When CNNs Outperform Transformers and Mambas: Revisiting Deep Architectures for Dental Caries Segmentation

Aashish Ghimire[1*], Jun Zeng[3*], Roshan Paudel[1], Nikhil Kumar Tomar[1], Deepak Ranjan Nayak[5], Harshith Reddy Nalla[1], Vivek Jha[4], Glenda Reynolds[2], Debesh Jha[1†]

[1]Department of Computer Science, University of South Dakota, USA
[2]Department of Dental Hygiene, University of South Dakota, USA
[3]School of Software Engineering, Chongqing University of Posts and Telecommunications, China
[4]Post Graduate Institute of Medical Education & Research, Chandigarh, India
[5]Malaviya National Institute of Technology, Jaipur, India

Abstract

Accurate identification and segmentation of dental caries in panoramic radiographs are critical for early diagnosis and effective treatment planning. Automated segmentation remains challenging due to low lesion contrast, morphological variability, and limited annotated data. In this study, we present the first comprehensive benchmarking of convolutional neural networks, vision transformers and state-space mamba architectures for automated dental caries segmentation on panoramic radiographs through a DC1000 dataset. Twelve state-of-the-art architectures, including VMUnet, MambaUNet, VMUNetv2, RMAMamba-S, TransNetR, PVTFormer, DoubleU-Net, and ResUNet++, were trained under identical configurations. Results reveal that, contrary to the growing trend toward complex attention based architectures, the CNN-based DoubleU-Net achieved the highest dice coefficient of 0.7345, mIoU of 0.5978, and precision of 0.8145, outperforming all transformer and Mamba variants. In the study, the top 3 results across all performance metrics were achieved by CNN-based architectures. Here, Mamba and transformer-based methods, despite their theoretical advantage in global context modeling, underperformed due to limited data and weaker spatial priors. These findings underscore the importance of architecture-task alignment in domain-specific medical image segmentation more than model complexity. Our code is available at: https://github.com/JunZengz/dental-caries-segmentation.

Introduction

Dental caries is among the most common chronic diseases worldwide (Abdalla, Elsayed, and Ahmed 2022; Lee, Kim, and Jeong 2021). Although largely preventable, dental caries remains a leading cause of tooth loss and oral discomfort (Hirata et al. 2023). Epidemiological data show that untreated dental caries affects over one-third of the global population. In children, tooth decay is the most common chronic dental condition, impacting approximately 514 million individuals worldwide (GBD 2017 Disease and Injury Incidence and Prevalence Collaborators 2018). Therefore, early and accurate detection of carious lesions is crucial. Conventional diagnostic methods, such as visual-tactile examination, intraoral radiography, approximal tooth separation, caries-detection dyes, fiber-optic transillumination (FOTI), and indices such as DMFT, are widely used. These approaches, however, are limited by subjective interpretation, inter-examiner variability, low sensitivity to early enamel lesions, false-positive staining, and difficulties in detecting early-stage or overlapping lesions (Abdalla, Elsayed, and Ahmed 2022; Srilatha et al. 2019; Abdelaziz 2023).

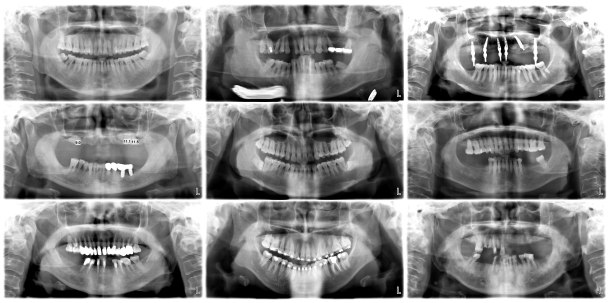

Figure 1: Variation in dental panoramic X-ray structures across the DC1000 dataset, showing patients with full dentition, missing teeth, and braces.

Deep neural networks have transformed medical image analysis in recent years. Convolutional Neural Networks (CNNs) have been especially successful in segmentation tasks (e.g., U-Net (Ronneberger, Fischer, and Brox 2015) and its variants), due to their ability to learn rich hierarchical features from images. More recently, Vision Transformers (ViTs) (Dosovitskiy et al. 2021; Chen et al. 2021) with self-attention mechanisms have emerged, showing promise in capturing global context dependencies and improving long-range reasoning. In parallel, a new family of state-space models has emerged and its medical adaptations, such as VM-UNet (Ruan, Li, and Xiang 2024a), VM-UNetV2 (Zhang et al. 2024) and Rmamba-s (Zeng et al. 2025), aim to balance efficiency and global modeling through selective state-space recurrence.

There are studies in the literature that have exhibited

*These authors contributed equally.
†Corresponding author: debesh.jha@usd.edu.

Table 1: Representative studies on DL approaches for dental segmentation. Reported performance corresponds to each study's best model configuration.

