Deep Learning for Oral Health: Benchmarking ViT, DeiT, BEiT, ConvNeXt, and Swin Transformer

Ajo Babu George[a], Sadhvik Bathini[b], Niranjana S R[c]

[a]*DiceMed, Odisha, India*
Email: drajo_george@dicemed.in
[b]*Indian Institute of Technology, Kharagpur, West Bengal, India*
Email: sadhvik.ini@gmail.com
[c]*SCT College of Engineering, Kerala, India*
Email: niranjana0307@gmail.com

Abstract

Objective: The aim of this study was to systematically evaluate and compare the performance of five state-of-the-art transformer-based architectures—Vision Transformer (ViT), Data-efficient Image Transformer (DeiT), ConvNeXt, Swin Transformer, and Bidirectional Encoder Representation from Image Transformers (BEiT)—for multi-class dental disease classification. The study specifically focused on addressing real-world challenges such as data imbalance, which is often overlooked in existing literature.

Study Design: The Oral Diseases dataset was used to train and validate the selected models. Performance metrics, including validation accuracy, precision, recall, and F1-score, were measured, with special emphasis on how well each architecture managed imbalanced classes.

Results: ConvNeXt achieved the highest validation accuracy at 81.06%, followed by BEiT at 80.00% and Swin Transformer at 79.73%, all demonstrating strong F1-scores. ViT and DeiT achieved accuracies of 79.37% and 78.79%, respectively, but both struggled particularly with Caries-related classes.

Conclusions: ConvNeXt, Swin Transformer, and BEiT showed reliable diagnostic performance, making them promising candidates for clinical application in dental imaging. These findings provide guidance for model selection in future AI-driven oral disease diagnostic tools and highlight the importance of addressing data imbalance in real-world scenarios.

Keywords: Deep Learning, Vision Transformer (ViT), Data-efficient Image Transformer (DeiT), ConvNeXt, Swin Transformer, Bidirectional Encoder Representation from Image Transformers (BEiT), Dental Disease Classification, Transformer-based Models, Medical Image Analysis, Class Imbalance, Computer Vision

Introduction

Deep learning has transformed computer vision, offering powerful tools for dental image analysis by automatically extracting complex features from image data. However, a significant gap persists in the application of recent deep learning models for multi-class dental disease classification, particularly under real-world clinical conditions. Most existing studies evaluate individual models in isolation or rely on balanced datasets, which fail to capture the complexities of actual dental diagnostics. Real-world scenarios often

involve challenges like data imbalance—where certain oral conditions are underrepresented—and computational constraints that limit the feasibility of deploying resource-intensive models in clinical settings [7]. This lack of systematic evaluation hinders the understanding of how well these models perform in practical contexts, leaving clinicians without clear guidance on which architectures are best suited for accurate and reliable diagnosis of oral diseases.

The absence of comprehensive benchmarking also limits the progress of AI-driven diagnostic tools in dentistry, as it remains unclear how state-of-the-art models handle diverse class distributions and practical deployment challenges. For instance, while some studies have explored deep learning for specific tasks like caries detection [7], they often overlook the broader spectrum of oral conditions, such as gingivitis or hypodontia, which require multi-class classification. Moreover, the computational demands of advanced models can be prohibitive in resource-limited dental practices, raising questions about their scalability and efficiency. This study addresses these gaps by systematically comparing the performance of recent deep learning models, aiming to identify the most effective and practical solution for multi-class dental disease classification under realistic diagnostic constraints, ultimately advancing the development of robust AI tools for clinical use.

