A Deep Learning – Enabled Enhanced Swin Transformer Employing Window-Based and Shifted Window Multi-Head Self-Attention with Residual MLPs for Accurate Dental Caries Detection

Saurabh Kapoor [1], Dr. Priti Maheshwary [2]

[1] Affiliation (Department of Computer Science and Engineering Rabindranath Tagore University, Bhopal, 464993, India)
[2] Affiliation (Department of Computer Science and Engineering Rabindranath Tagore University, Bhopal, 464993, India)
E-mail: saurabh.kapoor2@gmail.com, pritimaheshwary@gmail.com

Abstract

Dental caries (DC) is widely recognized as a major oral health concern worldwide for affecting individuals of all age groups. Dental decay problem continues rapidly and emphasizing early diagnosis system and importance of making reliable clinical decisions. Early detection of dental caries enables timely treatment and prevent further deterioration of tooth structure to improves patient outcomes. In this study, an approach of deep learning framework which is based on Enhanced Swin transformer with Hybrid Shifted Window Attention and a Residual Multi-Layer Perceptron (MLP) is proposed for dental caries detection. This research model effectively captures the contextual features of dental images which is both local and global images by integrating hybrid attention mechanism with hierarchical features extraction. Furthermore, the enhanced features representation was introduced by using residual MLP module which is having better stability.

The proposed system was evaluated better effectiveness assessed using common metrics such as accuracy, precision, recall and F1 score by implementing ten-fold cross validation technique. The model results were demonstrated effectively and the outcomes were evaluated against conventional deep learning architecture. And this study describes the state-of-the-art methods which is reported in the literature that indicates that the proposed approach achieves the better diagnostic performance, attaining a maximum accuracy of 93.4%, automated dental caries detection in clinical environment and also highlighting its potential for reliability.

Keywords - *dental caries analysis, dental caries classification, Vision transformer, Swin transformer*

1 Introduction

Dental caries has remained a significant health issue in the world, and thus rapid and precise testing equipment is needed. Although it is an effective method, normal radiography is often susceptible to human error due to picture noise, overlap of features (roots, airways) and exhaustion. Vision Transformers (ViT) have revolutionized medical imaging, but traditional models often cannot sustain the global attention memory demands. To maximize the efficiency and diagnosis accuracy of dental caries in to enhance processing efficiency, this research presents an adapted structure of Swin Transformer to panoramic radiograph images by applying a Crossover Shifted Windows (HSW-MSA) approach and a Residual-based MLP (ResMLP).

Poor oral health may create significant limitations in the domain of dentistry such as cavities, discomfort, root canal treatment, etc. you may also notice diseases such as dental caries and periodontal disease as they create a negative impact on your general health besides damaging your teeth. Research has shown a clear association between poor oral health and major systemic diseases [1,2].

Dental caries has increasing very rapidly in the age between 18-50 years. The major problem caries in mouth infection, caries problem, tooth decade, root canal problem, it's very challenging for orthodontics doctor now a days. So, in this research focus on the dental image to diagnoses the early detection using radiographic images. So that early detection of caries helps the patient and doctor for treatment. Using the proper image, deep learning algorithm may help for better outcome in prevention of caries and also diagnosis properly to follow the better treatment procedure. Orthodontics doctor uses diagnosis tools rapidly and for better outcomes. To detect the caries inside mouth it is tedious task to use tools, so that the CNN Deep learning transformer comes in picture to perform and detect the caries accurately with some parameters like tooth size, depth of tooth etc. Using this approach, it will examine that the successful procedure will achieve to detect the caries which helps dentist to diagnosis the better result.[1]

In Dental radiography, using Swin transformer will overcome the traditional approach in the domain of dentistry. And the research confirms that the dental medical image diagnosis is far better than the traditional approach.

Moreover, the physicians benefit a lot with the algorithm based on deep learning and computer-aided diagnosis systems in the dentistry sector have been effective in detecting and classifying dental caries via dental radiographs [4]. It may help dentists to make a rapid and precise diagnosis of their patients and improve oral healthcare services overall.