Paper (Year)	Modality	Architecture	Dataset (Size)	Results
Park et al. (2022), *BMC Oral Health* (Park et al. 2022)	Teeth / intraoral photographs	U-Net + ResNet-18 + Faster R-CNN(Combined)	In-house dataset: 2,348 images from 445 patients	**Segmentation-based DL** improved accuracy 0.813, AUC 0.837, Precision 0.874, Sensitivity 0.890.
Zhu et al. (2022), *Neural Comput. Appl.* (Zhu et al. 2023)	Teeth / panoramic radiographs	CariesNet (U-shape CNN with FSAA module)	1,159 images, 3,217 annotated caries lesions	**CariesNet** reached Dice 93.64% and accuracy 93.61% for caries detection.
Asci et al. (2024) (Asci et al. 2024)	Teeth / Pediatric panoramic radiographs	U-Net (CNN)	6,075 images (ages 4–14, mixed dentition)	**U-Net** achieved Sensitivity 0.827, Precision 0.912, F1-score 0.867.
Hao et al. (2024) (Hao et al. 2024)	Teeth / Panoramic radiographs	SemiTNet (Transformer-based, semi-supervised)	TSI15k: 1,589 labeled + 14,728 unlabeled ($\sim$16k total)	**SemiTNet** achieved IoU 94.41% and Dice 95.45% leveraging unlabeled data.
Lim et al. (2025), *Med. Image Anal.* (Lim et al. 2025)	Teeth / Panoramic radiographs	Dual-CNN: Faster R-CNN + U-Net (EfficientNet-B0)	597 images (443 train, 50 val, 100 test)	**Dual-CNN** obtained Dice 0.743, IoU 0.608, Recall 0.731, Precision 0.788.
Damrongsri et al. (2025), *Clin. Oral Imaging* (Pornprasertsuk-Damrongsri et al. 2025)	Teeth / Panoramic radiographs	YOLOv5 + Attention U-Net	500 images, 14,997 teeth (clinically validated GT)	**Attention U-Net** - Dice 0.85, IoU 0.75, Recall 0.96, Precision 0.77, Accuracy 0.93.

Abbreviations: FSAA – Full-Scale Axial Attention; GT – Ground Truth

strong performance. But most of them only use CNNs and do not systematically compare with Transformers, and Mamba-based models. In this field, it remains unclear whether architectural complexity truly translates into better performance for dental applications. To study the research question, we perform a comprehensive benchmark on the DC1000 dataset to rigorously study different segmentation architectures under the same experimental settings. Figure 1 shows the structural variability across the dataset, showcasing cases with full dentition, missing teeth, and orthodontic appliances. These examples highlight the real-world challenges of panoramic radiographs, where anatomical variation, metallic artifacts, and inconsistent contrast complicate automatic caries detection. The main contributions of this work are as follows:

- **Comprehensive benchmark:** This work presents a comprehensive benchmark evaluation of 12 segmentation architectures, spanning from CNNs, transformers, and Mamba-based models, on the DC1000 panoramic radiograph dataset for automated caries segmentation on the DC1000 dataset.
- **Results:** Our analysis reveals contrasting results, where convolutional networks consistently outperform Transformers and Mamba-based methods in dice coefficient, mIoU, precision and inference speed. Here, DoubleU-Net achieved the highest dice coefficient of 0.7345 and mIoU of 0.5978, while Mamba-based RMAMamba-S achieved mDSC of 0.6583 and transformer-based PVT-Former achieved the mDSC of 0.6733, establishing a comprehensive benchmark for dental caries segmentation. The 9% higher mDSC with significantly fewer parameters and faster inference highlights the importance of spatial inductive priors in data-limited medical imaging.
- **Clinical relevance and translational potential:** With this work, we highlight the relevance and translational potential of this problem into computer-aided dental diagnosis. We identify some of the strengths and limitations of current algorithms with qualitative and quantitative results, informing clinicians about key considerations in adopting AI-based diagnostic systems.

Related Works

Deep Learning is one of the fundamental components that aid automated diagnosis of dental imaging, especially for segmenting dental caries in panoramic X-rays. Initial research on the field has focused on convolutional architectures such as U-Net (Ronneberger, Fischer, and Brox 2015) and its variants. For instance, Park et al. (2022) (Park et al. 2022) used a U-Net to segment the tooth surface from Intraoral photograph and then passed the segmented output to the ResNet-18 (He et al. 2015) for caries image classification, and utilized Faster R-CNN (Ren et al. 2015) for the localization of carious lesions with improved accuracy from 0.758 to 0.813, and AUC from 0.731 to 0.837. Similarly, Asci et al.(2024) (Asci et al. 2024) used a U-Net model on more than 6000 pediatric radiographs, indicating consistent performance across primary, mixed, and permanent dentitions. Moreover, to tackle the challenge of automated diagnosis in panoramic x-rays, Hamamci et al. (2023) (Hamamci et al. 2023) introduced the DENTEX benchmark, which is the publicly available hierarchically annotated dataset to support

abnormal tooth detection through panoramic X-rays. However, it does not focus on segmentation tasks, and evaluation relies on metrics like AP/AR, which have known limitations in medical contexts such as class imbalance sensitivity, and interpretability.

