Dental Vision

Madhavan Balaji, Raghuveer Venkatesh, Vigneshwar Ravi Rao

1. Introduction

Oral health, affecting over 3.5 billion people globally, is often neglected despite its link to serious conditions like tooth loss and systemic diseases. Challenges such as limited specialist access and subjective diagnosis delay treatment. AI offers a solution by enabling faster, more accurate detection of oral diseases through dental imagery. This project aims to develop an AI-based diagnostic tool to improve efficiency, accuracy, and access to dental care.

2. Problem description

Oral diseases can significantly impact general health and quality of life, yet they often go undiagnosed in their early stages due to limited access to professional screening. Manual diagnosis from dental images is time-consuming and prone to human error. This project addresses the challenge by developing an automated image classification system using deep learning models to identify various oral conditions from dental images. The goal is to support dental professionals in diagnosis and to enable preliminary remote screening via a web interface. You can access the demo here, dentavision.onrender.com.

3. Description of the data

In this project, we utilized two publicly available datasets from Kaggle to build and evaluate our dental disease classification models. The first, the Oral Disease Dataset (link), contains labeled images of 6 dental conditions. The second, the Healthy Tooth Dataset (link), comprises images of normal, healthy teeth. There was a class imbalance, with conditions like Hypodontia and Calculus. Additionally, the datasets lacked angle diversity, limiting the model's ability to generalize across different views. Below find the breakdown of each class.

Condition	**Number of images**
Hypodontia	1251
Gingivitis	2349
Caries	2601
Mouth Ulcer	2806
Healthy	2680
Calculus	1296
Tooth Discoloration	2017

4. Methodology

The methodology was to build a deep learning-based classification pipeline capable of identifying 6 dental conditions and healthy teeth from real-world images. To accomplish this, a systematic approach was followed:

4.1 Data Preprocessing

To ensure consistency and compatibility across different models, the input images underwent several preprocessing steps:

- **Augmentation**: Since we got only 260 healthy images, we performed the following data augmentation, like random rotations, horizontal and vertical flips, random cropping and zooming, and then brightness, contrast, and saturation of the images were modified to get 2600 images, to balance the data.
- **Resizing**: All images were resized to a standard input size of 224×224 pixels.
- **Normalization**: Pixel values were first scaled to a range of [0, 1], and then normalized to a mean of 0 and standard deviation of 1 per channel (resulting in a range approximately [-1, 1]).
- **Tensor Conversion**: Images were transformed into 3D tensors of shape [3 x 224 x 224] suitable for input into PyTorch models.

4.2 Model Architectures

A total of five models were trained and compared

Vanilla CNN: The Vanilla Convolutional Neural Network (CNN) implemented in this project consists of a series of convolutional layers (Conv2D) followed by ReLU activation functions and MaxPooling layers, culminating in fully connected (dense) layers for classification. This simple yet effective structure enables the model to learn hierarchical features from dental images.

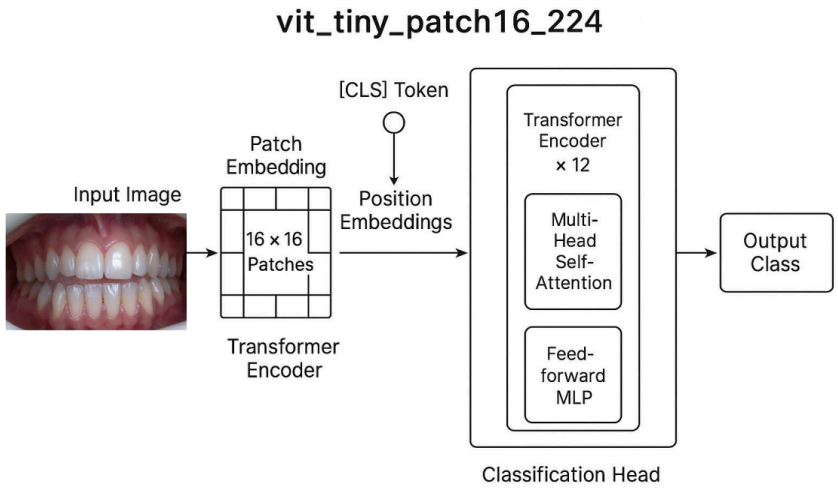
CNN with Attention: The CNN with Attention model enhances the convolutional architecture by integrating attention mechanisms after convolutional layers. These attention blocks help the network focus on the most relevant spatial regions within dental images, such as areas showing early-stage decay or subtle signs of hypodontia. By directing the model's capacity toward diagnostically significant features, this architecture improves classification accuracy and interpretability.