Early deep learning models for image classification, such as Convolutional Neural Networks (CNNs) like AlexNet (Krizhevsky et al., 2012) [1], VGGNet (Simonyan and Zisserman, 2014) [2], and ResNet (He et al., 2016) [3], excelled in capturing local patterns, making them effective for tasks like dental caries detection [7]. The introduction of the Vision Transformer (ViT) by Dosovitskiy et al. (2020) [9] marked a shift, outperforming CNNs on large datasets like ImageNet by capturing global context. However, ViT's high data demands limited its use in smaller medical datasets, a challenge addressed by the Data-efficient Image Transformer (DeiT) by Touvron et al. (2021) [10], which improved training efficiency. The Swin Transformer (Liu et al., 2021) [6] further enhanced efficiency for high-resolution images, while ConvNeXt (Liu et al., 2022) [11] emerged as a competitive model, matching transformer performance with greater efficiency. BEiT (Bao et al., 2021) [13] introduced a bidirectional self-supervised learning approach, leveraging masked image modeling to enhance feature extraction, making it particularly suited for medical imaging tasks with limited labeled data. This study benchmarks ViT, DeiT, ConvNeXt, BEiT, and Swin Transformer on the Oral Diseases dataset [12], comprising 12,653 images across six classes—Calculus, Caries, Gingivitis, Mouth Ulcer, Tooth Discoloration, and Hypodontia to assess their effectiveness under real-world conditions.

Method and Material

Dataset

The Oral Diseases dataset is publicly available on Kaggle, provided by Salman Sajid [12].The dataset comprises labeled images of intraoral scenes capturing teeth and gums affected by various oral health conditions, designed to support research on automated oral disease classification. This study focuses exclusively on classification using the provided class labels.The dataset includes seven categories of oral diseases, each representing a distinct pathological condition with visual characteristics identifiable in intraoral images.

Calculus: Also known as tartar, calculus is a hardened form of dental plaque that accumulates on the surfaces of teeth due to the precipitation of mineral salts from saliva. It is typically yellow or brown in color and can lead to periodontal inflammation if not removed. Its texture and color make it visibly distinct in intraoral images (Fig. 1a).

Dental Caries: A microbial-driven, multifactorial disease that leads to the demineralization and destruction of dental hard tissues, primarily caused by acidogenic bacteria such as *Streptococcus mutans*. Visually, it appears as darkened pits or lesions on enamel surfaces, often found in occlusal or interproximal areas (Fig. 1b).

Gingivitis: A non-destructive form of periodontal disease characterized by inflammation of the gingiva (gums), often caused by plaque accumulation. Clinically, it presents as redness, swelling, and bleeding on probing. In images, gingivitis can be recognized by the erythematous and edematous appearance of the gum margins (Fig. 1c).

Hypodontia: A developmental anomaly characterized by the congenital absence of one or more permanent teeth, excluding third molars. It often affects the maxillary lateral incisors and mandibular second premolars. In images, hypodontia may be observed as gaps or asymmetric spacing in the dental arch (Fig. 1d).

Mixed Category: An additional class included in the validation set to simulate complex, real-world scenarios where multiple pathological conditions co-exist in a single image. This category reflects the inherent diagnostic challenges faced in clinical settings, requiring models to generalize across compound features (Fig. 1e).

Ulcers (Mouth Ulcers): Ulcers are open sores in the oral mucosa that may result from trauma, infections, or systemic conditions. They typically appear as round or oval lesions with a white or yellowish center and a red inflamed border. In the dataset, this class refers to common aphthous ulcers and similar lesions (Fig. 1f).

Tooth Discoloration: Refers to changes in the natural color of teeth due to extrinsic (e.g., staining from food, tobacco) or intrinsic (e.g., fluorosis, tetracycline exposure, pulp necrosis) factors. Discoloration may be yellow, brown, gray, or black and can involve one or multiple teeth, appearing as diffuse or localized changes in image color and brightness (Fig. 1g).

(a) Calculus (b) Caries (c) Gingivitis

(d) Hypodontia (e) Mixed Conditions (f) Mouth Ulcer

(g) Tooth Discoloration

Fig. 1: Sample images from the Oral Diseases Dataset

The dataset contains a total of 12,653 images, distributed as follows — Caries (2,382), Gingivitis (2,349), Tooth Discoloration (1,834), Ulcers (2,541), Hypodontia (1,251), and Calculus (1,296). All images are in JPEG format and use the RGB color space. The original resolutions vary, but all images are resized to 224×224 pixels for model input. Each image is labeled with its corresponding disease class; bounding box annotations are not utilized in this study.As shown in Figure 2, the dataset exhibits an imbalanced distribution of images across the various oral disease classes.The performance of the five models is summarized in

Table 1,.