Recently, attention mechanisms and transformer-based architectures were introduced in the field to enhance the identification of small or low-contrast carious lesions. Zhu et al. (Zhu et al. 2023) proposed CariesNet, a U-shaped network with full-scale axial attention that outperformed conventional CNNs in multi-stage caries segmentation. Hao et al. (Hao et al. 2024) developed a semi-supervised transformer model called SemiTNet, which achieved state-of-the-art IoU and dice scores by utilizing unlabeled panoramic radiographs. These experiments show a growing transition from traditional CNN-based segmentation approaches to attention-driven and hybrid methods. In 2025, Lim et al. (Lim et al. 2025) integrated Faster R-CNN (Ren et al. 2015) with U-Net to improve lesion-level recall. On the other hand, Pornprasertsuk-Damrongsri et al.(2025) (Pornprasertsuk-Damrongsri et al. 2025) used a two-stage YOLOv5 + Attention U-Net pipeline, validated against bitewing-confirmed data, to achieve high recall for posterior lesions.

Table 1 summarizes representative DL studies on dental image segmentation on various image modalities, detailing their datasets, model types, and reported performance. Furthermore, many studies employ differing protocols and do not offer standardized toolkits or unified benchmarks, thereby limiting reproducibility and generalization. Additionally, the literature indicates that benchmarking efforts on the publicly available DC1000 dataset still remain sparse and lack systematic cross-architectural evaluation. To address these gaps, we present a first unified benchmark of 12 diverse DL models, including CNNs, transformer-based, and Mamba-based architectures. All of our models are evaluated under a consistent training pipeline. Our study aims to serve as a foundational resource guiding the development of robust, generalizable and clinically relevant caries segmentation models.

METHOD

Most of the medical image segmentation methods still depends on CNNs as their underlying foundation. The number of convolutional layers with a gradual reduction of the spatial resolution helps to extract local and global information. Architectures that use encoder-decoder structure like U-Net (Ronneberger, Fischer, and Brox 2015), ResUNet++ (Jha et al. 2019), DoubleU-Net (Jha et al. 2020) and ColonSegNet (Jha et al. 2021) are enhanced with skip connection, and thus are able to reconstruct minute detail. In addition to that, they can learn to increase abstract semantic representation. These designs can be explained as being characterized by effective training dynamics, stability on heterogeneous data, and high spatial resolution, which makes them useful in detecting small or low-contrast caries.

Transformer-based models replace localized convolutional processing with self-attention and in this way, allow

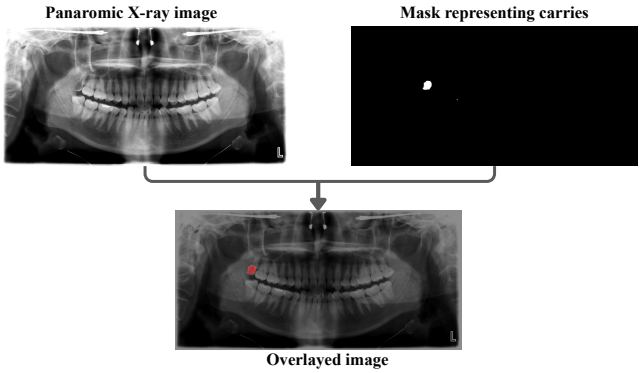

Figure 2: Illustration of dental caries annotation in the DC1000 dataset. The figure shows the original panoramic X-ray image, the expert-annotated binary mask highlighting carious regions, and the overlay visualization combining both.

the network to learn long-range interaction across an image. Canonical models include TransNetR, TransRUPNet, RSAFormer, and PVTFormer divide the image in smaller patches that are then described as tokens, and hierarchical feature constructions is made possible with multiple attention layers (Jha et al. 2024c,a; Yin et al. 2024; Jha et al. 2024b). These models are good in the context of the global context, but they usually demand large amounts of training data and large amounts of computational resources. As a result, they may not be able to reach their full potential in the application to small groups of medical imaging when the fine-scale information within the signal prevails.

The recent mamba-based architectures, such as VMUNet (Ruan, Li, and Xiang 2024b), VMUNetV2 (Zhang et al. 2024), MambaUNet (Wang et al. 2024), and RMAMamba-S (Zeng et al. 2025), provide a different paradigm of visual processing, which makes use of state-space modeling. They do not use attention but instead convert information into a sequence of sequential state transitions that enables the efficient management of long-range dependencies at a lower computing cost. The visual state Mamba model is tested in the current study as a lightweight option to transformer-based architectures. Although it shows some scaling potential, panoramic dental image performance is limited by the small size of the dataset used and the subtle changes in texture depending on the carious area, which highlights the importance of additional domain-specific adaptation.

Experimental setup

Dataset

In this study, we have used DC1000 dataset, which consists of 597 high resolution panoramic images. Each of them were annotated with pixel-level segmentation masks for dental caries(Wang et al. 2023). These radiographs were annotated by experienced dentists, and were obtained from the clinical sources, such that it ensures dataset's reliability and

Table 2: Performance comparison of 12 segmentation models on the DC1000 dataset. Top three results are highlighted in **red** (best), **blue** (second), and **green** (third).

