Kufa Journal of Engineering
Vol. 17, No. 2, April 2026, P.P. 375-386
Article history: Received 11 June 2025, last revised 14 July 2025, accepted 26 July 2025

A NOVEL APPROACH TO DENTAL X-RAY ANALYSIS: USING VISION TRANSFORMERS FOR DETECTING CARIES

Wasan Mueti Hadi[1], Zahraa K. Al-Sendi[2], Manar Hamza Bashaa[3], Ghosoon k. munahy[4], and Noor Abbas Khudhair[5]

[1] Department of Computer Science, College of Computer Science and Information Technology, University of Kerbala, Iraq, Email:wasan.m@uokerbala.edu.iq.

[2] Department of Computer Science, College of Computer Science and Information Technology, University of Kerbala, Iraq, Email:zahraa.k@uokerbala.edu.iq

[3] Department of Information Technology, College of Computer Science and Information Technology, University of Kerbala, Kerbala, Iraq, Email:manar.h@uokerbala.edu.iq.

[4] Department of Information Technology, College of Computer Science and Information Technology, University of Kerbala, Kerbala, Iraq, Email:ghosoon.k@uokerbala.edu.iq.

[5] University of Al_Ameed College of Dentistry, Kerbala, Iraq, Email:noorkhudhair419@gmail.com.

https://doi.org/10.30572/2018/KJE/170222

ABSTRACT

All age groups are affected by oral diseases that are common worldwide. The dentist relies on Dental radiographs were used to explore the characteristics of oral diseases. Dental X-ray image segmentation and analysis pose significant challenges compared with other medical images. This secondary challenge makes dental radiography challenging. Because dental images are captured at a lower resolution, the segmentation of a tooth and its related complications can be unreliable because they are not resolvable. Dental X-ray Image Segmentation (DXIS) is one of the most fundamental and important steps in obtaining relevant information concerning oral diseases. In dentistry, DXIS is an important step in obtaining many different pathologies of gingival tissues. The next proposed methodology uses Vision Transformers (ViTs) to identify dental caries from dental X-ray images. In contrast to traditional CNN-based approaches, this approach uses attention mechanisms to dissect each patch of the image in more detail and yields more accurate results with earlier detection of caries. The results showed that ViTs are better

than CNN, the proposed performance accuracy reached 95% compared to the accuracy of the convolutional neural network, which reached 86%.

KEYWORDS

Vision Transformers, Medical Image Analysis, Caries Detection, Deep Learning in Dentistry, Self-Supervised Learning.

1. INTRODUCTION

Worldwide, individuals experience cavities, commonly referred to as tooth decay, which is a chronic and frequent infectious oral disease. It is predicted that dental diseases afflict about 50% of the population, with 2.3 billion individuals experiencing permanent cavities. Tooth decay, albeit often labeled as cavities, is an oral disease in which oral bacteria form lactic acid, causing the enamel surface of the tooth to deteriorate. This may result in the formation of small crevice spaces between teeth, the neglect of which may lead to pain, infection, and tooth loss. Dental information technology is the latest topic in the field that can automate and simplify analysis procedures in dental clinics, save patients' time, and reduce their daily stress (Hasnain et al., 2024). Traditional clinical detection for dental caries is primarily based on the naked-eye observation by dentists, and it is experience-based, leading to false conclusions. Additionally, early caries cannot be detected visually, particularly on the interproximal tooth surface. Naked-eye inspection-based detection based on X-rays is the gold standard for detecting invisible carious teeth and can be used to assess the extent of caries. Detection performance depends on the experience of the stomatologist in the interpretation of X-ray images, and inexperienced young dentists fail to detect minor carious lesions. Compared to multiple images, this comparison increases the workload for physicians. It is crucial to create a superb repeatability and precise automatic detection scheme to offer objective caries diagnostic aid to clinical stomatologists. Researchers have put emphasis on computer-aided detection of caries via X-ray images. Tracy et al. examined and showed the usefulness in applying density analysis assistance in detecting and classifying caries (Ying et al., 2022).

