have longer training times [15, 16].

The aim of this systematic review is to compare the efficacy of CNNs and ViTs in the field of dental imaging, in order to examine in depth the potential, advantages, and limitations of both models in this domain.

Materials and methods

Search strategy

The literature search was conducted in accordance with predetermined guidelines [17]. Primary studies were identified in January 2025 by searching four data sources (Google Scholar, PubMed, Scopus, and IEEE Xplore), thereby capturing both clinical and technical literature. The search strings used in the study were "((“Vision Transformer” OR ViT OR “Transformer architecture”) AND (“Convolutional Neural Network” OR CNN OR ConvNet) AND (Dental OR Dentistry OR “Maxillofacial” OR “Oral Radiology”) AND (Image OR Imaging OR Radiograph))". The search was conducted in January 2025. The exact syntax and filters used for each database, along with the number of initial hits, are provided in Supplementary Material 1.

Inclusion criteria

The inclusion criteria for selecting articles were that they were written in English and published between January 2021 and January 2025. This date range was chosen because Vision Transformers had not been introduced until the end of 2020 [12]. In addition, studies had to make a direct comparison between CNNs and VITs on dental images and both architectures had to be used with their respective reported results. Studies proposing a hybrid architecture combining the two types of models were also considered for comparison. The selected studies needed to involve both deep learning architectures and be based on an image dataset appropriate for classification or segmentation tasks. While studies that performed validation with independent external datasets were prioritized, studies that only validated on their own dataset were also included.

Exclusion criteria

The exclusion criteria were defined as follows: *(1)* studies based on animal experiments; *(2)* studies with no full text available; *(3)* studies that did not include statistical data; *(4)* studies focusing on only one deep learning architecture (either ViT or CNN).

Study selection and data collection

After screening the titles and abstracts of all identified publications, two investigators (TF and RŞG) independently evaluated the full texts of all eligible articles and excluded those that did not meet the inclusion/exclusion criteria. Disagreements between the reviewers were resolved through discussion until consensus was reached or by involving a third reviewer (HT). From each publication, the following data were collected: the title of the study, study design (datasets, modality of dental images, machine learning algorithms, study factor and its definition, algorithm implementation, comparisons), primary outcomes, and overall findings.

Outcome measures and effect estimates

The primary outcome was the “F1-score” for classification tasks (to balance precision and recall). For segmentation studies, the main effect measure was the “Dice similarity coefficient” (DSC). When authors reported multiple metrics, we also extracted accuracy, area under the ROC curve (AUC), and Intersection-over-Union (IoU) to allow qualitative comparison. No statistical

pooling (meta-analysis) was attempted because of metric heterogeneity; therefore, effect estimates are presented narratively and in descriptive tables.

Assessment of risk of bias

The risk of bias in the included studies was qualitatively assessed by two independent reviewers (TF and HT) using a methodology adapted from the PROBAST (Prediction model Risk of Bias Assessment Tool) guidelines. While a formal PROBAST assessment was not feasible due to reporting heterogeneity, we evaluated each study across three key domains critical for AI-based diagnostic studies: (1) Dataset and Participants (public vs. private/proprietary data source, sample size, and clarity of patient selection criteria); (2) Validation (use of external validation on an independent population/source vs. reliance solely on internal validation such as train–test splits or cross-validation within a single dataset); and (3) Outcome Measurement and Metric Reporting (whether a balanced set of performance metrics—e.g., F1-score, Dice, AUC, precision/recall—was reported, noting that accuracy alone can be misleading. Each domain was judged as "low," "low-moderate," "moderate," or "high risk," and studies were also assigned an overall risk of bias using the same three levels (Supplementary Material 2).

Results

A first search of databases identified 2596 records (Google Scholar: 2530, PubMed: 13, Scopus: 18, IEEE Xplore: 35). After removing 106 duplicate records, 2490 articles were screened based on their titles and abstracts, of which 2430 were excluded as they did not meet the scope of this review. This left 60 reports that were sought for full-text retrieval. Of these, 22 reports could not be retrieved, leaving 38 for full-text eligibility assessment. At this stage, a total of 17 reports were excluded for the following reasons: study focused on only one deep learning architecture (Reason 4, $n = 13$), no full text was available (Reason 2, $n = 2$), study was based on animal experiments (Reason 1, $n = 1$), and the study did not include statistical data (Reason 3, $n = 1$). Ultimately, 21 studies met all inclusion criteria and were included in the qualitative synthesis, as detailed in the PRISMA flow diagram (Fig. 3). In reviewing the studies, they were evaluated across seven categories for comparison: task category, subject of the study, dataset, image type, pre-training, best-performing model, and which model was superior.

Table 1 presents a general overview of the 21 included studies [18–38]. Depending on the task category, of the 21 studies that were reviewed, 14 (66.7%) classification [21, 23–26, 29–32, 34–38], while 7 (33.3%) utilized segmentation [18–20, 22, 27, 28, 33]. Fig. 4A shows the distribution of task categories across the 21 studies. For segmentation tasks, the research topics included lesion detection, tooth loss, mandibular segmentation for implants, caries assessment, delineation of pathological conditions and infections related to the maxillary sinus, tooth segmentation, and cancerous lesion segmentation. As for classification tasks, the objectives included caries detection; detection and classification of mandibular fractures; classification of diseases of the dental region; identification of tooth crowns; early diagnosis of dental problems; detection of dens evaginatus; autonomous identification of dental implant brands; classification of midpalatal sutures at different maturation stages; classification of periodontal diseases; classification of radiographic images of dental structures; and detection of dental furcation involvement.

Figure 4B, which shows the distribution of imaging modalities across 21 eligible studies [18–38], shows that panoramic radiography is the most commonly used imaging modality ($n = 12$, 57.1%) [19, 23, 24, 26, 29–33, 35, 36, 38], followed by CBCT ($n = 5$, 23.8%) [18, 22, 27, 34, 37], periapical radiography ($n = 2$, 9.5%) [20, 21], MRI ($n = 1$, 4.8%) [28], and intraoral 3D scans ($n = 1$, 4.8%) [25].