Type	Model	mIoU (↑)	mDSC (↑)	Recall (↑)	Precision (↑)	F2 (↑)	HD (↓)
Mamba-based	VMUNet (Ruan, Li, and Xiang 2024b)	0.4652	0.6114	0.5981	0.6899	0.6003	2.2505
	VMUNetV2 (Zhang et al. 2024)	0.4039	0.5549	0.5516	0.6308	0.5475	2.3620
	MambaUNet (Wang et al. 2024)	0.4302	0.5759	0.6019	0.6503	0.5861	2.3653
	RMAMamba-S(Zeng et al. 2025)	0.5124	0.6583	0.6366	0.7348	0.6424	2.2225
Transformer-based	RSAFormer (Yin et al. 2024)	0.4465	0.5916	0.5880	0.6825	0.5863	2.3834
	TransNetR (Jha et al. 2024c)	0.4946	0.6400	0.6052	0.7411	0.6157	2.2513
	TransRUPNet (Jha et al. 2024a)	0.5278	0.6723	0.6460	0.7631	0.6526	2.1715
	PVTFormer (Jha et al. 2024b)	0.5291	0.6733	0.6564	0.7461	0.6594	2.1637
CNN-based	ResUNet++(Jha et al. 2019)	0.5197	0.6637	0.6463	0.7308	0.6503	2.1420
	ColonSegNet (Jha et al. 2021)	**0.5669**	**0.7044**	**0.6819**	**0.7779**	**0.6877**	**2.0686**
	U-Net (Ronneberger, Fischer, and Brox 2015)	**0.5836**	**0.7236**	**0.7159**	**0.7649**	**0.7161**	**2.0022**
	DoubleU-Net (Jha et al. 2020)	**0.5978**	**0.7345**	**0.7009**	**0.8145**	**0.7115**	**2.0260**

quality (Lee, Kim, and Jeong 2021). In addition to that, this dataset was collected from a wide range of population, as they have variation in caries size, intensity, and location, making it suitable for segmentation task (Ma et al. 2021). In this dataset, we have 497 training images, out of which 4 radiographs do not have mask, and 100 test images. However, this dataset do not provide any official validation split. Images in this dataset are stored in 8-bit grayscale format, and its corresponding masks are encoded as a binary map. Notations like 0 and 1 are used in this data, where 0 represents background and 1 denotes carious lesions. To maintain the balance of data, we have used the provided training set, and optionally, split it further into training and validation subsets at 80:20 ratio.

Experimental setup and configuration

To ensure consistency in runtime, throughput, and reproducibility, all models were trained on the PyTorch framework using a single NVIDIA V100 GPU with 32GB memory. Furthermore, the performance of all models was assessed using standard segmentation metrics, including mean Intersection over Union (mIoU), Dice coefficient(mDSC), Precision, Recall and F2-score. To ensure a fair comparison across architectures, all models were trained and evaluated under a unified experimental configuration using identical data splits, augmentations, and optimization settings. All experiments were conducted using a consistent training pipeline to ensure fair comparison across models.

Data augmentation and preprocessing

Data augmentations included horizontal flipping, random shifts, rotations, mild scaling, contrast-limited adaptive histogram equalization, and brightness–contrast adjustment. Validation images were only resized and normalized. Model performance was evaluated after each epoch. The best model checkpoint was selected based on the highest validation IoU to prevent overfitting. All metrics and losses were logged for quantitative analysis. Results were plotted to visualize convergence trends across different architectures. The training configuration is summarized as follows:

- **Optimizer:** Adam
- **Loss:** Binary Cross-Entropy with logits (pos_weight = 18.0) + Focal loss ($\alpha = 0.75$, $\gamma = 2.0$) combined in a 0.5 : 0.5 ratio
- **Learning Rate:** 1×10^{-4} with ReduceLROnPlateau scheduler (patience $= 5$)
- **Epochs:** 500 with early stopping based on validation Dice (patience = 50)
- **Batch Size:** 4
- **Input Size:** 384×384 (3-channel RGB panoramics)
- **Data Augmentations:**
  - Resize to 384×384
  - Random horizontal flip ($p = 0.5$)
  - Shift, scale, and rotation (shift $\pm 5\%$, scale $\pm 5\%$, rotate $\pm 15°$, border = constant, $p = 0.5$)
  - Random brightness and contrast adjustment ($p = 0.3$)

Results and discussion

Quantitative results

As seen in Table 2, DoubleU-Net achieved the highest segmentation performance over other models, with an mIoU of 0.5978, mDSC of 0.7345, Recall of 0.7009, and Precision of 0.8145 This indicates strong consistency between pixel level overlap and boundary description, highlighting the advantage of its dual-decoder design in capturing minute carious regions. Despite being a 2020 convolutional neural network architecture, DoubleU-Net outperformed more recent transformer-based approaches, suggesting that efficient multi-scale feature aggregation can be highly competitive for dental caries segmentation. Additionally, U-Net ranked second overall with mIoU of 0.5836, mDSC of 0.7236, having balanced recall of 0.7159 and precision of 0.7649, while maintaining a low boundary error, HD of 2.0022. This balance makes U-Net favorable for real-time or near–real-time diagnostic settings. In contrast, ColonSegNet achieved the third-best performance in terms of mIoU of 0.5669 and recall of 0.6819, also showing a third-lowest boundary deviation with HD of 2.0945, suggesting a potential trade-off between segmentation quality and contour precision.