EfficientNet-B0: The EfficientNet-B0 model utilized in this project leverages a compound scaling method to efficiently balance network depth, width, and input resolution. Pretrained on the ImageNet dataset, it was fine-tuned on the dental image dataset to adapt to the specific features relevant to oral disease detection.

Custom Vision Transformer: The Custom Vision Transformer (ViT) model was designed and implemented from scratch to explore the performance of a transformer-based architecture without the aid of transfer learning. The architecture included a patch embedding layer to divide and project image patches, positional encodings to maintain spatial awareness, multiple transformer encoder blocks to model long-range dependencies, and a fully connected classification head to predict the 7 oral disease classes. Initially trained on a dataset of 15,000 dental images, the model achieved only 10% accuracy. Since Transformers are data hungry, we expanded our dataset to 150,000 images by augmenting the data. This led to a modest improvement, achieving 19% accuracy.

Pre-Trained Vision Transformer (ViT): The Pre-Trained Vision Transformer (ViT) model used in this project is the vit_tiny_patch16_224, a compact and efficient transformer architecture that replaces convolutional layers with

self-attention mechanisms. It processes input images by dividing them into 16×16 patches, embedding each patch, and applying positional encodings before passing them through multiple transformer encoder layers. Initially pre-trained on ImageNet, the model was fine-tuned on our dental dataset to adapt to domain-specific visual patterns. As part of the adaptation, the original classification head was replaced with a custom linear layer designed to predict the 7 oral disease classes present in our dataset. Despite its lightweight design, the fine-tuned ViT-tiny model demonstrated strong performance, effectively capturing subtle and spatially distributed features critical for accurate dental diagnosis.



4.3 Training Strategy

Model	Loss function	Learning Rate	Optimizer	Weight Decay	Epochs	Batch Size
Vanilla CNN	Cross-Entropy Loss	3e-4	Adam	1e-5	10	80
CNN with Attention	Cross-Entropy Loss	1e-4	Adam	1e-5	5	80
EfficientNet-B0	Cross-Entropy Loss	3e-4	Adam	NA	5	80
Pre-Trained ViT	Cross-Entropy Loss	3e-4	Adam	NA	3	80
Custom ViT	Cross-Entropy Loss	3e-4	Adam	NA	10	80

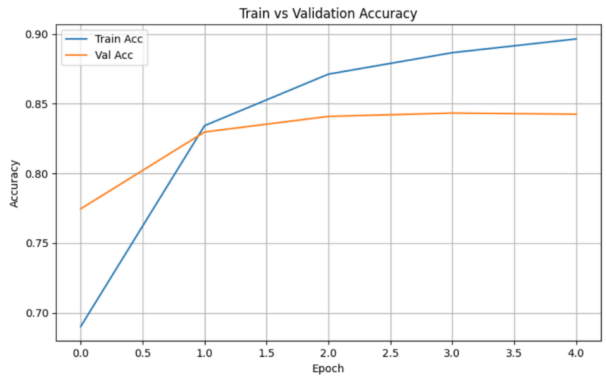
4.4 Evaluation Metrics

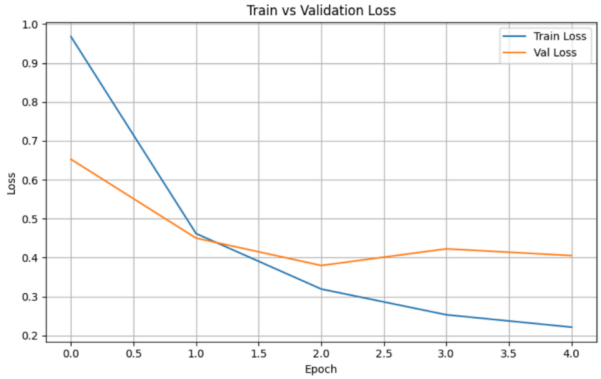
Each model was evaluated using the following metrics:

- **Accuracy**: Overall classification accuracy across all classes.
- **Precision, Recall, F1-Score**: Computed for each class to assess performance on imbalanced datasets.
- **Confusion Matrix**: Visualized misclassifications and helped identify class-specific weaknesses.