Color-based intraoral and radiation-based images can be classified into two categories. Intraoral color-based images have applications in the dental-care sector, whereas color-based images have applications in the same sector in the form of X-rays. Although there are threats involved in the X-ray imaging process in the form of harm, images formed through X-rays can be divided into three categories:

1. Bitewing
2. Periapical radiograph
3. Panoramic radiograph

Bitewing X-rays were used to identify the specific portions of the upper and lower dental arches. Periodontal disease and interproximal caries can be diagnosed by radiographic imaging. Periapical X-ray images display the entire depiction of the teeth from the surface enamel through the gingival zone, and Panoramic X-ray scans display the entire oral region, gums, and dental region. These photographs are frequently used.

This article discusses the periapical and panoramic radiographs. Fig. 1 shows a clear depiction of the three varieties on the dental radiographs (Liu et al., 2017).

Fig. 1. Different types of dental x-ray images.

Computer-aided diagnostic systems offer an improved solution for these issues. A computer-based mathematical model for disease diagnosis was developed based on the analysis and computing power of a computer. Furthermore, the identification, forecasting, and localization of lesions for such diseases can significantly reduce clinical practitioners' workload and ease difficulties. In recent years, artificial intelligence's revolutionary development means it's been increasingly applied for medical imaging, where deep learning is a subfield of machine learning that allows computers to analyze, learn from, and comprehend data via a hierarchical structure (Mohammed et al., 2022).

Deep learning is currently the most ubiquitous approach. This approach accepts a large quantity of medical images and uses convolutional neural networks (CNN) to learn and extract the image features (Zhu et al., 2023). Apart from the paper presented here, we propose in the paper at hand a vision transformer (ViT) that solves computer vision tasks applying the transformer architecture. ViT is better equipped for handling long-range dependencies and complex patterns in an image than conventional CNNs, because it divides the image into patches of the same size and treats them sequentially.

ViT is very useful in image classification because it captures the global context, particularly in areas where spatial relationships and minute details matter the most (Zhang et al., 2025).

2. RELATED WORK

The detection and analysis of dental caries from X-ray images have also witnessed tremendous progress with the application of deep learning-based approaches, specifically Vision Transformers (ViTs). Vision Transformers have demonstrated outstanding performance in medical image analysis based on their success in natural language processing, including the ability to model long-range dependencies and spatial relationships (Azad et al., 2024). On the other hand, CNN has been applied to predict the existence of Keratoconus. Inference in the CNN model trained with the GoogleNet parameters attained an accuracy of 96.6% on the

Keratoconus dataset (Abbosh et al., 2025). In this section, we recapitulate related work exploiting Vision Transformers for dental X-rays analysis and caries detection, and their contributions, methodology, and performance metrics work in (Zhou et al., 2023) presented an innovative Swin Transformer aided by tooth type information for the detection of caries in children on panoramic radiographs. In the model, differences among canine, molar, and incisor teeth were emphasized to improve diagnostic accuracy. The tooth-type enhanced swine transformer performed better than traditional CNN methods, with an accuracy value of 0.8557, precision value of 0.8832, and F1 value of 0.8567. This indicates the benefit of incorporating domain knowledge into Vision Transformers to achieve a better performance in dental tasks. In (Felsch et al., 2023), the authors presented the application of the Vision Transformer model (SegFormer-B5) for the detection and localization of MIH and caries in dental images. The model exhibited high-performance metrics, with IoU = 0.959, AP = 0.977, and accuracy = 0.978. This study demonstrated the feasibility of pixel-wise detection and localization of dental abnormalities using the Vision Transformer, with good performance for non-cavitation caries (IoU = 0.630) and dentin cavities (IoU = 0.692). The work in (Hossain et al., 2023) presented CaViT, an early caries detection system, from images obtained using a smartphone based on the Vision Transformer model. The model detected no caries, early caries, and advanced caries with 95%, 91%, and 100% sensitivity, respectively. Incorporating a U-Net for segmentation enhanced the system with the potential for low-cost and quick caries detection, without the need for face-to-face consultation. Paper (Gao et al., 2023) presented BTCN, an SSL model with a CNN and Transformer branch, to encourage regularization during the process of fine-tuning features. The parallel architecture in the BTCN learned global representations; hence, it could adapt with greater flexibility in the scenario for downstream detection models. The Multilayer Supervision Strategy (MSS) enhanced feature fusion with state-of-the-art performance on 1039 endoscopic caries image datasets. Jiang et al. (Jiang et al., 2021) introduced RDFNet, the FReLU activation function-based vision transformer approach, for the high-speed detection of caries. RDFNet was balanced in terms of both speed and accuracy and was hence portable-device-optimized. The performance of the model indicated the achievement of the application of the vision transformer, where detection was sped up immensely with no loss of accuracy. In (Sun and Chen, 2022), an attention-based transformer model for caries detection was created, and an innovative attention mechanism was presented for rich representation in varying channels. The model recorded an AP50 value of 63.81% and demonstrated the feasibility of detecting unlabeled caries in X-ray images, thus achieving the application of attention mechanisms in vision transformers for the representation of features. The work in (Li et al.,