The carious region covers just a fraction of each panoramic radiographs, and hence it amplifies the impact of

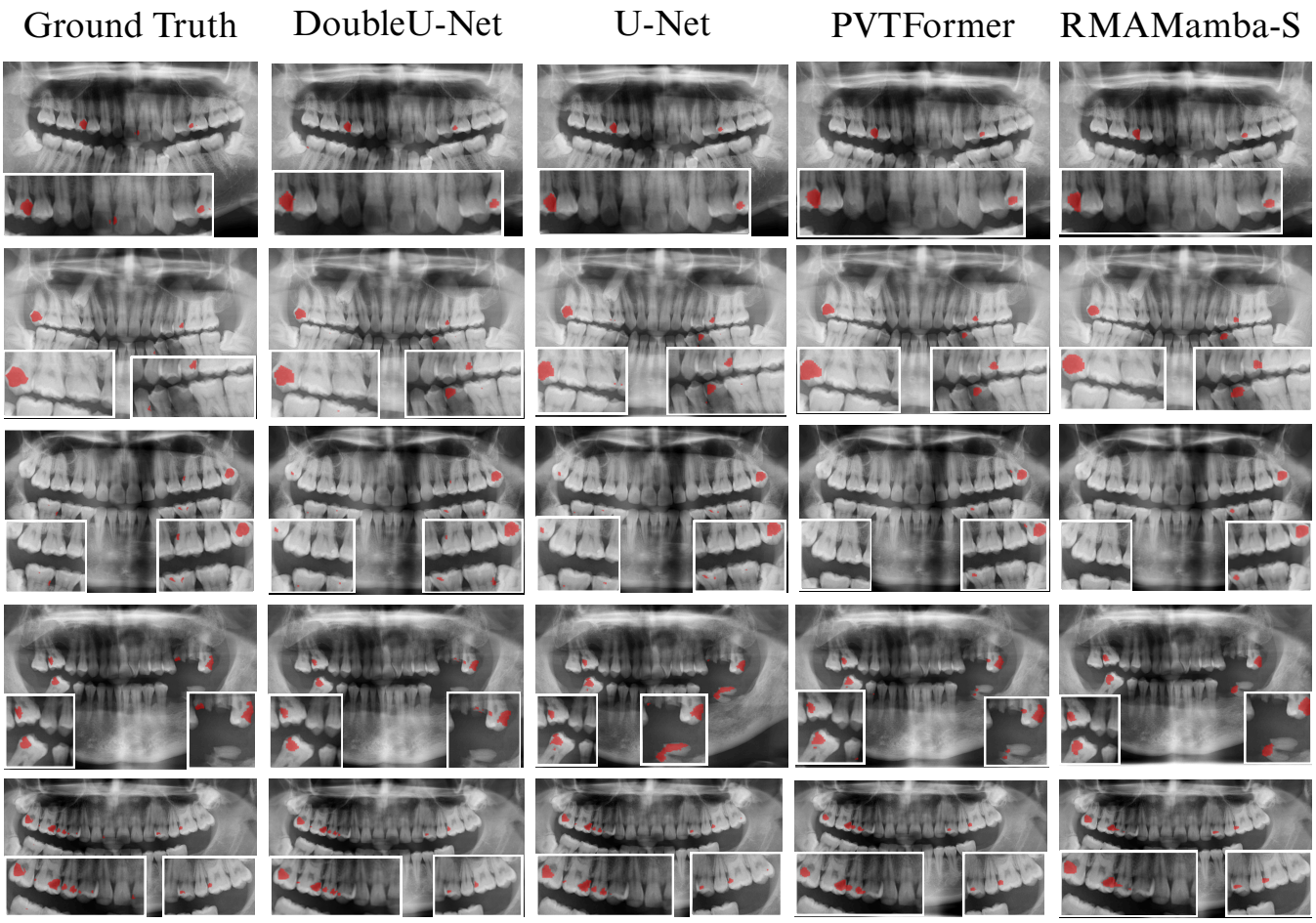

Figure 3: Qualitative comparison of segmentation outputs from four representative architectures on the DC1000 test set. The first column shows the ground truth masks, followed by predictions from DoubleU-Net, U-Net, PVTFormer, and RMAMamba-S. DoubleU-Net consistently yields the most accurate lesion outlines, preserving fine radiolucent boundaries and diffuse enamel–dentin transitions that other models often miss.

Table 3: Computational complexity and efficiency comparison of the evaluated segmentation models on the DC1000 dataset. Lower GFLOPs and fewer parameters indicate higher computational efficiency. The top three FPS values are highlighted in **red**, **blue**, and **green**.