Figure 5 presents the findings indicating which technique was considered most effective. The results of the studies indicate that the ViT-based model was observed to have the highest performance (58%). Additionally, some studies reported that in certain cases CNN and ViT each had scenarios where one performed better than the other.

Risk of bias in studies

Our qualitative assessment of the 21 included studies revealed several potential sources of bias that may impact the generalizability of their findings. A summary of the evaluation is provided in Supplementary Material 2. The main finding was the heavy reliance on private and institutional data sets; 17 out of 21 studies (81%) used data that were not public. This limits the reproducibility of the findings and may introduce selection bias, as the data may not be representative of the wider population. In addition, the vast majority of studies (18 out of 21, 86%) did not have external validation. They validated their models by using only internal train-test distribution or cross-validation using data from a single source, which increases the risk of overfitting and means that reported performance may not be consistent across different clinical settings. Although the majority of studies (19 out of 21, 90%) reported comprehensive metrics beyond precision, the most important methodological issue in the literature is the lack of external validation.

Discussion

This systematic review aimed to evaluate the effectiveness of CNNs and ViTs in the field of dental imaging. This study also examined how these two approaches are

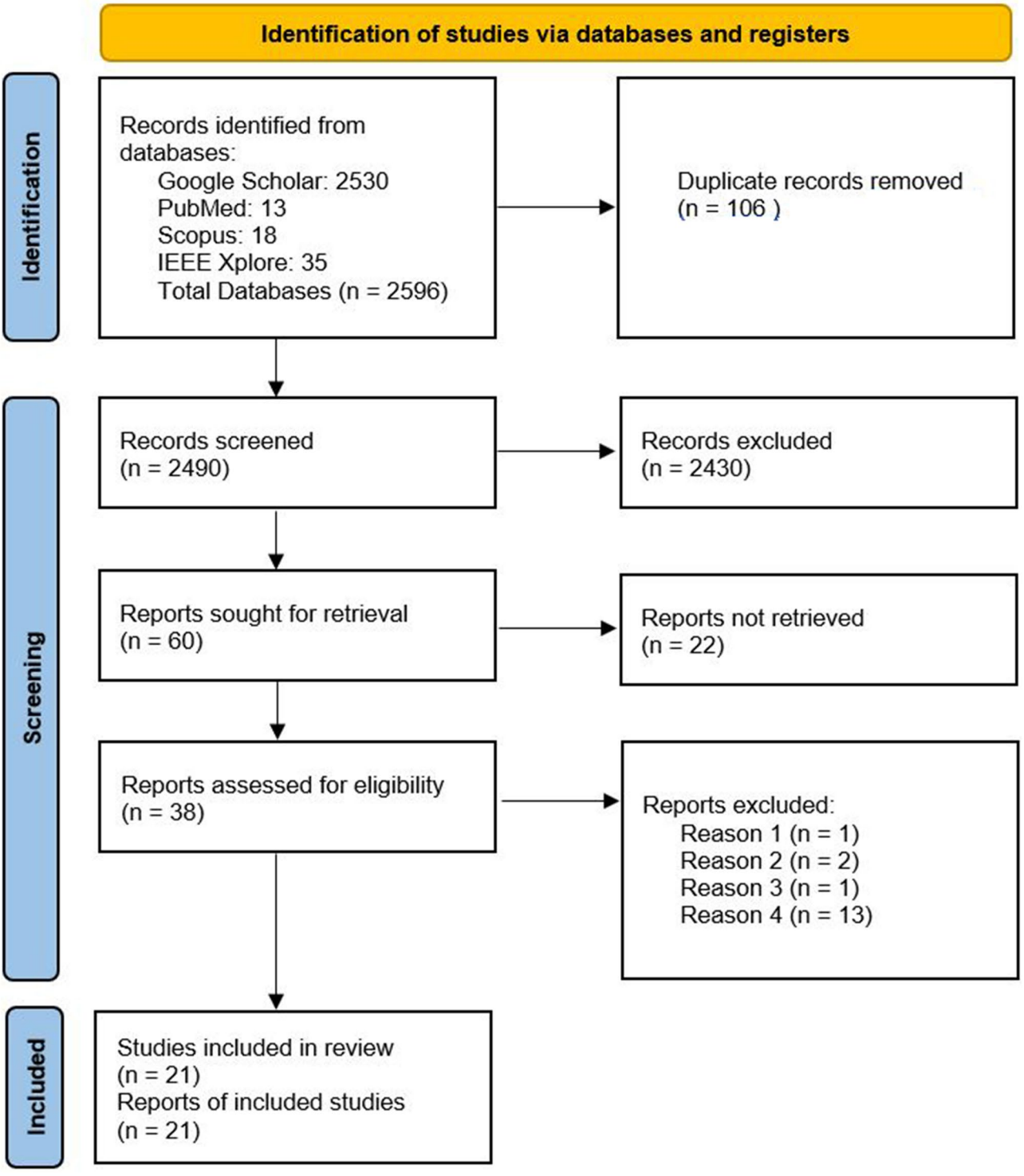


Fig. 3 Identification of new studies via databases and registers: Reason 1: studies based on animal experiments; Reason 2: studies with no full text available; Reason 3: studies that did not include statistical data; Reason 4: studies focusing on only one deep learning architecture (either ViT or CNN)

applied and in which types of dental imaging tasks they are more successful, yielding important implications for future research. Although several narrative and scoping reviews have explored artificial-intelligence applications in dentistry, none has undertaken a focused, systematic head-to-head appraisal of the field's two dominant deep-learning paradigms—convolutional neural networks (CNNs) and Vision Transformers (ViTs). The present review addresses this shortcoming by collating and critically evaluating all studies that explicitly compare these