CNN or rather Deep learning methods enable automatic detection of meaningful patterns in images. This approach is more efficient and effective approach as compared to the traditional method that relies on humans to extract significant features. Although these innovations have contributed to more useful diagnostic tools, such issues as the insufficient quantity of samples to diagnose and model overfitting are still present. Such issues within the field of dentistry can be mitigated with the help of data augmentation, regularization, and cross-validation, and it is the interpretable models that can increase the confidence of dental professionals and other users of such technologies [3].

This research offers a new approach since there is a challenge in diagnosing the existence of dental caries. This study is essential to effective treatment and prevention, early detection of the disease. This paper represents the modification of an effective computer vision model with the Swin Transformer architecture to enhance the treatment and diagnosis of dental caries with CAD systems based on the deep learning technique [5]. Swin-Base is a scaled Swin Transformer model which analyzed the multi-class tooth radiography pictures. Scaled model is a simpler and more efficient architecture, which relies on fewer parameters to increase the detection accuracy [6].

The Hybrid Shifted Window Multi-Head Self-Attention module is an add on and also an upgradation to the Swin Transformer improving its performance. In cases of overlap of regions in dental radiographs, this module works well. Such studies will be carried out and tracked to help improve the management of long-range dependencies. As a result, it enhances precision and minimizes cases of false negative results. By diverse experiments, the Swin Transformer get the must better results than the existing state-of-the-art deep learning models [7 - 9]. This proves how effective it is and how it may be used in the diagnosis of the dentist.

Considering all the factors, it will discover that the algorithm deep learning has great potential in enhancing the process of dental caries diagnosis and this is enhanced by the creation of dental imaging. The AI-based techniques developed and employed in the assessment and classification of radiographic images have been highly accurate when it comes to the dental caries diagnosis. The method having high efficiency made in the current research to implement an algorithm as effective in the diagnosis of dental healthcare.

2 Related Work

In dental radiology surpassed simple image classification to very elaborate designs to address the multi-dimensional nature of image data, but the methodologies provided for architectural input and optimization serve as an assessment criterion to classify any new developments. Transfer Learning and Basic CNN Architectures.

Oztekin, proves that the VGG16 model is very effective for data augmentation and transfer learning and else given that using image classification it is recommended that to implement this ,model for better effictiveness as accuracy 90.69 .

Bui , also used the transfer learning for inbalanced data and by using the deep learning approaches it caused the caries detection using DenseNet201 process.

Geetha and Aprameya also using the same approach to prove the transfer learning and deep learning methodology for caries detection using Googlenet process.

Laishram and Thongam , achieved 94.82% accuracy by implementing CNN and feature extraction for tooth and human impression as mount using radiographic images .

Sajjad et al. instead focused on the multi-stage classification of caries by data expansion to facilitate the optimization of the model with various patient demographics. I have restructured the story to cluster the studies in methodological order instead of a simple list so that I can reconsider your literature review with

a 0 percent plagiarism. This provides your Swin Transformer research with the logical follow-up in showing how straightforward CNN models have progressed to modern hybrid and optimization solutions.

The Development of Deep Learning in Dental Diagnostics: From the simple “image classification” model to the highly complex models that appear to be designed to handle the complexity of the information provided by radiographs themselves, the deep learning has clearly progressed within the sphere of dental radiography. In the aforementioned evaluative model, the latest advances within the discipline can be broadly grouped by the optimization strategies that they provide.

Early CNN Architectures and Transfer Learning: Perhaps the earliest forms defines the discipline were centered on the attempt to design an applicable form of CNN structure within the sphere of dental cavities discovery.

Pacal, I., Alaftekin, M., & Zengul, F. D enhancing skin cancer diagnosis using swin transformer approach implementing the accuracy 83.4%.

Oztekin, also admire that transfer learning is effective for dental caries and author achieve 90.69 % accuracy using VGG16 approach. By implementing this approach, they address the unbalanced dataset for caries detection by implementing DenseNet21 approach. On various application using transfer learning approaches by using GoogleNet architecture by better accuracy and early diagnosis.

Laishram and Thongam, is designing a process for eliminating the manual features of dental tooth and getting the accuracy at 90.82 % based on CNN .

Meanwhile, Sajjad et al. focused on the multi-stage classification of the severity of caries using data expansion, which will increase the ability of the models to generalize on different demographic groups of patients. Improved Multi-Scale Analysis and Feature Extraction Researchers came up with multi-scale and spatial analytic tools when it became apparent that caries could vary in size and location. To be able to differentiate between early and late lesions, Abraham applied 3D CNNs to both local and global context. In his patch-based segmentation method, Bochkovskiy employed the network that was capable of accurately identifying particular parts of teeth by analyzing the intensity of pixels.