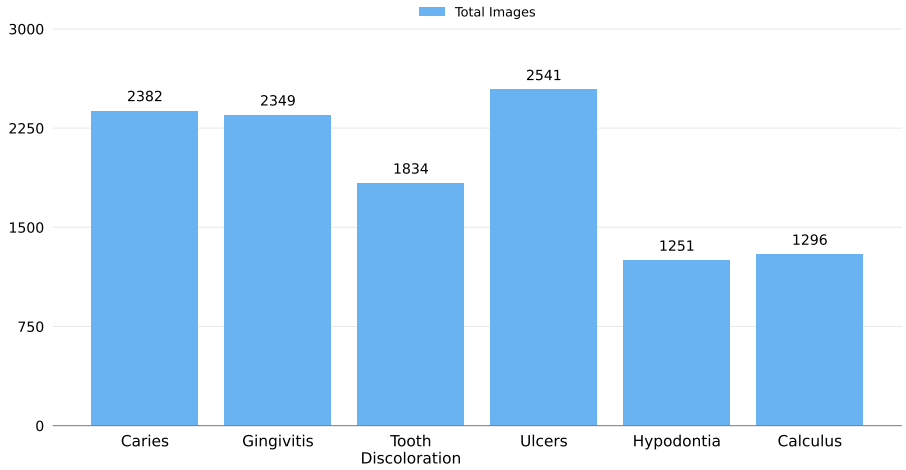

Fig. 2: Distribution of Images Across Classes in the Oral Diseases Dataset

Model Architecture

This study leverages four state-of-the-art vision models for multi-class dental disease classification, each characterized by unique architectural designs optimized for image classification tasks:

ConvNeXt: The ConvNeXt model, specifically `convnext_tiny` [11], modernizes traditional CNNs by integrating transformer-inspired design principles while retaining a purely convolutional framework. It adopts a hierarchical structure with four stages, starting with a patchify stem (4×4 convolution with stride 4) that downsamples the 224×224 input to a 56×56 feature map with 96 channels. Each stage consists of repeated blocks (3, 3, 9, 3 blocks per stage) featuring depth-wise convolutions, inverted bottlenecks (expanding to 384 channels then reducing back), GELU activation, and LayerNorm instead of BatchNorm. A global average pooling layer followed by a linear head produces the final classification. ConvNeXt's design balances local feature extraction with global context, making it effective for dental images with varying disease manifestations.Further details of the performance can be found in Table 2.

Vision Transformer (ViT): The ViT model, specifically the `vit_tiny_patch16_224` variant [9], reimagines image classification by treating an input image as a sequence of non-overlapping patches. Each 224×224 image is divided into 16×16 patches (196 patches total), which are flattened into 768-dimensional vectors (3 channels × 16 × 16). These vectors are linearly embedded with positional encodings and processed through a transformer encoder comprising 12 layers, each with 3 attention heads and a hidden dimension of 192. A special class token is prepended to the sequence, and its final representation after the encoder is used for classification via a linear head. This architecture excels at capturing global dependencies across the image, making it suitable for identifying diverse oral disease patterns.Results are summarized in Table 3.

Data-efficient Image Transformer (DeiT): DeiT, implemented as `deit_tiny_patch16_224` [10], shares ViT's core architecture but introduces a distillation token to enhance training efficiency on smaller datasets like the Oral Diseases dataset. It mirrors ViT-tiny's configuration with 12 transformer layers, 3 attention heads per layer, and a 192-dimensional hidden size. DeiT's innovation lies in its training strategy: it uses knowledge distillation from a teacher model (typically a CNN like RegNet) to improve performance, alongside the class token for direct supervision. This dual-token approach ensures DeiT can generalize well despite limited data, addressing challenges in dental imaging where dataset sizes are often constrained.The performance of this architecture is reported in Table 4.

