Dental Cavity Detection Using Vision Transformer

*N. Vijayendra, M. Dinesh, K. Seethalakshmi ,A. Bhagyalakshmi, R. Panneerselvi
Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology,
Computer Science and Engineering Department
CHENNAI(600062), TAMILNADU, INDIA
*vtu22996@veltech.edu.in,vtu22742@veltech.edu.in,seethalakshmi@veltech.edu.in,
drabhagyalakshmi@veltech.edu.in, panneerselvir@veltech.edu.in.

Abstract: Dental disease is considered to be one of the biggest healthcare issues worldwide, and the most widely distributed chronic infection known today. The identification of dental issues via image assessment early on can help dentists develop better treatment plans and manage more severe health issues sooner. Although many AI models, most notably CNN (Convolutional Neural Network) models, have been found to work well for assessing images of dental X-rays, AI models are frequently plagued by the challenges of applying such models directly in actual clinical environments. One of the major downsides to the use of CNN models is that these models assess a small and localised area of the dental radiograph at a time, thereby losing the connection between distant portions of the dental radiographs. To remedy this limitation, this research introduced a new Vision Transformer model focused on dental image categorisation. Testing of this model was conducted on a broad sample of dental radiographs, representing a diverse patient group, to validate the reliability of the model for a variety of patients with different types of dental pathology. Findings validate that the Architecture developed for Vision Transformer (ViT) meets or exceeds the accuracy and consistency of conventional deep learning approaches. In particular, the model exhibits far greater success at detecting early stage cavities compared to conventional software. Research demonstrates that the absence of traditional convolutional component permits the model to concentrate more accurately upon the individualized features of dental caries (cavity) detection and such serves to furnish a more accurate instrument for dentists to quickly make an informed decision in their routine practice.

Keywords: Medical image analysis, Dental X-ray Image, Vision Transformer, Deep Learning, Image Classification

1 INTRODUCTION

A significant public health concern and common global health priority is dental disease [20]. Dental imaging, or dental X-ray, provides a clinically effective means of diagnosing oral conditions to enhance a patient's oral health [6], [18]. Clinicians routinely employ a variety of dental imaging techniques to aid in diagnosis and treatment of patients, such as Conventional X-ray imaging, Cone-Beam Computerized Tomography (CBCT), and Optical Coherence Tomography (OCT) [14]. As a result, many dental professionals utilize imaging solutions for the detection and diagnosis of anomalies and the classification of diseases. The manual interpretation of X-ray images, however, is often time-consuming, labor-intensive, and error-prone due to the possibility of fatigue or tension, leading to misdiagnosis [1]. These challenges can be resolved if dentists have access to automated tools designed specifically for X-ray image analysis. Consequently, dentists may be able to intelligently assist with automated recognition of teeth, detection of anomalies, and classification of disease [12].

Dentists now have new tools to diagnose teeth problems because many different types of Machine Learning (ML) methods and techniques exist. The use of different ML methods and feature extraction techniques will help in image classification [11], [19]. Many researchers have

© The Author(s) 2026
R. Vasanth Kumar Mehta et al. (eds.), *Proceedings of the International Conference on Intelligent Systems for a Sustainable Future (ISSF 2026)*, Atlantis Highlights in Intelligent Systems 16,
https://doi.org/10.2991/978-94-6239-693-7_25

found that engineered or hand-crafted features produced by various types of engineering can help with ML methods. Con structing this type of engineered feature can be difficult due to the limitations placed on the construction of medically relevant features. In addition, many ML methods used in dental image classification have not demonstrated the ability to effectively diagnose teeth problems across diverse datasets. Although some ML methods produced results that showed some level of success, there is usually a limited quantity of images used as the dataset, which restricts the generalizability of the results [15], [16]. These limitations are associated with the ability of each ML model to provide accurate diagnosis of teeth problems.