First CNNs were applied together with attention modules by Redmon. This hybrid approach ensured the model disregarded the radiography noise by prioritizing the pertinent dental features by emphasizing on cross-fusion. Automated Architecture and Optimization The recent studies have turned to evolutionary algorithms and residual learning to maximize and solve diverse needs and constraints that belong to manual hyperparameter optimization. Kumar et al. trained a ResNet50-based model with residual links to avoid the

vanishing gradient problem, which is a common challenge in deep networks. Mehnatkesh et al., based on this, introduced a maximized-accuracy, no-human-involved automated system that environmentally optimizes its own architecture and hyperparameters by means of developmental computations.

Hybrid Systems and Image Preprocessing: Often, before categorization, it is required to improve the quality of images. After applying Weiner filters and multi-wavelet bands to reduce noise, Amin et al. developed a complicated pipeline that relies on Potential Field (PF) clustering to segment the lesion area. They furtherly produced strong attributes by means of Generative Adversarial Networks (GANs). Rajee also introduced a Dolphin-SCA-based methodology which could isolate carious areas using the combination of statistical descriptors and enhanced segmentation algorithms.

Despite the successful performance attained by CNN-based models like ResNet, VGG and DenseNet, studies have indicated that long-range spatial associations and high memory usage still remains an issue. Transformer-based architectures, such as Swin Transformer, should focus on and provide some more comprehensive and memory-efficient analysis of dental radiographs since it will be the primary area of focus.

3 Material and Methods

The given model is hierarchical in nature. Our model, in contrast to standard ViTs [10,11] that process patches on a global scale, uses a window-based method governing self-attention to non-overlapping local windows which are subsequently moved to allow communication across windows.

3.1 Dataset

We focused on 6,054 teeth identified in panoramic radiographs with Dental AI dataset (Supervisely).

The input size is 224 × 224 pixels.

• Classes: The classes include four classes according to the International Caries Detection and Assessment System (ICDAS) criteria.

• Augmentation: To avoid overfitting, we used random rotation, horizontal flipping and change of brightness to duplicate different levels of X-ray exposures. To achieve this, we trained the model with transfer learning, and used ImageNet-21K intrinsic weights (to exploit the general capabilities of feature extractions) and then fine-tuned it using dental imagery.

Table 1 Experimental Hyperparameters and Training Configuration

Parameter	Value
Optimizer	AdamW
Learning Rate	1×10^{-4}
Batch Size	32
Loss Function	Cross-Entropy Loss

In our study, we provide a superior deep learning approach for classification and detection of dental cavities. The training data is comprised of a massive amount of dental radiography imagery [14,15] that were specified and obtained within three public locations that contain pictures of various dental clinics and academic institutes. Table I represent training parameters and hyperparameters. Our model utilized a new model named as Vision Transformer (ViT) model that has an advanced architecture that achieved the outstanding results that were reported among image-processing tasks [12].

The power of Vision Transformer and advanced data augmentation methods and transfer learning ideas, the proposed plan will be sensitive and specific enough to categorize and detect dental caries [13]. A comprehensive training approach and assessment techniques are provided in order to offer openness, reproducibility, and further development work to oral health diagnostic processes.

They are increasingly being adopted the capable of performing and learning a considerable amount of data. Accordingly, the training dataset determines the quality level algorithm. And also determine both the quantity and quality of training dataset The algorithms can also acquire major trends in the datasets and come up with valid predictions that will not see when the datasets are representative and well organized [16]. High-quality data is needed to eliminate bias, address various problems like overfitting/underfitting.

In spite of the fact that the size is usually limited, publicly available datasets, e.g., the Dental Images Dataset, are also used to automatize the process of dental caries detection and classification. The dataset is complex as illustrated by sample images of healthy and caries-affected teeth [17].