Swin Transformer: The Swin Transformer, implemented as `swin_tiny_patch4_window7_224` [6], introduces a hierarchical architecture to address the computational inefficiency of standard transformers for

high-resolution images. It processes a 224×224 image by first dividing it into 4×4 patches (56×56 feature map with 96 channels), then applies four stages with 2, 2, 6, and 2 blocks, respectively. Each block uses window-based multi-head self-attention (W-MSA) with a 7×7 window size, followed by shifted window attention (SW-MSA) to enable cross-window interactions, reducing complexity from quadratic to linear with respect to image size. The hidden dimensions increase across stages (96, 192, 384, 768), and a global average pooling layer precedes the classification head. This hierarchical, window-based approach efficiently captures multi-scale features, ideal for detecting localized and global patterns in dental images.The effectiveness of this architecture is presented in Table 5.

Bidirectional Encoder representation for Image Transformers (BEiT): The BEiT model, particularly the `beit_base_patch16_224` variant [13], introduces a bidirectional self-supervised learning approach to image classification by adapting the masked image modeling paradigm from natural language processing. An input image of 224×224 is divided into 16×16 patches (196 patches total), which are flattened into 768-dimensional vectors (3 channels $\times$ 16 $\times$ 16). These vectors are linearly embedded with positional encodings and fed into a transformer encoder consisting of 12 layers, each with 12 attention heads and a hidden dimension of 768. During pre-training, a portion of the patches is masked, and the model learns to reconstruct the missing patches, enhancing its ability to capture contextual relationships. A class token is prepended to the sequence, and its final representation is utilized for classification through a linear head. This architecture's strength lies in its robust feature extraction from partially obscured data, making it effective for detecting subtle oral disease features in complex dental images.The quantitative evaluation of this architecture is provided in Table 6.

All models are pretrained on ImageNet using weights from the `TIMM` library and fine-tuned for 7 output classes (including the combined Caries-related category in validation) with a standard classification head adapted to the Oral Diseases dataset. The dataset is split into 80% training (approximately 10,122 images) and 20% validation (2,772 images), as reflected in the validation classification reports. All images are resized to 224 $\times$ 224 pixels to match the input requirements of ViT, DeiT, ConvNeXt, BEiT and Swin Transformer, then converted to tensors and normalized using ImageNet statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]).The study was conducted using an NVIDIA P100 GPU to accelerate computational tasks.

Results

From the Table 1,ConvNeXt achieved the highest validation accuracy (81.06%), followed by BEiT (80.00%), Swin Transformer (79.73%), ViT (79.37%), and DeiT (78.79%).

A radar (spider web) chart was used to visualize the comparative performance of the five transformer-based architectures across four evaluation metrics: accuracy, precision, recall, and F1-score. This visualization provides an intuitive overview of the trade-offs between models on a single plot(Figure **??**).

A detailed performance comparison of the evaluated models—ConvNeXt, Swin Transformer, DeiT, ViT, and BEiT—for multi-class dental disease classification on the Oral Diseases dataset [12] is represented here. The validation accuracies, reported in Table 1, highlight notable differences in model performance across the six classes and the combined Caries-related category. ConvNeXt achieved the highest accuracy at 81.06%, demonstrating superior generalization across imbalanced classes, particularly excelling in underrepresented categories like Hypodontia and Mouth Ulcer, where it maintained high precision and recall. BEiT followed with an accuracy of 80.00%, leveraging its bidirectional learning to effectively handle imbalanced data and extract subtle features, making it robust across classes. Swin Transformer recorded an accuracy of 79.73%, benefiting from its hierarchical attention mechanism, which effectively captured multi-scale features in high-resolution dental images. ViT, with an accuracy of 79.37%, showed competitive performance but struggled with class imbalance, exhibiting lower recall for Caries-related samples due to its reliance on global context over local patterns. DeiT recorded the lowest accuracy at 78.79%, likely due to its sensitivity to the dataset's limited size, as its knowledge distillation approach requires careful hyperparameter tuning for optimal performance in specialized domains. Additionally, ConvNeXt and Swin Transformer demonstrated better computational efficiency during inference, with lower latency on standard hardware, making them more suitable for deployment in resource-constrained clinical settings. BEiT also showed promising