Table 1 Summary of 21 papers selected based on the search criteria

Title	Task category	Task	Dataset	Image Type	Pre training	Best architecture and its performance	Which is better?
Leveraging Pre-trained Transformers for Efficient Segmentation and Lesion Detection in Cone-Beam Computed Tomography Scans[18]	Segmentation	3D CBCT Segmentation and Lesion Detection	138 anonymized CBCT scans from Penn Dental Medicine dataset	CBCT	Swin-UNETR-PRETRAIN	Swin-UNETR-PRETRAIN (Transformer-based) – Lesion Dice = 85.12%, Restorative Material Dice = 82.82%, Bone Dice = 91.78%, Tooth Structure Dice = 90.29%; Lesion Detection Sensitivity = 100%, Specificity = 92%	Transformer
A new dataset of oral panoramic xray images and parallel network using transformers for medical image segmentation[19]	Segmentation	Mandible and teeth segmentation for tooth loss and mandibular implant therapy	MRDataset (open access) & DRIVE dataset	Panoramic	Unknown	CBTrans, 90.39% ACC, 91.87% IOU on MRDataset	CBTrans (hybrid CNN and Swin Transformer
Caries segmentation on tooth X-ray images with a deep network[20]	Segmentation	Dental Caries Segmentation	Private dental radiograph dataset (153 images, augmented to 800)	Periapical Radiography	ImageNet	Proposed hybrid U-Net (dice: 0.7487, precision: 0.7443)	Proposed hybrid U-Net (outperformed U-Net, Trans-UNet, Swin-UNet)
Performance comparison of multifarious deep networks on caries detection with tooth X-ray images[21]	Classification	Caries Detection	Zhejiang Univ. Periapical Radiographs (n = 392)	Periapical Radiography	ImageNet (YOLOv5, DETR, Trans-UNet)	YOLOv5 achieved the highest caries detection performance with F1-score = 0.87; precision = 0.93; sensitivity = 0.82; specificity = 0.94;	CNN
Maxillary sinus detection on cone beam computed tomography images using ResNet and Swin Transformer-based UNet [22]	Segmentation	Maxillary Sinus Pathology & Infection Segmentation	Private dental clinic CBCT scans (n = 298)	CBCT	ResNet	Res-Swin-UNet – F1 = 91.72%, Acc = 99%, IoU = 84.71%	Transformer
Detection of mandibular fractures on panoramic radiographs using deep learning[23]	Classification	Mandibular Fracture Detection & Classification	Charité Berlin Panoramic Radiographs (n = 6404; 1624 fractured, 4780 intact)	Panoramic	ImageNet (Swin-Transformer)	Faster R-CNN + Swin-Transformer Prec = 93.5%; Rec = 96%; F1 = 94.7%; AUC = 0.977	Hybrid model
A Hybrid Transfer Learning Approach to Teeth Diagnosis Using Orthopantomogram Radiographs [24]	Classification	Dental Region Disease Classification	Kaggle OPG radiographs (n = 1,262; 9,283 annotated teeth; augmented to ~10,859 instances)	Panoramic	ImageNet, COCO (YOLOv10)	Vision Transformer Accuracy = 0.96; F1-score = 96%	Transformer

Table 1 (continued)

Title	Task category	Task	Dataset	Image Type	Pre training	Best architecture and its performance	Which is better?
An Application of 3D Vision Transformers and Explainable AI in Prosthetic Dentistry[25]	Classification	3D scan classification of prepared teeth for partial crown type	34 deidentified tooth specimens	Intraoral 3D scans	ImageNet21k and fine-tuned on ImageNet1k	CNN + ViT hybrid clearly outperforms both standalone models (90% vs. ViT's 80% and 3D CNN's 60% acc.)	Hybrid model
Identification of Dental Lesions Using Self Supervised Vision Transformer in Radiographic X-ray Images[26]	Classification	Dental lesion identification	1 418 manually collected panoramic radiographs	Panoramic	Self-supervised on ImageNet-ISLVRC	Dual-channel Vision Transformer Sensitivity = 96.8%, Specificity = 95.9%, Accuracy = 97.2 %, F1 = 97.1%	Transformer
CTA-UNet: CNN-transformer architecture UNet for dental CBCT images segmentation [27]	Segmentation	Dental CBCT image segmentation	CBCT scans from 200 patients; 45 volumes annotated (440 slices each)	CBCT	CTAMIM self-supervised	CTA-UNet (CNN-Transformer UNet) with CTAMIM –DSC = 87.12%, IoU = 78.90%, HD95 = 0.525 mm, ASSD = 0.199 mm	Hybrid model
Auxiliary Segmentation Method of Osteosarcoma MRI Image Based on Transformer and U-Net [28]	Segmentation	Auxiliary osteosarcoma MRI image segmentation	4,000+ osteosarcoma MRI scans from the Monash University AI Research Center	MRI	No (Scratch)	OSTRansnet (U-Net + channelized Transformer + edge-enhancement) achieved IoU = 90.4%, DSC = 94.9%, Precision = 92.4%, Recall = 98.1%, F1 = 95.1%	Hybrid model
Tooth Type Enhanced Transformer for Children Caries Diagnosis on Dental Panoramic Radiographs [29]	Classification	Tooth-level caries diagnosis	6 028 extracted tooth images from 304 panoramic radiographs	Panoramic	No (Scratch)	Tooth-Type Enhanced Swin Transformer (T2S-Transformer) – Acc = 85.5%; Prec = 88.3%; Rec = 83.1%; F1 = 85.6%; AUC = 0.9223	Transformer
ResNet-Transformer deep learning model aided detection of dens evaginatus [30]	Classification	Dens evaginatus detection	Private clinical panoramic radiographs from West China Hospital of Stomatology (Sichuan University) – 1 410 images	Panoramic	No (Scratch)	ResNet-Transformer hybrid – Sens = 0.86; Spec = 0.92; PPV = 0.9149; NPV = 0.8679; F1 = 88.6%; AUC = 0.956	Hybrid model
Position Weighted Convolutional Neural Network for Unbalanced Children Caries Diagnosis [31]	Classification	Unbalanced children caries diagnosis on individual tooth crops	6 028 extracted tooth images from 304 children's panoramics (private)	Panoramic	No (Scratch)	Position-Weighted CNN (PW-CNN) – Accuracy = 88.6%; Precision = 88.7%; Recall = 89.3%; F1 = 89%; AUC = 0.9315	CNN

Table 1 (continued)