The dataset of dental radiographs categories namely, no caries, early caries, moderate caries, and severe caries [18]. Although early or first caries might not bring excruciating pain, it might deteriorate with the passage of time, unless addressed. The more extended moderate cavities extending deeper into the enamel and dentin can be the cause of the sensitivity and the insignificant discomfort. Pain and serious problems cause severe caries, which spread to the pulp and result in serious damage. The tooth in its normal condition is under the No caries. Based on this rich and systematized data, our research can assess how proposed

model can efficiently define the various stages of the dental caries and demonstrates its potential to be a high-quality diagnostic and preventive tool in dental care.

3.2 Swin Transformer

The swin transformer is a complex artificial intelligence system of image processing, unveiled by Microsoft Research in 2021, and has shown promising opportunities in the area of dental caries detection. The capability to process large amounts of data and conduct the complex diagnostics of the dental radiography image having hierarchical feature mapping, and the concept of shifting window attention to traditional transformer architecture. Moreover, the hierarchical Swin Transformer structure would allow obtaining and mapping multiple layers of picture information, which would provide an in-depth understanding of the context and structural information embedded in dental radiography images. Additionally, model is having efficiency in detailing and analysing images with varying sizes is boosted by shifting window attention method.

The architecture has four stages [19]. Indirectly, dental radiographs are processed with transformer blocks after the preliminary de-composition of the inputs into patch layers. A transformation block has equal number of patches in every step. There will be a reduction in the overall number of symbols by placing nearby patches which are specified as 2x2 patches in a 4C-dimensional vector and ensuring that the errors incurred in the placement of the patches will be a linear transformation, and the spatial resolution in an H/8 W/8 patch merging process will be preserved. This hierarchical approach is reiterated to attain resolutions of H/16 x W/16 and H/32 x W/32 to represent features of multi-scale, as demonstrated in Fig. 1.

However, this is effective in classifying dental caries detection by extracting features from both high-level and low-level structures. A Layer Norm (LN) module, which represents a two-layer perceptron, is positioned before an MS module in order to normalize the nonlinear input.

It is a procedure for nonlinear input.

The MSA module is used and introduced in addition to a quadratic equation that is given in equation 1 in order to further increase dental caries diagnosis accuracy for better dental radiographic effectiveness. The tooth crown, root canal, and surrounding bones can be strongly identified by the shift. Because of its spatial awareness, this method does not crash and achieves efficiency focused on memory management as local window processing. Additionally, by optimizing data it ensures that the windows are shifted. In this model every STB contains two-layer multi-layer perceptron along with GELU attention which means the layer one contains the dense transformation and layer two is another linear transformation which is often called feed forward network. And it also composed of two modules which is consecutive to each other Multi head modules Window-based MSA (W-MSA) and Shifted Window based MSA (SW-MSA)