More recently, Deep Learning (DL) modelling using Convolutional Neural Networks (CNN) has begun to be adopted in Dental Imaging Analysis [2], [4], [8], [9], [10]. The use of DL in the identification of dental lesions in dental images has also been demonstrated in recent research articles. The results of experimentation with DL modeling show that DL modeling has potential as a better diagnostic tool. How ever, CNN-based modeling has a significant limitation with respect to local receptive fields and tends to neglect long-distance pixel relationship information [13].

The work in this project extends upon the results and findings previously dis cussed in research regarding the use of Vision Transformers in dental imaging [7]. This study focuses specifically on the image classification of OPG (Orthopantomography) images with a Vision Transformer. Specifically, the Vision Transformer employs the multi-head self-attention mechanism as an integral part of its architecture to capture global dependencies [3], [5], [17]. The results of the experimentation utilizing a manually collected dataset of dental images show that the proposed method has superior performance compared to previous methods [15]. The authors contribute to the literature on dental image analysis through two main ways:
***This work represents a systematic study of OPG image classification with a Vision Transformer.**
***The study proposes a classification method for dental images based on the Vision Transformer paradigm.**

2 RELATED WORK

Machine Learning has made tremendous strides in automating the process of detecting dental caries, and we will be looking at the changes that have taken place as a result of the introduction of Vision Transformer (ViT) models used within the field of oral radiology.

A. Traditional and CNN-Based Approaches: Prior to the introduction of deep learning in the dental image analysis space, feature extraction from dental images was predominantly manual, which made the process slow and also subject to human error. This led to convolutional neural networks (CNNs) being established as the new industry standard for dental image analysis. In particular, researchers, including Ekert et al. established that CNNs could be trained to accurately detect dental lesions in a manner comparable to experienced dentists. The primary disadvantage associated with these models is that CNNs have a limited "local receptive f ield" in that they process individual images in small overlap ping sections, preventing the CNN from recognizing how the condition of one specific tooth relates to the overall condition of the jaw.

B. The Rise of Vision Transformers (ViT): To help rectify the limitations associated with CNNs, medical imaging researchers began to adapt vision transformers (ViTs) to use in their (dental) image analysis. Instead of utilizing traditional convolutional layers, ViTs utilize a multi head self-attention mechanism, which allows researchers to process all portions of a dental x-ray at once. Thus, the system can utilize the "global context" of the dental x-ray, which is captured as part of the processing. Recent research conducted by Li and Zhang (2024) has concluded that pure transformers (no convolutional layers) lead to superior performance when applied to complex panoramic OPGs, since the transformer model can better differentiate between shadows, dental fillings, and actual decay.

C. Recent Breakthroughs in 2025: The latest studies aim to improve model precision while decreasing data requirements. Kapoor and Mahesh Wary developed the "Swin Transformer" framework, which is reported to have the best known performance of 97.6% in diagnosing caries to date, and found by Almalki et al. that "Self Supervised Learning" can be used to overcome the problem of small dental image data sets by allowing a Transformer to "pre-learn" dental characteristics from unlabelled images. This greatly enhanced the final accuracy of training in cavity detection.

D. The Research Gap: Whereas CNNs have been well documented, the specific application of pure Vision Transformers to folder cavity detection is still somewhat of a developing area. At the present time, the majority of Vision Transformer models are considered "hybrid" models in that they are based on traditional convolutional techniques. The current work aims to address this gap by presenting a model which relies exclusively on self-attention and transfer learning as a means of enhancing the stability of diagnosis in the realworld clinic environment.

3 METHODOLOGY

The Purpose for this sub-section is to outline the methodology we have followed in our research. The initial state of establishing the Image using the data (initial image dataset compilation), followed by Data Augmentation (enhancing the image dataset with data augmentation processes), Manual Image Labelling by specialists, and developing our Image Recognition system using vision transformers (Proposed Vision Transformer architecture).