2024) introduced SPGTNet, a spatial prior-guided transformer method for tooth instance segmentation in panoramic X-ray images. Using tooth positional characteristics and distant contextual information, SPGTNet has achieved state-of-the-art performance on public data. Proper detection and analysis of tooth structure using this model provides valuable information for treatment planning and dental diagnosis. Ying et al. (Ying et al., 2024), the detection performance of recent deep networks, including YOLOv5, DETR, UNet, and Trans-UNet, for caries detection. The highest F1-score (0.87) was attained via YOLOv5, followed by 0.86 using Trans-UNet, and 0.82 through DETR. AbstractOur results show that Vision Transformers can be used to obtain competitive performance compared to the state-of-the-art object IoU detection networks. Tooth numbers and caries were also detected using a cascade R-CNN architecture in (Yoon et al., 2024). The model achieved a value of 0.880 for detection of the number of teeth and 0.769 for caries detection. The application of Vision Transformers to large-scale intraoral data has demonstrated their viability in direct clinical applications. In (Hashim&Mazinani et al., 2025) A novel CNN model has been developed that, in contrast to conventional CNNs, can effectively process limited quantities of real data previously subjected to PCA, yielding high accuracy outcomes. The suggested system operates in three stages: pre-processing, which encompasses cropping, converting colored photos to greyscale, histogram equalization, and resizing; feature extraction; and classification utilizing machine learning methods, including Naive Bayes (NB), K-Nearest Neighbors (KNN), AdaBoost (ADA), Decision Trees (DT), and Convolutional Neural Networks (CNN).

3. METHODOLOGY & MATERIALS

3.1. Dataset Description

The main aim of this study aims to Detect Caries where the Dental Panoramic X-rays were collected from the Kerbala Institute of Dentistry The collection from our private repository has 374 images of patients with dental disorders (tooth decay, fillings, and implants).

3.2. Data preprocessing

The dataset consists of images of [different classes it represents]. The data used for classification was preprocessed using a multistage preprocessing pipeline. Initially, ResNet50, a deep convolutional neural network pre-trained on ImageNet, was used to extract high-level feature vectors from each image. One of the final layers of ResNet50 that holds the most significant visual features of an input image was used to transform each of them into a 768-dimensional vector.

In order to reduce the computational complexity and overfitting, we applied Principal

Component Analysis (PCA) to the feature vectors, reducing their dimensions from 768 to 512 while maintaining the most significant properties of the data. Since image labels were not available, the K-means clustering algorithm was then used to divide the feature vectors into two clusters (k=2). Because those clusters were then used as pseudo-labels during the training phase, they effectively supported a semi-supervised methodology as well (ie, explicit labels were not provided — only the structure was used to infer the labels).