Title	Task category	Task	Dataset	Image Type	Pre training	Best architecture and its performance	Which is better?
A Comparative Analysis of Deep Learning-Based Approaches for Classifying Dental Implants Decision Support System [32]	Classification	Autonomous dental implant brand identification	1 258 panoramic images from Erciyes University (Straumann 106, Bilimplant 121, Implance 236, Dyna 248, Megagen 253, Dentium294)	Panoramic	ImageNet	ConvNeXt (small) – Acc = 94.2%, Prec = 95.6%, Rec = 93.3%, F1 = 94.2%	CNN
Transformer-Based Deep Learning Network for Tooth Segmentation on Panoramic Radiographs[33]	Segmentation	Tooth tissue segmentation	PLAGH-BH with external validation on PDX public set	Panoramic	ImageNet	Swin-Unet (pure Transformer) on PLAGH-BH: F1 = 63.7%; IoU = 46.9%; Acc = 88.5%	Transformer
Prediction of midpalatal suture maturation stage based on transfer learning and enhanced vision transformer[34]	Classification	Midpalatal suture maturation stage classification	2 518 CBCT images (Stages A–E: 416, 168, 514, 598, 822)	CBCT	ImageNet	CNN-enhanced ViT_B/16 – Acc = 95.75%, Prec = 95.15%, Rec = 93.15%, F1 = 94.13%, Macro-AUC = 97.89%, Micro-AUC = 98.36%	Hybrid model
Multi-label dental disorder diagnosis based on MobileNetV2 and swin transformer using bagging ensemble classifier[35]	Classification	Multi-label diagnosis of fillings, implants, impacted teeth, and cavities	1 272 panoramics cropped into 4 023 labeled images (2 609 fillings, 910 implants, 301 impacted teeth, 203 cavities)	Panoramic	ImageNet	MobileNetV2 + Swin Transformer with bagging ensemble – Acc = 95.6%, Prec = 95.7%, Rec = 95.4%, Spec = 95.7%, DSC = 95.5%	Hybrid model
Unsupervised few shot learning architecture for diagnosis of periodontal disease in dental panoramic radiographs[36]	Classification	Periodontal disease diagnosis via few-shot learning	Three de-identified panoramic radiograph sets—Tufts Dental Database (1 000 images), HUSHDD (256 images), NMICDD (116 images)	Panoramic	No (Scratch)	UNet-CVAE + DBSCAN, TDD: Acc = 85.4%, Prec = 85.2%; HUSHDD: Acc = 94.9%, Prec = 97.4%; NMICDD: Acc = 88.4%, Prec = 94.1%	CNN

Table 1 (continued)

Title	Task category	Task	Dataset	Image Type	Pre training	Best architecture and its performance	Which is better?
Preparing for Downstream Tasks in AI for Dental Radiology: A Baseline Performance Comparison of Deep Learning Models[37]	Classification	CBCT feature classification (sinus, incisors, mental foramen, root–nerve relation)	2 653 anonymized slices from CBCT volumes (1098 sinus, 705 incisors, 400 premolar, 450 third-molar images)	CBCT	No (Scratch)	ViT achieved a median accuracy of 0.98 in the maxillary vs. mandibular incisor classification. CNN achieved a median accuracy of 0.99 in the maxillary sinus "opaque" appearance classification	Comparable
Enhancing Furcation Involvement Classification on Panoramic Radiographs with Vision Transformers[38]	Classification	Furcation involvement classification	1 568 tooth images from 506 panoramic radiographs	Panoramic	No (Scratch)	Vision Transformer (ViT) – Precision = 98%; Recall = 92%; F1 = 95%; Accuracy = 92%; AUC = 0.98	Transformer

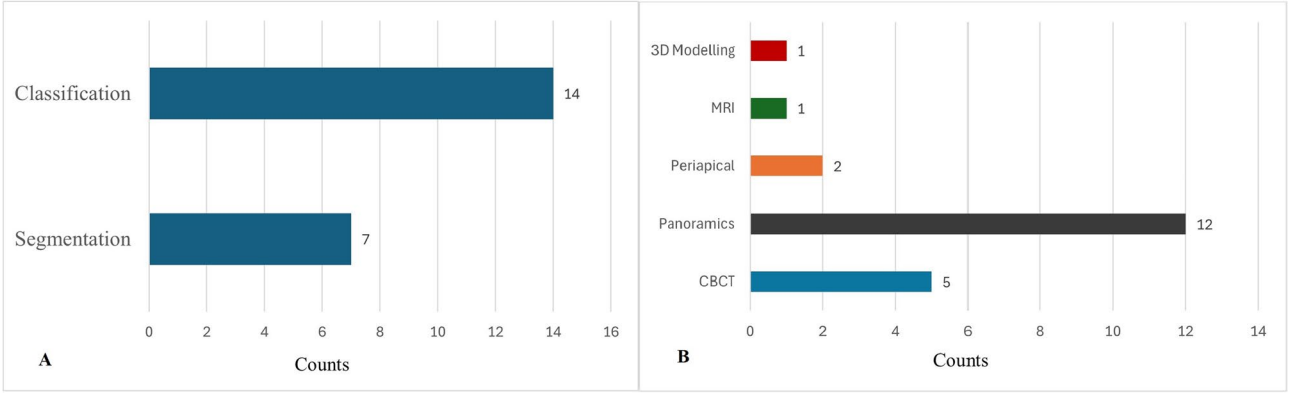


Fig. 4 Summary of the 21 included studies, showing their distribution by (**A**) AI task and (**B**) dental imaging modality

architectures across diverse dental-imaging tasks. In doing so, it delivers three key contributions: (i) new, consolidated evidence on performance patterns favouring either CNNs, ViTs, or their emerging hybrid variants; (ii) insight into how the discipline is pivoting from models optimised for local feature extraction to those capable of capturing long-range contextual cues; and (iii) practical guidance for designing next-generation diagnostic tools that balance accuracy, data requirements and computational cost. Ultimately, this study results were intended to steer both researchers and clinicians toward more informed choices when developing or adopting deep-learning solutions for dental radiology.

In this systematic review, the results reported for classification tasks generally indicated a trend that ViT–based models tend to exhibit superior performance.