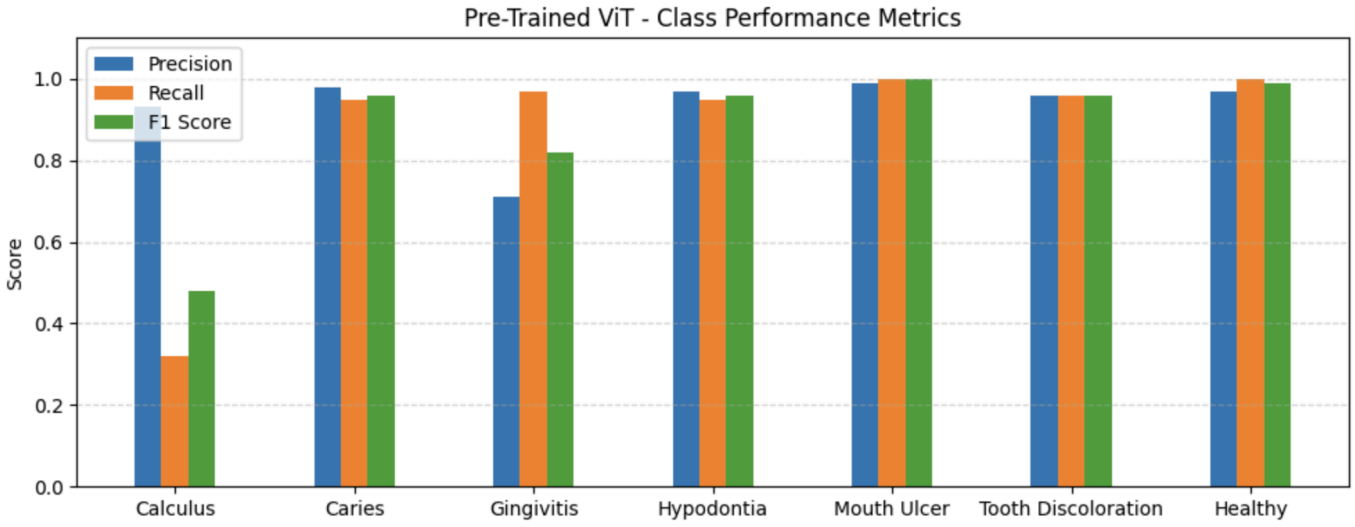
- **Loss and Accuracy Curves**: Training and validation metrics were plotted to analyze overfitting, underfitting, and learning dynamics.

Below find the accuracy and loss curves for the pre-trained ViT.





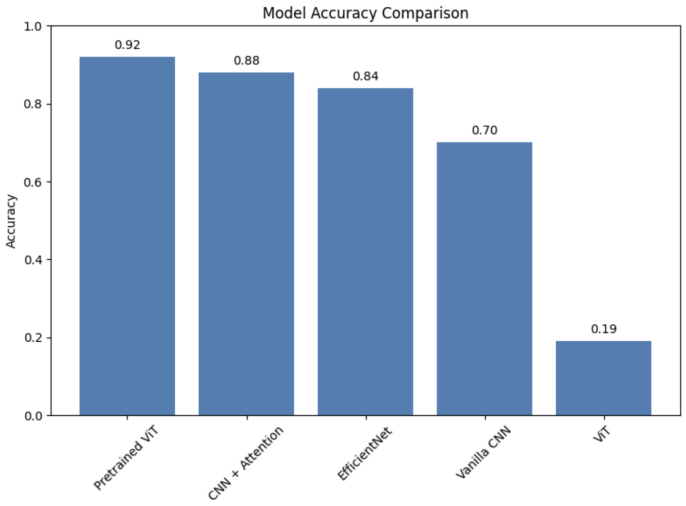
Below plot shows the class wise precision, recall and F1 score for the highest performing model, pre-trained ViT.



Results

Among all the models evaluated, the **Pre-Trained Vision Transformer (ViT)** achieved the highest accuracy and overall performance across all metrics. It consistently outperformed other models such as Vanilla CNN, CNN with Attention, EfficientNet-B0, and the Custom Vision Transformer, particularly in terms of precision, recall, and F1-score on the test dataset.