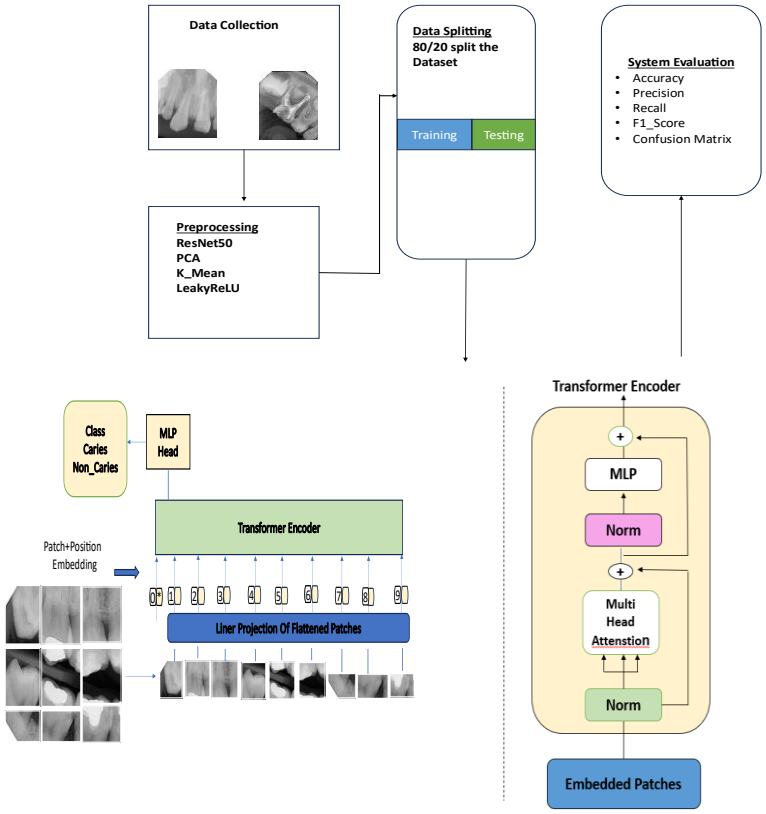

Fig. 2. The proposed model

3.3. Vision Transformer (ViT)

Vision Transformers, presented in 2020 (11) by Google Research, have been extremely successful in representing global dependencies and relations in images, and their success has led to enhanced performance in a broad set of computer vision tasks. In ViT, the self-attention mechanism supplants convolutions, allowing the model to capture long-range dependencies and contextual information. This is especially beneficial in domains such as medical imaging, where nuanced connections between remote picture regions are frequently essential for precise diagnosis.

This study conducts a comprehensive comparison of CNNs and ViTs across multiple dimensions: architectural differences, training efficiency, performance on benchmark datasets, and adaptability to various image processing tasks. We examine the training and validation loss

patterns and assess both models utilizing the Brain MRI dataset based on critical metrics such as accuracy, precision, recall, and ROC-AUC. The ViT model achieved an accuracy of 88.5%, whilst the CNN model demonstrated an accuracy of 85.5%. Moreover, we see that Vision Transformers (ViTs) exhibit superior efficacy in capturing long-range correlations within images compared to Convolutional Neural Networks (CNNs) in extracting localized characteristics. Our findings elucidate the merits and demerits of each strategy, hence highlighting their significance in various computer vision contexts. This work aims to assist academics and practitioners in identifying the most suitable model architecture for certain applications, particularly in medical imaging and complex vision tasks that necessitate precise feature recognition (Mittal et al., 2024).

Building to these findings, ViTs process images in patches sequentially and model the interactions between patches via a self-attention mechanism, whereas CNNs process images in the form of local convolution. Consequently, as a result of architectural dissimilarity, ViTs better capture local and global features and, in the process, find applications in medical image analysis where contextual information and minor patterns play a significant role in making accurate diagnoses.

The base model employed the `google/vit-base-patch16-224-in21k` model, the Vision Transformer model that was pre-trained on the ImageNet21k dataset. The image was split into 16 * 16pixel patches and the hidden dimension was 768 pixels. The model consists of a 3072-layer multilayer perceptron, 12 transformer layers, and 12 multihead self-attention heads. A dropout rate of 0.1, was used during training to avoid overfitting; the attention operations were not dropped. To preserve spatial information, the model trained 1D positional embeddings and was configured for binary classification into two classes (Mail-Sharifa, 2025).

4. RESULT

With the latest validation loss of 0.0999, the training and validation performance reached 95% and 95%, respectively. Class 0 achieved a precision rate of 0.96, recall of 0.95, F1-Score of 0.96, and support of 1228 samples, whereas Class 1 achieved a precision of 0.95, recall of 0.96, F1-Score of 0.95, support of 1124 samples, and the accuracy of the model was 0.95%. From the training graphs, we can see a gradual increase in accuracy, along with a consistent loss curve.