Performance analysis by task category

Classification tasks

Of the 21 studies examined [18–38], 14 focused on classification tasks [21, 23–26, 29–32, 34–38], and these studies most frequently used panoramic radiographic images [23, 24, 26, 29–32, 35, 36, 38]. In many classification studies, ViT architectures provided higher accuracy or F1 scores compared to the CNN architectures they were evaluated against. For example, in a study on mandibular fracture detection, a Swin-Transformer–based model demonstrated extremely high performance with an F1 score of 94.7% and AUC = 0.977 [23]. Similarly, the Swin-Transformer model used to detect caries in children achieved approximately 82% accuracy, which was superior to the CNN comparator [29]. In the early diagnosis of dens evaginatus lesions, a proposed ResNet-Transformer hybrid model also yielded a strong result with an F1 score of 88.6% [30]. However, this review has shown that CNN models have also shown better results in some scenarios. In particular, CNNs continue to be successful in situations where local features are critical or data are limited. Indeed, a study on caries detection in periapical radiographs reported that a YOLOv5-based CNN model, with an F1 score of 87%, outperformed the Transformer alternatives [21]. Likewise, in the autonomous

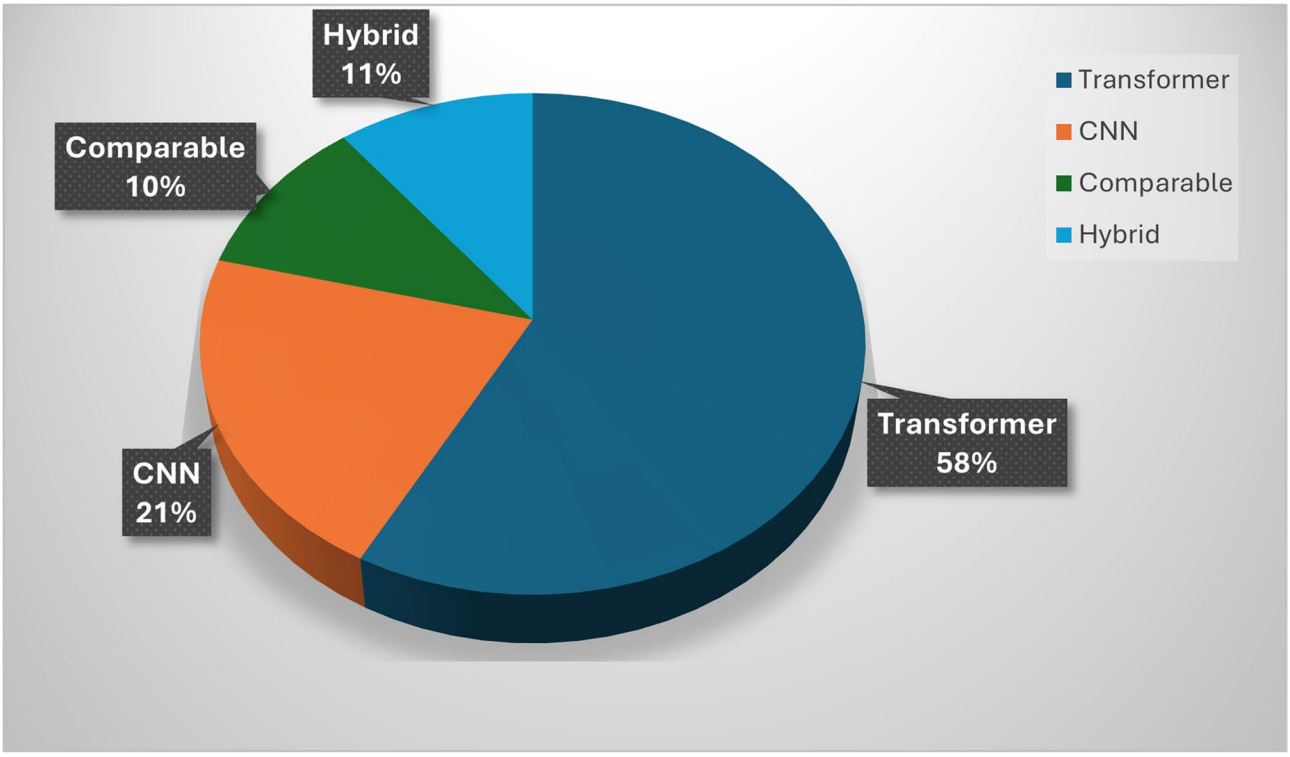


Fig. 5 Percentage distribution of the best-performing architectures from 21 studies, categorized as Transformer, CNN, Hybrid, or Comparable

identification of dental implant brands, a conventional CNN variant (ConvNeXt) achieved ~ 94% accuracy, surpassing the ViT-based approaches [32]. These findings suggest that in classification tasks, both architectures can each be superior under certain conditions. However, considering the overall picture, it can be said that Transformer-based architectures offer higher accuracy and generalization ability in especially complex classification problems, whereas CNNs still provide a strong standard in specific cases.

Segmentation tasks

A similar situation is true for the segmentation studies included in this review. A total of 7 articles addressed segmentation tasks [18–20, 22, 27, 28, 33], three of which focused on segmentation problems in CBCT images [18, 22, 27]. Transformer-based models generally stood out in dental image segmentation tasks as well, often achieving higher metrics such as Dice and IoU. For example, the OSTransnet model (a hybrid Transformer + U-Net) developed by Liu et al. achieved a Dice score of 94.9% and an IoU of 90.4% on test data for bone tumor MRI segmentation, which corresponds to an improvement of about 6% compared to a pure U-Net model [28]. In the same study, this model outperformed all common CNN-based segmentation methods such as U-Net, FCN, and PSPNet, achieving a segmentation sensitivity of up to 98% (Recall 98.1%). The success of Transformer-based architectures is also striking in segmentation studies on dental radiographs. The CBTrans hybrid architecture, for example, performed well in the segmentation of all teeth in panoramic radiography, with 90.39% accuracy and 91.87% IoU [19]. The proposed hybrid U-Net model for segmenting carious regions in periapical radiographs obtained a precision of 0.7443 and a dice score of 0.7487 [20]. On the other hand, there have also been specific cases in segmentation tasks where CNN approaches were superior. Notably, evaluations on different datasets have yielded interesting findings: In the panoramic tooth segmentation study by Sheng et al. [33], a Transformer-based Swin-Unet model achieved the highest F1 score (63.7%) on a private dataset. However, in the same study, on a public dataset called PDX, a lighter CNN architecture (LinkNet) delivered a higher F1 score (87.76% vs. 82.03% for Swin-Unet) an absolute difference of 5.73% points and a relative improvement of ~ 7%. These contrasting results highlight the impact of the dataset's composition and scope on model performance. Transformer-based models can yield very strong results when they are attuned to a specific data distribution, whereas well-optimized CNNs can be advantageous on more diverse or larger population datasets. Nevertheless, the overall trend in segmentation problems indicates that the global context information provided by Transformer architectures makes an important contribution to delineating the boundaries of complex objects like dental

lesions and structures. However, it should be remembered that each architecture can outperform the other depending on the data conditions.