Model	**Architecture Type**	**GFLOPs (↓)**	**Params (M) (↓)**	**FPS (↑)**	**Efficiency Insight**
VMUnet (Ruan, Li, and Xiang 2024b)	Classical encoder–decoder Mamba	**9.25**	22.04	25.17	Low GFLOPs; moderate accuracy.
VMUnetV2 (Zhang et al. 2024)	Vision Mamba UNet with SDI block	**9.90**	17.91	21.02	Complex vision mamba design; poor accuracy.
MambaUnet (Wang et al. 2024)	UNet-like pure visual Mamba	**10.35**	**15.48**	26.24	Baseline Mamba; low parameters.
RMAMamba-S (Zeng et al. 2025)	Reverse-attention Vision Mamba	27.95	55.21	11.43	High parameters; moderate performance.
RSAFormer (Yin et al. 2024)	Attention-enhanced dual decoder	31.69	65.76	9.95	High complexity but poor performance.
TransNetR (Jha et al. 2024c)	Encoder–decoder hybrid Transformer	22.70	15.03	31.17	High GFLOPs; transformer-heavy.
TransRUPNet (Jha et al. 2024a)	Transformer–residual U-Net hybrid	89.40	25.64	23.82	High computational cost; moderate accuracy.
PVTFormer (Jha et al. 2024b)	Hierarchical transformer encoder	97.64	45.51	16.96	Large model; powerful global reasoning.
ResUNet++ (Jha et al. 2019)	Residual U-Net with SE blocks	35.58	**4.06**	**34.22**	Lightweight; strong efficiency with moderate accuracy.
ColonSegNet (Jha et al. 2021)	Lightweight CNN encoder–decoder	139.86	**5.01**	2.31	Computationally heavy due to multi-branch decoding.
U-Net (Ronneberger, Fischer, and Brox 2015)	Classical encoder–decoder CNN	147.43	34.53	**40.83**	Baseline CNN; higher FLOPs despite simple structure.
DoubleU-Net (Jha et al. 2020)	Stacked dual- decoder CNN	121.40	29.29	**33.09**	Balanced trade-off between accuracy and cost.

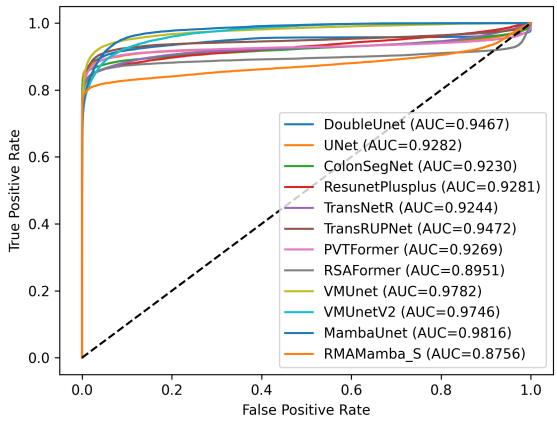

Figure 4: ROC curves comparing pixel-level discriminative performance of several CNN, Transformer, and Mamba architectures.

missing pixels on the mIoU and Dice coefficients. To some extent, this sparsity explains the recall versus precision differences between the models and it highlights the sensitivity of the boundary based measures, such as the Hausdorff Distance. Between the transformer-based models, PVTFormer has the best scores with mIoU of 0.5291, mDSC of 0.6733, closely followed by TransRUPNet with a mIoU of 0.5278 and mDSC of 0.6723, while RSAFormer delivered a slightly lower mIoU of 0.4465 with the highest HD of 2.3834. While discussing about the state-space models, the RMAMamba achieved the best in terms of mIoU of 0.5124 and better precision and recall with the lowest boundary error with HD of 2.2225 among Mamba architectures. To conclude, CNN-based models currently provide better segmentation results, transformer-based models maintain a balance with enhanced global context reasoning, and Mamba models has a compromise between accuracy and efficiency, showing the strengths and trade-offs of each approach.

Qualitative results

Figure 3 shows qualitative results comparison four representative algorithms DoubleUNet (Jha et al. 2020), UNet (Ronneberger, Fischer, and Brox 2015), PVTFormer (Jha et al. 2024b) and RMAMamba-S (Zeng et al. 2025) due to limited space. DoubleU-Net consistently provides the most accurate delineation of carious lesions, preserving both subtle radiolucent boundaries and diffuse enamel–dentin transitions. U-Net shows strong structural localization but slightly under-segments low-contrast regions. PVTFormer captures broader lesion context yet occasionally over-segments due to sensitivity to intensity variations and metallic artifacts. RMAMamba-S produces smoother, less precise boundaries and frequently misses fine lesion details. From a clinical perspective, accurate boundary localization is critical for assessing lesion severity and treatment needs; thus DoubleU-Net's superior performance aligns most closely with clinically actionable segmentation.

Computational efficiency analysis

Table 3 compares the computational efficiency of the evaluated models in terms of GFLOPs, parameter count, and inference speed. Here, Mamba-based architectures exhibit lower GFLOPs and smaller parameter counts overall, with VMUNet of 9.25 GFLOPS, VMUNetv2 of 9.90 GFLOPs. However, they have limited boundary precision, especially in RMAMamba-S. ResUNet++, ColonsegNet and MambaUNet are the smallest model with least number of parameters, with 4.06M, 5.01 M and 15.48M parameters, respectively.