Smooth curves, consistent learning, and low overfitting indicate the verification of metric training data.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (2)$$

$$\text{F1} = 2 \times \frac{(\text{Precision}) \times (\text{Recall})}{(\text{Precision}) \times (\text{Recall})} \quad (3)$$

$$\text{Accuracy} = \frac{\text{TP+FP}}{\text{TP+TN+FP+FN}} \quad (4)$$

TP = true positive; this reflects the fact that the model predicted a positive outcome and the actual outcome was positive.

Here, FP denotes a false positive, in which the value is actually false and the model predicts the value as positive.

TN stands for true negative; the model predicts the value as negative, and the actual value is negative.

FN is a false negative; the result is negative, and the model prediction is negative.

Precision: classifier's ability to detect useful data points Eq.1.

Recall: This is the measurement criterion for determining how well the model identifies positive classes in the dataset. Eq.2:

F1: Accuracy value of the measured model compared with the dataset. Eq.3:

Accuracy: This metric was used to assess the efficiency of the model in identifying patterns and correlations among features in the dataset Eq.4.

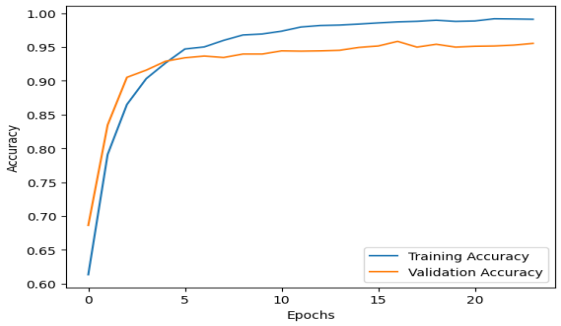

Fig. 3. Recognition accuracy

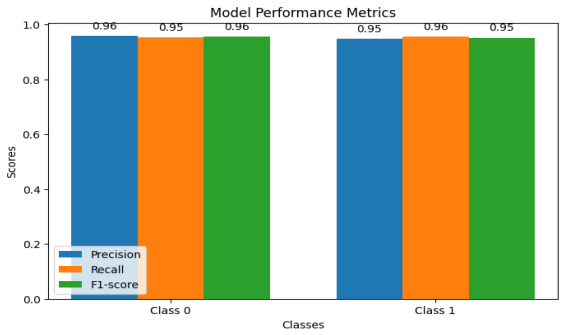

Fig. 4. Model Performance Metrics

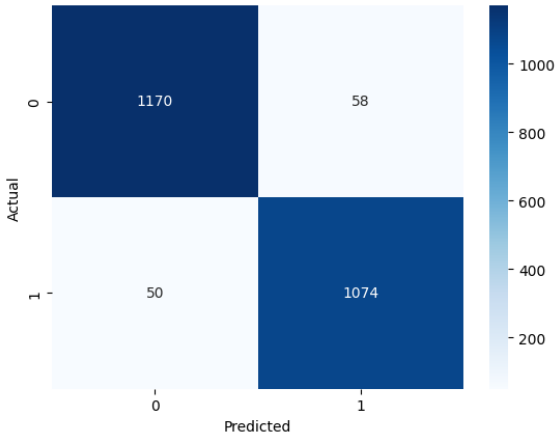

Fig. 5. Confusion Matrix of the ViT Classification

Table1. Classification Report for Vision Transformer

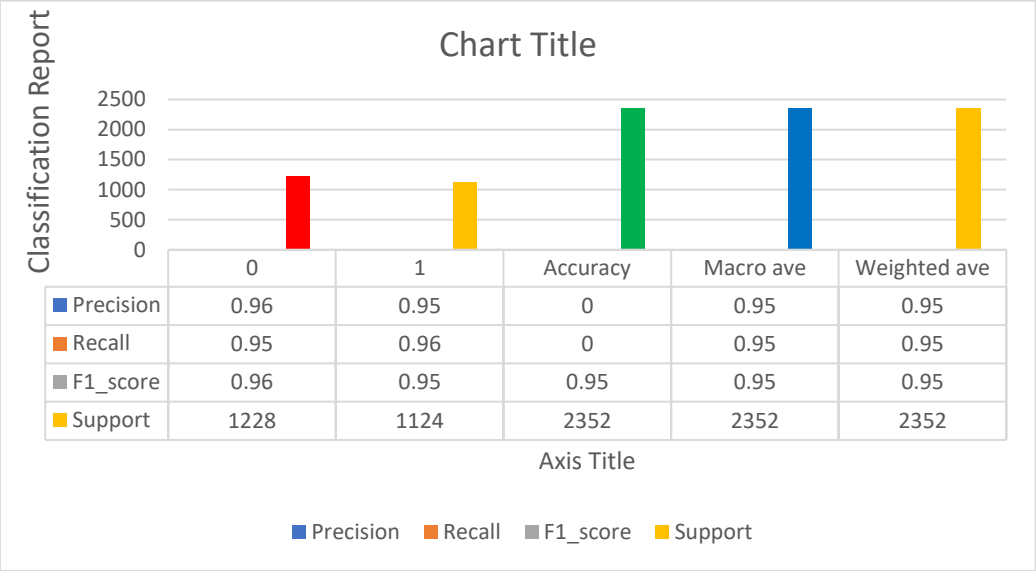

Table 2. Comparison between Other Methods and proposed model using Vision Transformer

	Model	Task	Accuracy/F1/Other
Our Method (ViT)	ViT	Caries Classification	Accuracy=0.953, F1=0.96
Zhou et al. (2023)	Swin Transformer	Caries Detection (Children)	Accuracy=0.8557, F1=0.8567
Felsch et al. (2023)	SegFormer-B5	MIH/Caries Localization	Accuracy=0.978, IoU=0.959
Hossain et al. (2023)	CaViT (ViT + U-Net)	Early Caries Detection	Sensitivity:91–100%
Sun and Chen (2022)	Attention-based Transformer	Unlabeled Caries Detection	AP50=63.81%