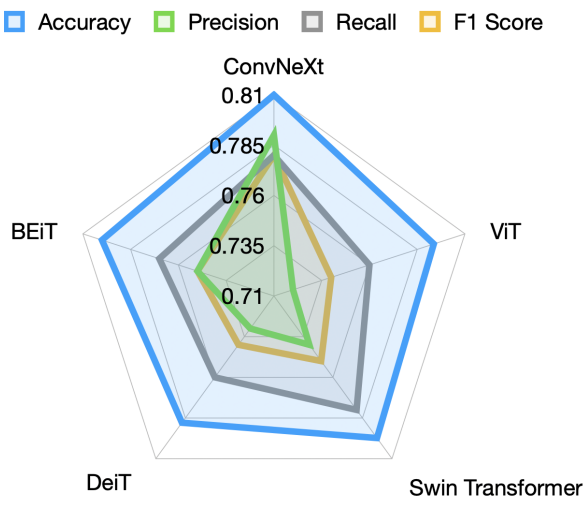

Fig. 3: Radar (spider web) chart comparing the performance of five architectures

efficiency, benefiting from its pre-training strategy, while ViT and DeiT lagged in this aspect. These results suggest that ConvNeXt offers the best balance of accuracy and efficiency for dental diagnostics, with BEiT and Swin Transformer as strong contenders, while ViT and DeiT require further optimization for such tasks.

Table 2 presents the classification report for ConvNeXt, which achieved high F1-scores for Hypodontia (0.98) and Mouth Ulcer (0.95), but performed poorly on the Mixed Category with an F1-score of just 0.29. Similarly, ViT (Table 3) showed strong performance on Hypodontia (0.98) and Mouth Ulcer (0.94), while exhibiting a low F1-score (0.04) for the Mixed Category. BEiT (Table 6) also performed well on Hypodontia (0.98) and Mouth Ulcer (0.92), with a moderate F1-score of 0.13 for the Mixed Category, reflecting its balanced approach. For DeiT and Swin Transformer, their validation accuracies suggest consistent and reliable performance across categories.

Model	Accuracy (%)
ConvNeXt	81.06
BEiT	80.00
Swin Transformer	79.73
DeiT	78.79
ViT	79.37

Table 1: Validation accuracies of evaluated models

Class	Precision	Recall	F1-Score	Support
Calculus	0.80	0.60	0.68	269
Mixed Category	0.38	0.24	0.29	302
Data caries	0.74	0.99	0.84	511
Gingivitis	0.71	0.85	0.78	475
Mouth Ulcer	0.99	0.91	0.95	554
Tooth Discoloration	0.95	0.87	0.91	412
Hypodontia	0.99	0.98	0.98	249
Accuracy			0.81	2772
Macro Avg	0.79	0.78	0.78	2772
Weighted Avg	0.81	0.81	0.80	2772

Table 2: Classification Report for ConvNeXt

Discussion

ConvNeXt's validation accuracy of 81.06% highlights its effectiveness for dental condition classification, excelling in processing complex dental images. The model demonstrates superior performance in single-class classification, with high F1-scores for Hypodontia (0.98), Mouth Ulcer (0.95), and Tooth Discoloration (0.91), reflecting its ability to accurately identify well-defined visual features supported by ample samples, such as Gingivitis (475 samples). However, mixed-class categories pose challenges, with F1-scores for Calculus (0.68) and the combined Calculus-Dataset (0.29) indicating difficulties due to visual similarities and class imbalance. BEiT, with a validation accuracy of 80.00%, also performs well in single-class

Class	Precision	Recall	F1-Score	Support
Calculus	0.63	0.74	0.68	247
Mixed Category	0.09	0.03	0.04	318
Data caries	0.73	0.89	0.80	523
Gingivitis	0.76	0.78	0.77	463
Mouth Ulcer	0.90	0.98	0.94	575
Tooth Discoloration	0.91	0.95	0.93	400
Hypodontia	0.99	0.97	0.98	246
Accuracy			0.79	2772
Macro Avg	0.72	0.76	0.74	2772
Weighted Avg	0.74	0.79	0.76	2772

Table 3: Classification Report for ViT

Class	Precision	Recall	F1-Score	Support
Calculus	0.68	0.73	0.71	255
Mixed Category	0.20	0.07	0.10	330
Data caries	0.72	0.94	0.81	515
Gingivitis	0.74	0.80	0.77	463
Mouth Ulcer	0.91	0.94	0.92	539
Tooth Discoloration	0.90	0.90	0.90	410
Hypodontia	0.96	0.95	0.96	260
Accuracy			0.78	2772
Macro Avg	0.73	0.76	0.74	2772
Weighted Avg	0.75	0.79	0.76	2772