Overall performance summary and architectural considerations

In approximately 58% of the 21 studies reviewed, the best-performing model was a ViT architecture, whereas in 21% of the studies the best results were achieved with a CNN-based model. Hybrid CNN–Transformer architectures were identified as the most successful approach in about 11% of the studies; moreover, a small portion (~10%) of the studies reported that performance varied case by case, with CNN and ViT approaches each outperforming the other in different scenarios. These findings indicate that ViT-based methods provide a distinct advantage over CNNs in the majority of dental imaging tasks. Indeed, the data from this review show that ViT models have come to the forefront in various tasks thanks to their high accuracy and generalization capability. In one comparative study, for instance, a CNN model classified maxillary sinus opacities with ~99% accuracy, while a ViT model distinguished maxillary vs. mandibular incisors with ~98% accuracy [37]. This indicates that both approaches can perform better in certain subtasks. Therefore, neither model provides absolute superiority in every scenario; however, in the overall picture, ViT architectures are seen to offer a clear performance advantage.

CNNs have been effective in local feature extraction for many years, and they can detect details in small regions of interest (ROIs) quickly and accurately [5, 21]. In addition, ConvNeXt achieved a ~94% accuracy for the independent identification of dental implant brands, which was the highest [32]. This strong inductive bias for locality makes them highly efficient for tasks where local patterns are determinative, such as identifying the distinct texture of a specific dental implant brand or detecting the subtle density changes of incipient caries on a periapical radiograph. This success can be attributed to CNNs' ability to strongly capture local structural information while maintaining parameter efficiency and low computational cost, thanks to optimized convolutional kernels [8]. However, due to their limited receptive field, CNNs can be insufficient in learning the global context across an entire image [6]. For example, in panoramic tooth segmentation, while the Swin-Unet model gave the best result on a private dataset, on a larger public dataset (PLAGH-BH) a lighter LinkNet-based CNN outperformed Swin-Unet with an F1 score of approximately 88% [33]. This indicates that although locally focused CNNs can achieve high generalization performance in some scenarios, they may struggle in tasks that require capturing broad contextual relationships.

ViTs can effectively model long-range dependencies and global context through the self-attention mechanism [13]. This capability is particularly advantageous in tasks that require a holistic understanding of the entire image, such as classifying a mandibular fracture on a panoramic radiograph where the relationship between distant points of the jaw is crucial [23], or segmenting a diffuse lesion in a CBCT scan that spans multiple anatomical structures [18]. A dual-channel ViT model for classifying dental lesions on panoramic radiographs achieved 97% accuracy and an F1 score of 97% [26]. In furcation involvement classification, a ViT was the most successful model with 98% precision and an F1 score of 95% [38]. In segmentation tasks, Swin-UNETR reported a lesion sensitivity of 100% and a specificity of 92% in CBCT scans [18], and OSTransnet delivered a ~6% higher Dice score (94.9%) than the classic U-Net in osteosarcoma MRI segmentation [28]. On the other hand, ViT-based models generally require large-scale pre-training data and high computational resources [15, 16]. To mitigate this need, many studies have augmented ViTs with CNN components pre-trained on large datasets like ImageNet or have employed self-supervised approaches [12].

ViTs are known to be successful not only in the field of dental imaging but also in other fields of medical imaging. For instance, in a study by Tyagi et al., a Vision Transformer model outperformed a traditional CNN model in detecting pneumonia on chest radiographs with approximately 96% accuracy [39]. Similarly, the ViT-based classifier achieved higher performance than the CNN in the classification of lesions on ultrasound images of breast despite a limited number of training examples [40]. In dermatology, one study found that the VIT model was superior to convolutional networks for the classification of dermosculotic skin lesions [41]. These examples demonstrate that thanks to their unique self-attention mechanism, ViT architectures can also deliver superior results in various clinical imaging tasks.

Hybrid CNN–Transformer architectures offer synergistic performance by combining local feature extraction with global context modeling capabilities. For example, a hybrid model based on three-dimensional dental images achieved a 90% accuracy, representing a significant performance gain over both standalone architectures (a 12.5% relative improvement over the pure ViT's 80% and a 50% relative improvement over the pure 3D CNN's 60%) [25]. In a different study, a hybrid U-Net architecture achieved a higher Dice score [20] in tooth decay segmentation than classical U-Net, TransUNet, and Swin-Unet. CTA-Unet al.so demonstrated a strong performance in dental CBCT segmentation with an 87% Dice score [27]. Another advantage of hybrid models is their ability to enable the detection of multiple conditions simultaneously. For example, in one study, an ensemble

model based on MobileNetV2 and Swin Transformer achieved over 95% sensitivity and specificity in jointly classifying multiple conditions such as fillings, implants, impacted teeth, and caries on panoramic radiographs [35]. This demonstrates that a complex task that would be difficult for a single architecture to accomplish on its own can be successfully overcome with a hybrid approach. These results indicate that hybrid approaches are particularly advantageous in complex dental imaging tasks that require detailed local analysis and comprehensive global understanding. This synergy is especially powerful for complex segmentation tasks; for instance, precisely delineating a single tooth requires the CNN component to identify its sharp edges and local texture, while the Transformer component helps by contextualizing the tooth's position relative to the alveolar bone and adjacent teeth, preventing anatomically implausible segmentation. Hybrid models can process multi-dimensional data holistically by combining local and global feature extraction, enabling analysis that a single architecture alone could not achieve. In this way, while a small but critical lesion in a dental radiograph is detected, the broader context of the relevant anatomical structure can be evaluated at the same time. The clinical advantage of this combined approach is an increase in diagnostic accuracy and a reduction in the likelihood of false negatives.