Ying et al. (2024)	Trans-UNet / YOLOv5	Object Detection for Caries	F1: YOLOv5=0.87, Trans-UNet=0.86

5. CONCLUSION

The body of work surrounding vision transformers in the analysis of dental X-rays, including caries detection, builds a case in which these models improve the accuracy and efficiency relative to existing options. These studies suggest that vision transformers are viable alternatives to traditional methods, with good performance as reliable automated methods for detecting early caries. The results of the vision transformers showed a high sensitivity for early caries detection, achieving high rates in terms of sensitivity (up to 95%).

6. REFERENCES

Abbosh, Y.M., Ali, S.M., Ali, D.M. and Alhummada, I.A.(2025). “Keratoconus detection using deep learning”. Kufa Journal of Engineering, 16(2).

Azad, R., Kazerouni, A., Heidari, M., Aghdam, E.K., Molaei, A., Jia, Y., Jose, A., et al. (2024), "Advances in medical image analysis with vision transformers: a comprehensive review", Medical Image Analysis, Elsevier, Vol. 91, p. 103000, doi: 10.1016/j.media.2023.103000.

Felsch, M., Meyer, O., Schlickenrieder, A., Engels, P., Schönewolf, J., Zöllner, F., Heinrich-Weltzien, R., et al. (2023), "Detection and localization of caries and hypomineralization on dental photographs with a vision transformer model", NPJ Digital Medicine, Nature Publishing Group UK London, Vol. 6 No. 1, p. 198.

Gao, N., Li, Y., Liang, R., Chen, P., Tang, J. and Liu, T. (2023), "Btcn: Bridging the gap between pre-trained and downstream models for endoscopic caries detection", 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, pp. 857–862, doi: 10.1109/BIBM58861.2023.10385358.

Hashim, A.A. and Mazinani, M.(2025). "Detection of keratoconus disease depending on corneal topography using deep learning". Kufa Journal of Engineering, 16(1).

Hasnain, M.A., Ali, Z., Maqbool, M.S. and Aziz, M. (2024), "X-ray Image Analysis for Dental Disease: A Deep Learning Approach Using EfficientNets", VFAST Transactions on Software Engineering, Vol. 12 No. 3 SE-Articles, pp. 147–165, doi: 10.21015/vtse.v12i3.1912.