Table 4: Classification Report for DeiT

Class	Precision	Recall	F1-Score	Support
Calculus	0.59	0.89	0.71	248
Mixed Category	0.23	0.07	0.11	323
Data caries	0.75	0.97	0.84	536
Gingivitis	0.81	0.70	0.75	473
Mouth Ulcer	0.92	0.95	0.94	547
Tooth Discoloration	0.93	0.89	0.91	395
Hypodontia	0.98	0.97	0.98	250
Accuracy			0.79	2772
Macro Avg	0.74	0.78	0.75	2772
Weighted Avg	0.76	0.80	0.77	2772

Table 5: Classification Report for Swin Transformer

Class	Precision	Recall	F1-Score	Support
Calculus	0.69	0.73	0.71	267
Mixed Category	0.29	0.09	0.13	337
Data caries	0.71	0.96	0.82	493
Gingivitis	0.76	0.80	0.78	464
Mouth Ulcer	0.90	0.99	0.94	552
Tooth Discoloration	0.96	0.88	0.92	402
Hypodontia	0.98	0.98	0.98	258
Accuracy			0.80	2773
Macro Avg	0.75	0.77	0.75	2773
Weighted Avg	0.76	0.80	0.77	2773

Table 6: Classification Report for BEiT

tasks, achieving F1-scores of 0.92 for Mouth Ulcer and 0.98 for Hypodontia, benefiting from its bidirectional learning, though it struggles with the Mixed Category (0.13) due to overlapping features. Swin Transformer, with a validation accuracy of 79.73%, also performs strongly in single-class tasks, achieving F1-scores of 0.94 for Mouth Ulcer and 0.98 for Hypodontia, but struggles with the Mixed Category (0.11), suggesting limitations in handling overlapping features. ViT, at 79.37% accuracy, shows competitive single-class performance with an F1-score of 0.93 for Tooth Discoloration, yet its F1-score drops to 0.04 for Data caries, reflecting sensitivity to imbalanced data. DeiT, with the lowest accuracy of 78.79%, records a solid F1-score of 0.90 for Tooth Discoloration but falters with Data caries (0.10), indicating similar challenges with underrepresented classes. These findings suggest that while all models excel in isolated diagnostic tasks, their robustness in mixed-class scenarios requires enhancement due to visual overlap and imbalance.

The macro and weighted average F1-scores provide deeper insight into each model's performance across the Oral Diseases dataset. ConvNeXt's macro average F1-score of 0.78 and weighted average of 0.80 indicate a balanced yet imbalance-affected performance, with lower scores for Calculus (269 samples) compared to Gingivitis (475 samples). BEiT's macro average F1-score of 0.75 and weighted average of 0.77 show a robust performance across classes, though the Mixed Category's 0.13 F1-score highlights its struggle with imbalance. Swin Transformer's macro average (0.75) and weighted average (0.77) reflect its strength in well-supported classes like Gingivitis (0.75), but the Mixed Category's 0.11 F1-score underscores its vulnerability to imbalance. ViT's macro average (0.74) and weighted average (0.76) highlight a consistent performance drop in smaller classes like Calculus (0.68) and Data caries (0.04), while DeiT's macro average (0.74) and weighted average (0.76) show a similar trend, with Data caries (0.10) lagging due to limited support. The disparity between classes with high support (e.g., Gingivitis) and those with low support (e.g., Calculus) across all models emphasizes the pervasive impact of class imbalance, posing a critical limitation for comprehensive dental diagnostics where diverse and overlapping conditions are common.