U-Net achieves the fastest inference at 40.83 FPS. ResUNet++ and DoubleUNet achieve second and third with FPS of 34.22 and 33.09, respectively. Thus CNN based architecture exhibits strong efficiency, making it well-suited for real-time clinical workflows. Overall, the efficiency results reinforce that CNN-based models provide the most favorable trade-off between accuracy and computational cost.

ROC Analysis

Figure 4 shows the ROC curve for all the architectures. Here, all models exhibit high discriminative ability ($AUC > 0.87$), indicating that all the methods are capable of reliably classifying individual pixels as carious or non-carious. Among the CNN-based architectures, *DoubleUNet* obtained the highest AUC score of 0.9467. In the Transformer family, *PVTFormer* achieved the highest AUC of 0.9269. Interestingly, Mamba-based architecture performed impressively well with VMUNet obtaining AUC of 0.9782, VmUNetV2 of 0.9782 and MambaUNet obtaining the overall highest AUC of 0.9816.

However, a high AUC does not guarantee accurate segmentation, as segmentation quality depends not only on discriminability but also on the model's ability to produce anatomically meaningful masks. This distinction is important in medical imaging. It explains why Mamba and Transformer models, despite better AUCs, struggle with mDSC and mIoU. Their prediction often shows oversegmentation and fails to capture lesion counters as compared to CNN-based methods. Here, DoubleU-Net and U-Net are able to translate their discriminative ability into much sharper and clinically consistent mask boundaries, maintaining a strong balance between true positives and false positives in spatial context.

Key limitation

Although this research offers a detailed benchmark of the DC1000 dataset, it has certain limitations. While the panoramic radiographs are diverse, they remain relatively small compared to the natural image benchmark. This might be one of the possible reasons why Transformers and Mamba-based models are performing low. The dataset also contains class imbalance and numerous low-contrast or subtle lesions, which may have limited the performance.

Conclusion

This work presents a comprehensive analysis of twelve SOTA CNN, Transformer, and Mamba-based architectures

for dental caries segmentation on the DC1000 panoramic radiograph dataset. Among all the models, the CNN-based DoubleU-Net achieved the highest overall mDSC of 0.7345, mIoU of 0.5978, and Precision of 0.8145. In contrast, Transformer and Mamba variants, despite strong discriminative ability reflected in high AUC values, often fail to translate this into precise boundary localization. These results highlight that architectural success in medical imaging depends less on model complexity and more on spatial inductive priors, data scale, and the ability to capture fine-grained texture cues. The computational analysis further shows the practicality of CNNs, which showed more favorable trade-offs than their Transformer and Mamba counterparts. Future work includes developing more datasets in the field through multi-institutional data collection, and expanding the problem to more classes, such as impacted tooth, periapical radiolucency, etc, in addition to caries and testing our algorithm on more heterogeneous cohorts.

References

Abdalla, A.; Elsayed, A.; and Ahmed, S. 2022. Deep learning for dental caries detection in panoramic radiographs: A systematic review. *BMC Oral Health*, 22(1): 385.

Abdelaziz, M. 2023. Detection, Diagnosis, and Monitoring of Early Caries: The Future of Individualized Dental Care. *Diagnostics*, 13(24): 3649.

Asci, E.; Kilic, M.; Celik, O.; Cantekin, K.; Bircan, H. B.; Bayrakdar, I. S.; and Orhan, K. 2024. A Deep Learning Approach to Automatic Tooth Caries Segmentation in Panoramic Radiographs of Children in Primary Dentition, Mixed Dentition, and Permanent Dentition. *Children (Basel, Switzerland)*, 11(6): 690.

Chen, J.; et al. 2021. TransUNet: Transformers make strong encoders for medical image segmentation. In *MICCAI*.

Dosovitskiy, A.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. 2018. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, 392(10159): 1789–1858.

Hamamci, I. E.; Er, S.; Simsar, E.; Yuksel, A. E.; Gultekin, S.; Ozdemir, S. D.; Yang, K.; Li, H. B.; Pati, S.; Stadlinger, B.; Mehl, A.; Gundogar, M.; and Menze, B. 2023. DENTEX: An Abnormal Tooth Detection with Dental Enumeration and Diagnosis Benchmark for Panoramic X-rays. arXiv:2305.19112.

Hao, J.; Wong, L. M.; Shan, Z.; Ai, Q. Y. H.; Shi, X.; Tsoi, J. K. H.; and Hung, K. F. 2024. A Semi-Supervised Transformer-Based Deep Learning Framework for Automated Tooth Segmentation and Identification on Panoramic Radiographs. *Diagnostics*, 14(17): 1948.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385.

Hirata, A.; Sato, Y.; Hayashi, T.; and Okada, K. 2023. Deep learning-based dental caries detection in panoramic radiographs: A comprehensive evaluation of CNN architectures. *Scientific Reports*, 13: 1145.