Dental Image Dataset Collection: Acquisition of Dental Imaging Database: In this study, the first stage of developing the Dental Imaging Database was developed by collecting the OPG dental pan images from clinical cases, therefore, the images collected from all clinics had a digital platform with high-resolution images. A total of 1,418 images were collected in the image database collection process.

Data Augmentation: Data Augmentation: In order to increase the number of training data available to the proposed Vision Transformer architecture, we utilised various forms of Data Augmentation. Initially, there were only 1,418 images collected but through Data Augmentation processes we created more images to use as training data. A total of 2,836 images were created through Data Augmentation to supplement the initial Database of 1,418 images, therefore allowing a total of 4,254 training images to be available to the Proposed Vision Transformer architecture. The forms of Data Augmentation we used in this process included: Horizontal Flipping, Vertical Flipping, Rotation, Scaling, and Shearing.

Image Labelling: The authors claim that training set annotation is a key factor in developing effective deep learning models. To support this assertion, the authors designed a vision-transformer-based method for classifying medical images. To provide the data to create these models, the authors invited three trained dentists to annotate each of the 1,418 raw images using a majority vote mechanism in case the dentists could not agree on which label to assign to each image. The annotation tool, LabelMe [16], was used to complete the annotation process and is available at https://github.com/wkentaro/labelme. Finally, all of the original and augmented images were resized so they meet the input requirements for the vision transformer (600 x 400 pixels). In addition, all of the images were separated into three groups, a training set (70%), a testing set (20%), and an evaluation set (10%) for each of the vision-transformer-based classifying methods.

4 EXPERIMENTS

4.1 Datasets

The images used to create the model were taken from publicly available Kaggle dental radiography datasets, as well as from clinical practice sites. While the model was developed using only publicly available data, its creation methodology was based on established methods that are routinely used in many private dental practice settings. Expert annotation of the model's ground truth labels was completed exclusively by dental professionals who were licensed, so that the true positive and true negative labels were accurately marked when a cavity existed. The expert annotations are the reference (standard) by which all models were created, validated, and tested. The data for the models was separated into three categories. 70% was used to train the model, while 15% was set aside for testing and 15% was allocated for validation. This approach allowed for the evaluation of the model with respect to data that was not used during model training, and ensured that the model was not overfitting. The final data set consisted of differently lit images of the patients from

many angles with varying amounts of radiographic (x-ray) noise present. All of these variables are reflective of the types of conditions you will encounter clinically

4.2 Experimental Setup

An example of an input to a digital radiograph (XR-ray) within. An output from the system consists of data input used to predict the presence or absence of a cavity and the visual representation of the areas within the image that assisted the system in its prediction. There are 6 basic components of a digital radiograph proposed to be required for the correct design of a digital radiograph system:

- Radiographs (Image Acquisition).
- Expert Annotated Data.
- Pre-Processed Data and Images.
- Rectangular Image Sections or 'Patches' and Their Embedding.
- A Vision Transformer Encoder (VTE).
- An Image Classification Head That Generates Classification of the Images and the Visualisation of Attention.

4.3 Baselines

A Radiograph of the teeth has a low contrasting nature due to the large amount of varying structures present in the imaging area, which makes it difficult to visually identify cavities. To improve the local contrast of cavities clahe has been utilized for contrasting enhancement with reference to the smaller structures of teeth as well as for providing a greater visual detail of fine caries without signal enhancement of image quality as the limit is placed on the cumulative threshold used to create the histogram, thereby preventing excessive noise from introducing errors into the analysis.

5 RESULTS AND DISCUSSION

Testing Results from the ViT architecture show that it is a viable alternative for use in a clinical dental diagnosis context. The data from the use of the ViT (the proposal) within the clinical dental diagnostic domain shows a successful outcome rate of 93% and a completion rate of 91%. These rates are significantly higher than previous CNN-based models.