In summary, the findings of this review translate into practical guidance for clinical adoption. The choice of architecture should be task-dependent: CNNs remain a computationally efficient choice for well-defined local tasks, whereas ViTs are powerful for complex diagnostics requiring global context. However, the clinical relevance of the model depends on both its accuracy and its applicability. The significant computational costs associated with ViTs may be a significant obstacle to implementation in the real world, often making CNN optimization a more pragmatic option. The success of hybrid models is therefore very important, and it points to a key future direction: the development of architectures that combine the contextual power of transformers with the power of CNNs to create powerful but also practical clinical tools.

Limitations

A notable limitation of the present review is the absence of a formal meta-analysis. The available evidence was too heterogeneous for a valid statistical aggregation. The studies diverged in several critical respects: the clinical task under investigation (classification versus segmentation), dataset size and provenance (predominantly proprietary collections acquired with differing imaging modalities), the performance metrics reported (accuracy, F1-score, Dice coefficient, IoU, etc.), and the specific ways in which CNN and ViT architectures were implemented. Because these variations violate the assumptions required for meaningful pooling and risk generating misleading summary estimates, we elected to provide a structured narrative synthesis that captures overarching trends while respecting the underlying diversity of the data. A further significant limitation, identified through our risk of bias assessment, is the methodological quality of the primary studies. The majority of the included literature relied on small, private datasets and lacked external validation, which poses a high risk to the generalizability of the reported findings. Therefore, while our review identifies a trend favouring ViT-based architectures, these results should be interpreted with caution until they are replicated on larger, more diverse, and publicly available datasets with robust external validation. Restricting inclusion to English-language publications may have introduced language bias, and our search strategy may have missed relevant studies in the gray literature. Excluding studies without statistical data—required for quantitative assessment—may have precluded relevant qualitative or theoretical contributions.

Conclusions

This systematic review suggests that ViT-based deep learning models tend to exhibit higher performance in many dental image analysis scenarios compared to traditional convolutional neural networks. The findings reveal that especially in classification and segmentation tasks, ViT architectures can provide higher accuracy, sensitivity, and in most cases better generalization ability. However, CNN-based models still provide a strong performance benchmark in dental imaging and continue to be effective in tasks where data are limited or local features are determinative. Therefore, the choice of model should be adapted to the nature of the task and the characteristics of the dataset. The overall picture indicates a paradigm shift in dental imaging, but in practice it suggests that CNN and ViT approaches can be used in a complementary manner rather than in competition with each other.

Future research should focus on innovative hybrid architectures combining CNN and Transformer components, few-shot learning approaches, and advanced transfer learning techniques to improve both model performance and data efficiency. In conclusion, when appropriately developed and applied, these advanced deep learning–based models can significantly increase diagnostic accuracy and speed in the field of dental imaging, thereby providing a positive contribution to patients' clinical care.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12903-025-07036-5.

Supplementary Material 1. Database-Specific Search Strategies.

Supplementary Material 2. Summary of Qualitative Risk of Bias Assessment

Supplementary Material 3.

Registration and protocol
The protocol for this systematic review was retrospectively registered with OSF Registries on September 3, 2025, and is publicly available at osf.io/kgnav.

Author contributions
RŞG, TF and HT made significant contributions to the planning and design of the study. The literature search, study selection, and quality assessment were conducted by RŞG, TF and HT supervised the overall study. RŞG, TF and HT participated in the manuscript writing process. All authors critically reviewed the manuscript, agreed on its final version for publication, and accepted responsibility for all aspects of the study.

Funding
No financial support was received for this study.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
This systematic review was conducted based on the analysis of previously published studies. Since no primary data collection involving human participants was performed, ethical approval was not required.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 7 July 2025 / Accepted: 22 September 2025
Published online: 15 October 2025