Jha, D.; Ali, S.; Tomar, N. K.; Johansen, H. D.; Johansen, D.; Rittscher, J.; Riegler, M. A.; and Halvorsen, P. 2021. Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. *Ieee Access*, 9: 40496–40510.

Jha, D.; Riegler, M. A.; Johansen, D.; Halvorsen, P.; and Johansen, H. D. 2020. DoubleU-Net: A Deep Convolutional Neural Network for Medical Image Segmentation. In *Proceedings of the 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, 558–564.

Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Johansen, D.; De Lange, T.; Halvorsen, P.; and Johansen, H. D. 2019. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*, 225–2255. IEEE.

Jha, D.; Tomar, N. K.; Bhattacharya, D.; Biswas, K.; and Bagci, U. 2024a. TransRUPNet for Improved Polyp Segmentation. In *Proceedings of the 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 1–4.

Jha, D.; Tomar, N. K.; Biswas, K.; Durak, G.; Medetalibeyoglu, A.; Antalek, M.; Velichko, Y.; Ladner, D.; Borhani, A.; and Bagci, U. 2024b. Ct liver segmentation via pvt-based encoding and refined decoding. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 1–5.

Jha, D.; Tomar, N. K.; Sharma, V.; and Bagci, U. 2024c. TransNetR: Transformer-based residual network for polyp segmentation with multi-center out-of-distribution testing. In *Proceedings of the Medical Imaging with Deep Learning*, 1372–1384.

Lee, J. H.; Kim, D.-H.; and Jeong, S. H. 2021. Automated detection and classification of dental caries in panoramic radiographs using deep learning. *Diagnostics*, 11(3): 486.

Lim, Y.-S.; Chun, D.; Kim, J.; Lee, J.-y.; Ju, M. J.; Park, J. H.; and Jeon, H.-J. 2025. Dual Convolutional Neural Network Framework for Segmenting Dental Caries in Panoramic Radiographs. *The Journal of Prosthetic Dentistry*.

Ma, J.; Wang, Y.; An, X.; Ge, C.; Wang, Y.; and Zhang, Y. 2021. Benchmarking Deep Learning Models for Biomedical Image Segmentation: A Comparative Study. *Frontiers in Medicine*, 8: 648240.

Park, E. Y.; Cho, H.; Kang, S.; Jeong, S.; and Kim, E.-K. 2022. Caries detection with tooth surface segmentation on intraoral photographic images using deep learning. *BMC Oral Health*, 22(1): 573.

Pornprasertsuk-Damrongsri, S.; Vachmanus, S.; Papasratorn, D.; Kitisubkanchana, J.; Chaikantha, S.; Arayasantiparb, R.; and Mongkolwat, P. 2025. Clinical Application of Deep Learning for Enhanced Multistage Caries Detection in Panoramic Radiographs. *Scientific Reports*, 15(1): 33491.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241. Springer.

Ruan, J.; Li, J.; and Xiang, S. 2024a. VM-UNet: Vision Mamba UNet for Medical Image Segmentation. *ACM Transactions on Multimedia Computing, Communications, and Applications*.

Ruan, J.; Li, J.; and Xiang, S. 2024b. Vm-unet: Vision mamba unet for medical image segmentation. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Srilatha, A.; Doshi, D.; Kulkarni, S.; and Reddy, M. 2019. Conventional diagnostic aids in dental caries. *Journal of Global Oral Health*, 2.

Wang, X.; Gao, S.; Jiang, K.; Zhang, H.; Wang, L.; Chen, F.; Yu, J.; and Yang, F. 2023. Multi-level uncertainty aware learning for semi-supervised dental panoramic caries segmentation. *Neurocomputing*, 540: 126208.

Wang, Z.; Zheng, J.-Q.; Zhang, Y.; Cui, G.; and Li, L. 2024. Mamba-unet: Unet-like pure visual mamba for medical image segmentation. *arXiv preprint arXiv:2402.05079*.

Yin, X.; Zeng, J.; Hou, T.; Tang, C.; Gan, C.; Jain, D. K.; and García, S. 2024. Rsaformer: A method of polyp segmentation with region self-attention transformer. *Computers in Biology and Medicine*, 172: 108268.

Zeng, J.; Jha, D.; Aktas, E.; Keles, E.; Medetalibeyoglu, A.; Antalek, M.; Lewandowski, R.; Ladner, D.; Borhani, A. A.; Durak, G.; et al. 2025. A reverse mamba attention network for pathological liver segmentation. *Proceedings of the IEEE ICECCME 2025*.

Zhang, M.; Yu, Y.; Jin, S.; Gu, L.; Ling, T.; and Tao, X. 2024. VM-UNET-V2: rethinking vision mamba UNet for medical image segmentation. In *International symposium on bioinformatics research and applications*, 335–346. Springer.

Zhu, H.; Cao, Z.; Lian, L.; et al. 2023. CariesNet: A Deep Learning Approach for Segmentation of Multi-Stage Caries Lesion from Oral Panoramic X-Ray Image. *Neural Computing and Applications*, 35: 16051–16059.