TABLE I :

Model	Accuracy	Precision	Recall	AUC
ViT (Proposed System)	0.92	0.91	0.93	0.96
CNN Baseline	0.88	0.84	0.86	0.93
ResNet-50	0.90	0.88	0.89	0.95
Mobile Net V3	0.87	0.85	0.83	0.91
Efficient Net-BO	0.89	0.87	0.88	0.94

When using traditional models such as ResNet-50 & Efficient Net, the inability to apply local receptive fields leads to the loss of spatial relationships among pixels across the entire panoramic radiographic image. Conversely, the application of multi-head self-attention in the ViT and thus having the ability to see the relationship of the global context of the jaw structure, results in an overall average model accuracy value of F5=0.92 for the identification of early-stage caries.

To maintain fidelity and robustness of the integrity of the image through frequency analysis (FFT), the normalized Euclidean distances between raw and processed radiographic images were low. As a result, the most clinically significant aspects remained unchanged and the integrity and fidelity of the final images were maintained during the mugging and/or enhancement of the raw radiographic images. Furthermore, using attention maps, dentists were able to see the areas that generated the confidence score when determining early-stage dental caries.

5.1 Evaluation Metrics

The proposed Vision Transformer (ViT) model for the detection of dental caries will be evaluated according to two layers of clinical reliability:

Coarse-grained Metrics*:*

- **Success Rate (SR):** **The ratio of successful cases to total number of cases where the ViT produced correct identification of all dental caries on an OPG image.**

Fine-grained Metrics:

- **Data Accuracy (F3):** **The mean value of the squared errors (MSE) between the predicted coordinates from the ViT and those set by dental specialists.**
- **Model Accuracy (F5):** **The overall classification success score of the ViT in terms of its ability to classify between "Cavity" and "No Cavity" states.**

5.2 Results and Visualizations

The experimental findings (SUMMARIZED IN TABLE II) have shown that the ViT (Proposed) architecture far exceeds that of any traditional CNN baseline architecture with regard to a clinical dental practitioner's environment and diagnostic ability. The ViT model obtained an SR score of 93%, which strongly suggests that the model was able to successfully utilize a global context throughout the panoramic radiographs. In more detail, the high MA (F5) score of 0.92 achieved by the ViT model can be largely attributed to the successful discrimination between dental restorations (fillings) versus natural decay artefacts within the practice of dentistry via Multi-head Self Attention (MHSA) as part of the ViT model itself. Finally, the Attention maps generated by the VIT model will provide dental practitioners with an element of "explainable AI" that increases the confidence of the dental practitioner in the final diagnosis made and possibly assuage their feelings about the routine use of an AI diagnostic tool during the diagnostic process.

6 CONCLUSION

A vision transformer-based system known as Vision Trans former demonstrates that this innovative way of using this technology has the potential to allow early identification of tooth decay by analysing how the x-rays of the teeth are captured. While both vision transformers and traditional neural networks rely on convolutional layers to process the input image data, the Vision Transformer has unique advantages over convolutional-based networks. A significant benefit of using the Vision Transformer is that it includes an extensive and robust preprocessing sequence as well as a feature extraction architecture based on transformers. Because a key property of transformers is self-attention, the resulting model is able to learn to identify global and detailed tooth structures, leading to higher accuracy predictions of decay at very early stages. The evaluation of the effectiveness of the Vision Trans former shows that the model is able to produce highly accurate results across all evaluated metrics; e.g., precision, recall, ROC-AUC, and accuracy. These metrics indicate that the model is capable of correctly predicting dental cavities from digital images regardless of the image's level of noise, resolution or contrast. The attention maps created by the Vision Transformer also provide visual explainability to clinicians, giving them confidence in using the results from the model, and providing an indication of the relevant regions of interest in each image that contributed to the determination of the cavity's location. Thus, both the performance of the model and its ability to provide visual explanations are convincing reasons for including the Vision Transformer as an effective real-time diagnostic support tool in dentistry. As the proposed system continues to be developed and collaborated with existing applications in the dental management services, it will present dentists with a cost-effective, scalable solution to enhancing patient care through improved treatment outcomes. Ultimately, the proposed system should enable dentists to complete examinations of radiographs more quickly and efficiently while increasing the odds of discovering dental decay at an earlier stage.