References

1. Singh NK, Raza K. Progress in deep learning-based dental and maxillofacial image analysis: a systematic review. Expert Syst Appl. 2022;199:116968.
2. Yu Y. Machine learning for dental image analysis. arXiv. 2016; 1611.09958.
3. Turosz N, Chęcińska K, Chęciński M, Brzozowska A, Nowak Z, Sikora M. Applications of artificial intelligence in the analysis of dental panoramic radiographs: an overview of systematic reviews. Dentomaxillofac Radiol. 2023;52:20230284.
4. Kuwada C, Ariji Y, Fukuda M, Kise Y, Fujita H, Katsumata A, et al. Deep learning systems for detecting and classifying the presence of impacted supernumerary teeth in the maxillary incisor region on panoramic radiographs. Oral Surg Oral Med Oral Pathol Oral Radiol. 2020;130:464–9.
5. Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK. Medical image analysis using convolutional neural networks: a review. J Med Syst. 2018;42:1–13.
6. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
7. Krois J, Garcia CA, Chaurasia A, Patil R, Chaudhari PK, Gaudin R, et al. Generalizability of deep learning models for dental image analysis. Sci Rep. 2021;11:6102.
8. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;25:1106-14.
9. Putra RH, Doi C, Yoda N, Astuti ER, Sasaki K. Current applications and development of artificial intelligence for digital dental radiography. Dentomaxillofac Radiol. 2022;51:20210197.
10. Schwendicke F, Golla T, Dreher M, Krois J. Convolutional neural networks for dental image diagnostics: a scoping review. J Dent. 2019;91:103226.
11. Takahashi S, Sakaguchi Y, Kouno N, Takasawa K, Ishizu K, Akagi Y, et al. Comparison of vision transformers and convolutional neural networks in medical image analysis: a systematic review. J Med Syst. 2024;48:84.
12. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv. 2020;11929:2020.
13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998-6008.
14. Benz P, Ham S, Zhang C, Karjauv A, Kweon IS. Adversarial robustness comparison of vision transformer and mlp-mixer to cnns. arXiv. 2021;02797:2021.
15. Maurício J, Domingues I, Bernardino J. Comparing vision transformers and convolutional neural networks for image classification: a literature review. App Sci. 2023;13:5521.
16. Murphy ZR, Venkatesh K, Sulam J, Yi PH. Visual transformers and convolutional neural networks for disease classification on radiographs: a comparison of performance, sample efficiency, and hidden stratification. Radiology: Artificial Intelligence. 2022;4:e220012.
17. Snyder H. Literature review as a research methodology: an overview and guidelines. J Bus Res. 2019;104:333–9.
18. Chen RQ, Lee Y, Yan H, Mupparapu M, Lure F, Li J, et al. Leveraging pretrained Transformers for efficient segmentation and lesion detection in Cone-Beam computed tomography scans. J Endod. 2024;50:1505–e15141.
19. Chen P, Zhang J, Jiang Y, Li Y, Song L, Hu F, et al. A new dataset of oral panoramic x-ray images and parallel network using transformers for medical image segmentation. J Stomatol Oral Maxillofac Surg. 2024;125:101700.
20. Ying S, Wang B, Zhu H, Liu W, Huang F. Caries segmentation on tooth X-ray images with a deep network. J Dent. 2022;119:104076.
21. Ying S, Huang F, Shen X, Liu W, He F. Performance comparison of multifarious deep networks on caries detection with tooth X-ray images. J Dent. 2024;144:104970.
22. Çelebi A, Imak A, Üzen H, Budak Ü, Türkoğlu M, Hanbay D, et al. Maxillary sinus detection on cone beam computed tomography images using ResNet and Swin Transformer-based UNet. Oral Surg Oral Med Oral Pathol Oral Radiol. 2024;138:149–61.
23. Vinayahalingam S, van Nistelrooij N, van Ginneken B, Bressem K, Tröltzsch D, Heiland M. Detection of mandibular fractures on panoramic radiographs using deep learning. Sci Rep. 2022;12:19596.
24. Aljabar AA, Raisan Z, Adnan M, Dhou S. A hybrid transfer learning approach to teeth diagnosis using orthopantomogram radiographs. IEEE Access. 2024. https://doi.org/10.1109/ACCESS.2024.3507925.
25. Sifat FA, Arian MSH, Ahmed S, Farook TH, Mohammed N, Dudley J. An application of 3D vision transformers and explainable AI in prosthetic dentistry. Applied AI Letters. 2024;5:e101.
26. Li Y, Zhao H, Yang D, Du S, Cui X, Zhang J. Identification of dental lesions using self-supervised vision transformer in radiographic X-ray images. Comput Aided Des Appl. 2024;21:332–42.
27. Chen Z, Chen S, Hu F. CTA-UNet: CNN-transformer architecture UNet for dental CBCT images segmentation. Phys Med Biol. 2023. https://doi.org/10.1088/1361-6560/acf026.
28. Liu F, Zhu J, Lv B, Yang L, Sun W, Dai Z, et al. Auxiliary segmentation method of osteosarcoma MRI image based on transformer and U-net. Comput Intell Neurosci. 2022;2022:9990092.
29. Zhou X, Yu G, Yin Q, Yang J, Sun J, Lv S, et al. Tooth type enhanced transformer for children caries diagnosis on dental panoramic radiographs. Diagnostics. 2023;13:689.
30. Wang S, Liu J, Li S, He P, Zhou X, Zhao Z, et al. Resnet-transformer deep learning model-aided detection of dens evaginatus. Int J Paediatr Dent. 2025;35(4):708–16.
31. Zhou X, Feng X, Li Q, Yin Q, Yang J, Yu G, et al. Position weighted convolutional neural network for unbalanced children caries diagnosis. IEEE Access. 2023;11:77034–44.
32. Lubbad MAH, Kurtulus IL, Dervis Karaboga, Kilic K, Alper Basturk B, Akay, et al. A comparative analysis of deep Learning-Based approaches for classifying dental implants decision support system. J Imaging Inf Med. 2024;37(5):2559–80.
33. Sheng C, Wang L, Huang Z, Wang T, Guo Y, Hou W, et al. Transformer-based deep learning network for tooth segmentation on panoramic radiographs. J Syst Sci Complex. 2023;36:257–72.
34. Tang H, Liu S, Tan W, Fu L, Yan M, Feng H. Prediction of midpalatal suture maturation stage based on transfer learning and enhanced vision transformer. BMC Med Inform Decis Mak. 2024;24:1–16.

35. Alsakar YM, Elazab N, Nader N, Mohamed W, Ezzat M, Elmogy M. Multi-label dental disorder diagnosis based on MobileNetV2 and Swin transformer using bagging ensemble classifier. Sci Rep. 2024;14:1–23.
36. Kim MJ, Chae SG, Bae SJ, Hwang KG. Unsupervised few shot learning architecture for diagnosis of periodontal disease in dental panoramic radiographs. Sci Rep. 2024;14:1–15.
37. Fernandes FA, Ge M, Chaltikyan G, Gerdes MW, Omlin CW. Preparing for downstream tasks in artificial intelligence for dental radiology: a baseline performance comparison of deep learning models. Dentomaxillofac Radiol. 2025;54:149–62.
38. Zhang X, Guo E, Liu X, Zhao H, Yang J, Li W, et al. Enhancing furcation involvement classification on panoramic radiographs with vision transformers. BMC Oral Health. 2025;25:153.
39. Tyagi K, Pathak G, Nijhawan R, Mittal A. Detecting Pneumonia using Vision Transformer and comparing with other techniques. Proceedings of the 5th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2021. 2021;12–6. https://doi.org/10.1109/ICECA52323.2021.9676146.
40. Gheflati B, Rivaz H. Vision Transformer for Classification of Breast Ultrasound Images. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS. 2022;480–3. https://doi.org/10.1109/EMBC48229.2022.9871809.
41. Xin C, Liu Z, Zhao K, Miao L, Ma Y, Zhu X, et al. An improved transformer network for skin cancer classification. Comput Biol Med. 2022;149:105939.

Publisher's Note

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