Hossain, M.S., Rahman, M.M., Syeed, M.M.M., Hannan, U.H., Uddin, M.F. and Mumu, S.B. (2023), "Cavit: Early stage dental caries detection from smartphone-image using vision transformer", 2023 4th International Conference on Artificial Intelligence, Robotics and Control (AIRC), IEEE, pp. 9–14, doi: 10.1109/AIRC57904.2023.10303012.

https://github.com/wasanhadi/My-dataset/tree/main

Jiang, H., Zhang, P., Che, C. and Jin, B. (2021), "Rdfnet: A fast caries detection method incorporating transformer mechanism", Computational and Mathematical Methods in Medicine, Wiley Online Library, Vol. 2021 No. 1, p. 9773917.

Li, P., Gao, C., Lian, C. and Meng, D. (2024), "Spatial Prior-Guided Bi-Directional Cross-Attention Transformers for Tooth Instance Segmentation", IEEE Transactions on Medical Imaging, IEEE, doi: 10.1109/TMI.2024.3406015.

Liu, W., Zhou, X., Durrani, S. and Popovski, P. (2017), "A novel receiver design with joint coherent and non-coherent processing", IEEE Transactions on Communications, IEEE, Vol. 65 No. 8, pp. 3479–3493.

Mail-Sharifa, E. (2025), “From Image to Insight: Using Vision Transformers to Revolutionize Dental Caries Assessment in Radiographic Imaging”, SEEJPH, Vol. XXVI No. S2, pp. 494–501.

Mittal, P., Sharma, B. and Yadav, D.P.(2024). “Comparative analysis between CNN and ViT using brain MRI dataset”. In: 2024 Eighth International Conference on Parallel, Distributed and Grid Computing (PDGC), December 2024, pp.290–295. IEEE.

Mohammed, H.A., Kareem, S.W. and Mohammed, A.S. (2022). “A comparative evaluation of deep learning methods in digital image classification”. Kufa Journal of Engineering, 13(4).

Sun, C. and Chen, H. (2022), “An attention-based transformer model for dental caries detection”, International Conference on Electronic Information Engineering, Big Data, and Computer Technology (EIBDCT 2022), Vol. 12256, SPIE, pp. 673–679, doi: 10.1117/12.2635362.

Ying, S., Huang, F., Shen, X., Liu, W. and He, F. (2024), “Performance comparison of multifarious deep networks on caries detection with tooth X-ray images”, Journal of Dentistry, Elsevier, Vol. 144, p. 104970, doi: 10.1016/j.jdent.2024.104970.

Ying, S., Wang, B., Zhu, H., Liu, W. and Huang, F. (2022), “Caries segmentation on tooth X-ray images with a deep network”, Journal of Dentistry, Elsevier, Vol. 119, p. 104076, doi: 10.1016/j.jdent.2022.104076.

Yoon, K., Jeong, H.-M., Kim, J.-W., Park, J.-H. and Choi, J. (2024), “AI-based dental caries and tooth number detection in intraoral photos: Model development and performance evaluation”, Journal of Dentistry, Elsevier, Vol. 141, p. 104821, doi: 10.1016/j.jdent.2023.104821.

Zhang, X., Guo, E., Liu, X., Zhao, H., Yang, J., Li, W., Wu, W., et al. (2025), “Enhancing furcation involvement classification on panoramic radiographs with vision transformers”, BMC Oral Health, Springer, Vol. 25 No. 1, p. 153, doi: 10.1186/s12903-025-05431-6.

Zhou, X., Yu, G., Yin, Q., Yang, J., Sun, J., Lv, S. and Shi, Q. (2023), “Tooth type enhanced transformer for children caries diagnosis on dental panoramic radiographs”, Diagnostics, MDPI, Vol. 13 No. 4, p. 689.

Zhu, H., Cao, Z., Lian, L., Ye, G., Gao, H. and Wu, J. (2023), “CariesNet: a deep learning approach for segmentation of multi-stage caries lesion from oral panoramic X-ray image”, Neural Computing and Applications, Springer, pp. 1–9.