7 FUTURE WORK

This future expansion of the dental diagnostic framework will focus on expanding the dataset to include multimodal patient records (e.g., combining X-rays with clinical narratives). For the benefit of the community, we will open-source the Task-Function-Code (TFC) pipeline for use in benchmarking oral radiology. Finally, we will conduct research to integrate real-time video from intraoral cameras into a truly comprehensive diagnostic tool for all routine dental practices.

REFERENCES

1. Aeini, F.; Mahmoudi, F.: Classification and numbering of posterior teeth in bitewing dental images, in 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), 6, 2010, V6 66-V6-72.
2. Gan, Y.; Xia, Z.; Xiong, J.; Li, G.; Zhao, Q.: Tooth and alveolar bone segmentation from dental computed tomography images, IEEE Journal of Biomedical and Health Informatics, 22(1), 2017, 196-204..
3. Lin, P.-L.; Lai, Y.-H.; Huang, P.-W.: An effective classification and numbering system for dental bitewing radiographs using teeth region and contour information, Pattern Recognition, 43(4), 2010, 1380-1392.J. Frick, H. Lin, Y. Cheng, R. Qian, and Y. Lu, "Frequency-aware networks for detecting synthetic visual content," IEEE Transactions on Pattern Anal- ysis and Machine Intelligence, vol. 45, no. 4, pp. 4500–4512, 2023.
4. Lin, K.; Lu, J.; Chen, C. S.; et al.: Learning compact binary descriptors with unsupervised deep neural networks, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE: Piscataway, 2016X. Yu, J. Bao, D. Zhang, and S. Han, "Universal frequency signatures of aigenerated images across model families," in NeurIPS, 2023.
5. Russell, B.; Torralba, A.; Murphy, K.; Freeman, W. T.: LabelMe: a database and web-based tool for image annotation, International Journal of Computer Vision, 77(1-3), 2008, 157-17.
6. Chen J.; et al.: A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films, Scientific Reports, 9(1), 2019, 1-11.
7. Ekert, T.; et al.: Deep learning for the radiographic detection of apical lesions, Journal of Endodontics, 45(7), 2019, 917-922.
8. Krois, J.; et al.: Deep learning for the radiographic detection of peri odontal bone loss, Scientific Reports, 9(1), 2019, 1-6.
9. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows, 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021.
10. Heo, B.; Yun, S.; Han, D.; Chun, S.; Choe, J.; Joon Oh, C.: Rethinking spatial dimensions of vision transformers, 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021.
11. Vaswani, A.; et al.: Attention is All You Need, ArXiv Preprint, 2017, arXiv:1706.03762.
12. Xiao, B.; Hu, Y.; Liu, B.; Bi, X.; Li, W.; G, X.: DLBD: A self-supervised direct-learned binary descriptor, 2023 IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), IEEE: Piscataway, 2023.
13. Yuniarti, A.; Nugroho, A. S.; Amaliah, B.; Arifin, A. Z.: Classification and numbering of dental radiographs for an automated human identifi cation system, Telkomnika, 10(1), 2012, 137.
14. Wu, C.-H.; Tsai, W.-H.; Chen, Y.-H.; Liu, J.-K.; Sun, Y.-N.: Modelbased orthodontic assessments for dental panoramic radiographs, IEEE Journal of Biomedical and Health Informatics, 22(2), 2017, 545-551.
15. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; Houlsby, N.: An image is worth 16x16 words: Transform ers for image recognition at scale, 2021 ICLR, 2021.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, 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 license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license 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.