Given its superior accuracy and generalization capability, we selected the Pre-Trained Vision Transformer as the final model for further analysis and deployment. The remaining models, while helpful for comparative evaluation, showed lower performance and were not considered for final implementation.



Discussion

Despite the promising results from the *Dental Vision* project, several limitations were identified that restrict its performance and applicability in real-world settings. One of the major concerns was the model's performance on phone-captured or user-submitted images. These images often suffer from poor lighting, blur, and unconventional angles, which differ significantly from our training dataset. As a result, the model struggled to generalize when exposed to such inputs.

Another key limitation lies in the lack of diversity in the training data. Most images were captured from frontal or standard viewpoints and lacked sufficient representation of side angles, variations in oral structure. This limited variation negatively impacted the model's performance when processing real-world scenarios involving different oral anatomies and image perspectives.

Class imbalance was also a significant challenge. Conditions such as Hypodontia and Calculus were underrepresented compared other classes. This imbalance skewed the model's predictions, with the model favoring majority classes and performing poorly on minority ones, particularly in terms of recall and F1-score. Detecting these rare but clinically important conditions requires a more balanced and comprehensive dataset.

Additionally, the system assumes that every input is a valid dental image. It lacks a pre-filtering or rejection mechanism to flag irrelevant or non-dental images. Consequently, when faced with edge cases like images with lipstick, braces, or partial occlusion, the model will still classify the image, often producing misleading results.

Finally, the current models support only single-label classification and do not offer any localization capabilities. In real-world scenarios, multiple oral conditions can co-exist within a single image, and being able to detect and localize each issue is crucial for clinical use. The absence of multi-label and object localization functionality limits the model's practical utility in comprehensive dental assessments.

References

[1] Chen, K., Wang, J., Chen, L.C., Gao, H., Xu, W. and Nevatia, R., 2015. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv preprint* arXiv:1511.05960.

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

[3] Felsch, M., Meyer, O., Schlickenrieder, A., Engels, P., Schönewolf, J., Zöllner, F., Heinrich-Weltzien, R., Hesenius, M., Hickel, R., Gruhn, V. and Kühnisch, J., 2023. Detection and localization of caries and hypomineralization on dental photographs with a vision transformer model. *NPJ digital medicine*, 6(1), p.198.

[4] Gao, Y., Zhang, P., Xie, Y., Han, J., Zeng, L., Ning, N., Zheng, Q., Li, H., Chen, X. and Chen, Z., 2024. Application of transformers in stomatological imaging: A review. *Digital Medicine*, 10(3), pp.e24-00001.

[5] Ibrahimovic, E., 2023, May. Optimizing Vision Transformer Performance with Customizable Parameters. In *2023 46th MIPRO ICT and Electronics Convention (MIPRO)* (pp. 1721-1726). IEEE.

[6] Islam, N., Hasib, K.M., Joti, F.A., Karim, A. and Azam, S., 2024. Leveraging Knowledge Distillation for Lightweight Skin Cancer Classification: Balancing Accuracy and Computational Efficiency. *arXiv preprint* arXiv:2406.17051.

[7] Priya, J., Raja, S.K.S. and Kiruthika, S.U., 2024. State-of-art technologies, challenges, and emerging trends of computer vision in dental images. *Computers in Biology and Medicine*, 178, p.108800.

[8] Segu, M., Tonioni, A. and Tombari, F., 2023. Batch normalization embeddings for deep domain generalization. *Pattern Recognition*, 135, p.109115.

[9] Shorten, C. and Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1), pp.1-48.

[10] Ward, D.H., 2007. The vision of digital dental photography. *Dentistry Today*, 26(5), p.100.

Appendix 1: Contributions from each member

Madhavan Balaji contributed by implementing the EfficientNet B0 model and enhancing a CNN architecture with attention mechanisms for improved performance.
Raghuveer Venkatesh was responsible for collecting and preprocessing the dataset, building the Vanilla CNN model, and developing the website (link) to demonstrate the model's predictions.
Vigneshwar Ravi Rao used the pretrained Vision Transformer (ViT) model and fine-tuned it and built a custom ViT architecture for the dental image classification task.