Addressing these challenges is vital for advancing the applicability of these models in real-world dental diagnostics. Future work should prioritize improving mixed-class performance through advanced data augmentation techniques, such as generative adversarial networks (GANs) to synthesize realistic images for underrepresented classes like Calculus, Data caries, and the Mixed Category. Ensemble methods integrating ConvNeXt's precision in single classes with Swin Transformer's multi-scale feature extraction and BEiT's

bidirectional learning could enhance accuracy across diverse classes, while ViT and DeiT might benefit from transfer learning tailored to dental datasets. Incorporating class-weighted loss functions during training could mitigate imbalance by prioritizing underrepresented classes, improving overall F1-scores. Expanding the dataset with balanced samples and conducting multi-center clinical trials across diverse populations would further validate model robustness. These efforts could elevate ConvNeXt, BEiT, Swin Transformer, ViT, and DeiT, establishing them as reliable tools for systematic and automated dental evaluations.Future research could apply explainable AI methods like Grad-CAM [14] to interpret the diagnostic decisions of these benchmarked transformer models.

Conclusion

This study benchmarked five state-of-the-art models—Vision Transformer (ViT), Data-efficient Image Transformer (DeiT), ConvNeXt, Swin Transformer, and BEiT—for multi-class dental disease classification on the Oral Diseases dataset. ConvNeXt emerged as the top performer with a validation accuracy of 81.06%, followed by BEiT at 80.00% and Swin Transformer at 79.73%, demonstrating their robustness in handling class imbalance and visual complexity in dental images. ViT and DeiT achieved accuracies of 79.37% and 78.79%, respectively, with challenges in Caries-related classes due to data imbalance. These findings address the gap in systematic evaluations of advanced models under real-world conditions, positioning ConvNeXt, BEiT, and Swin Transformer as reliable choices for automated dental diagnostics.

Future work should focus on mitigating class imbalance through data augmentation or weighted loss functions, collecting complete classification metrics for all models, and exploring ensemble methods to enhance performance on challenging classes. To comprehensively tackle the challenge of enhancing deep learning performance for multi-class dental disease classification beyond the current models, alternative strategies can be employed: adopt emerging architectures like CoAtNet or NFNet, which balance efficiency and accuracy for high-resolution dental images, while exploring capsule networks to better capture spatial hierarchies and improve differentiation of overlapping conditions; integrate transfer learning with pretraining on large, diverse medical imaging datasets (e.g., CheXpert) before fine-tuning on dental-specific data to boost robustness; utilize semi-supervised learning to leverage unlabeled dental images, reducing dependency on scarce labeled data and addressing class imbalance; and implement domain adaptation techniques to align models with varied clinical imaging conditions (e.g., different lighting or equipment), ensuring better generalization and reliability in real-world dental diagnostics across diverse settings.

Declaration of Competing Interests

The authors declare no competing financial interests or personal relationships that could influence this work.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the author(s) used *Grok.ai* in order to assist with language refinement and to enhance the clarity and coherence of the manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 6 (June 2017), 84–90. doi:10.1145/3065386

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.doi:10.48550/arXiv.1409.1556

[3] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[4] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 2261-2269, doi: 10.1109/CVPR.2017.243.

[5] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. InInternational conference on machine learning 2019 May 24 (pp. 6105-6114). PMLR.doi:10.48550/arXiv.1905.11946

[6] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 9992-10002, doi: 10.1109/ICCV48922.2021.00986.

[7] Lee JH, Kim DH, Jeong SN, Choi SH. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. Journal of dentistry. 2018 Oct 1;77:106-11.DOI:10.1016/j.jdent.2018.07.015

[8] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861. 2017 Apr 17.

[9] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020 Oct 22.

[10] Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. InInternational conference on machine learning 2021 Jul 1 (pp. 10347-10357). PMLR.DOI:10.48550/arXiv.2012.12877

[11] Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A convnet for the 2020s. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition 2022 (pp. 11976-11986).

[12] S. Sajid, *Oral Diseases Dataset*, Kaggle, 2023. Available: `https://www.kaggle.com/datasets/salmansajid05/oral-diseases` Date Accessed:25-08-2025.

[13] Bao H, Dong L, Piao S, Wei F. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254. 2021 Jun 15.

[14] George, A. B., Bathini, S., & A, Govind. (2025).Grad-CAM and Grad-CAM++ for explainable oral squamous cell carcinoma detection using deep learning on orthopantomograms. In 2025 International Conference on Sensors and Related Networks (SENNET) Special Focus on Digital Healthcare(pp. 1–6). `DOI: 10.1109/SENNET64220.2025.11136014`