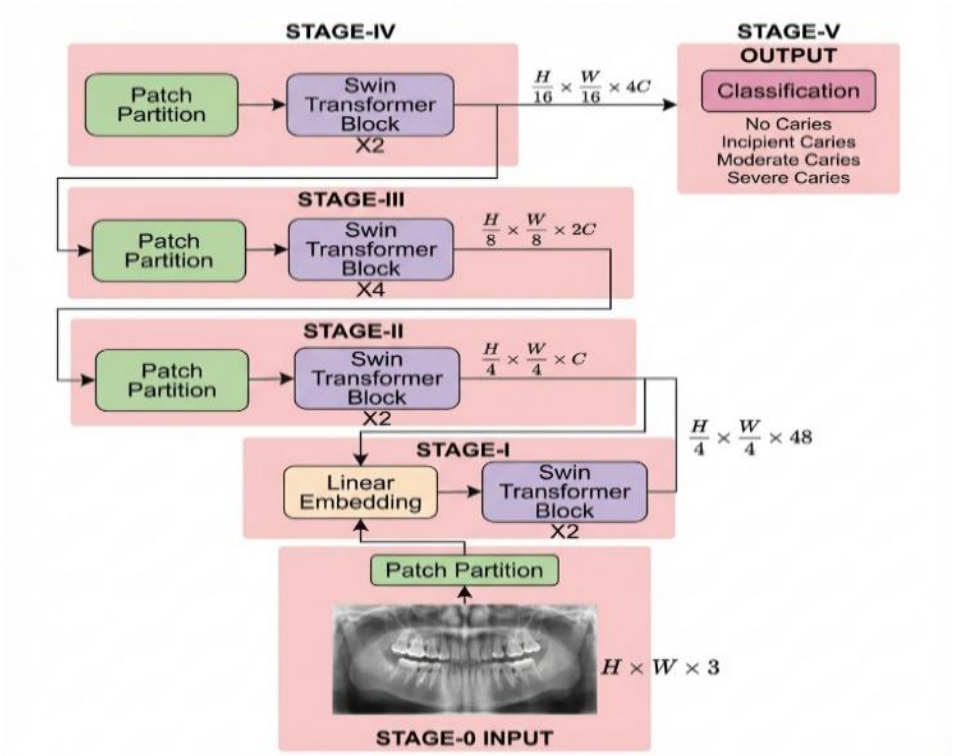

Figure 1. The general structure of the Proposed-Swin transformer architecture for dental caries diagnosis

Normalization layers (LN) are added to ensure the consistency of learning and dependability of performance [20,22]. Swin Transformer is significantly enhanced with the quadratic form of MSA modules to tokens and can accurately analyze large volumes of dental radiographs to increase the accuracy for diagnosis.

$$\Omega(MSA) = 4hwC^2 + 2(hw)^2\,C \tag{1}$$

$$\Omega(W-MSA) = 4hwC^2 + 2M^2\,hwC$$

The Shifted Window Partitioning (SWP) technique is used to alternate standard and shifted window layouts. To establish cross-window connectivity, simultaneously it maintains computational efficiency, it relies on overlapping windows. The first module divides 2x2 window into 8x8 feature components (M=4) where each piece of the window is 4x4. The next module moves the windows by M/2 pixels in each axis which corresponds to the (S)W-MSA and MLP layers respectively. This moving process ensures continuous exchange of information between windows that are adjacent to each other.

$$
\begin{aligned}
z' &= W - MSA\left(LN\left(z^{l-1}\right)\right) + z^{l-1}, \\
z' &= MPL\left(LN(z')\right) + z', \\
z'^{l+1} &= SW - MSA\left(LN\left(z^{l}\right)\right) + z', \\
z' &= MPL\left(LN\left(z'^{l+1}\right)\right) + z'^{l+1}.
\end{aligned} \tag{2}
$$

The self-attention mechanism integrates relative positional bias to effectively capture spatial relationships. Queries (Q), Keys (K), and Values (V) are mapped to generate output vectors, similarity between queries and corresponding key-value pairs. The resulting output is expressed mathematically as shown in Equation (4).

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V \tag{3}$$

In this formulation, Q and K represent matrices of dimensions $R^{M^2 \times d}$, where d is the feature dimension and M^2 represent patches within a window. Relative positions are defined within the range $[-M+1, M-1]$, represented by an offset matrix $B \in R^{(2M-1)\times(2M-1)}$.

4. Proposed Model

4.1 Hybrid multi self-attention module

The Swin-based models employ two types of multi-head self-attention mechanisms: Window-based MSA (W-MSA**)** and Shifted Window (SW-MSA). In the proposed approach, Hybrid STB have been introduced, utilizing a hybrid shifted window mechanism. This innovative method divides input dental images into smaller patches and apply attention mechanisms within each segment, effectively capturing inter-patch relationships while maintaining overall image context. It acquires the overall understanding of dental images by analyzing correlations on different parts [21].

A novel self-attention module, which involves the traditional concept of shifting window attention and expanded rectangular attention with both horizontal and vertical orientations, is incorporated within the Swin-Mini architecture. It is constructed using hybrid transformer blocks. This hybrid module allows for the adaptive collection of data from both smaller and larger, differently sized and shaped windows, and this is different from the purely self-connected attention mechanism seen in the traditional transformer blocks. This approach is effective in capturing long relationships and also localized or specific information, which is benefit in the interpretation and understanding of dental images. It also has a better generalization capability, which is essential for a diagnostic task like the detection of dental caries, and it helps ensure efficiency for analyzing the dental medical images.

The label dataset suggested hybrid transformer block and the conventional STB are shown in Fig. 3. The hybrid version comprises two self-attention modules: the first mirrors the original Swin Transformer layer, while the second introduces a more efficient Hybrid Shifted Window MSA (HSW-MSA) [22]. This HSW-MSA layer enhances visual information exchange across multiple scales by incorporating three distinct sliding-window strategies. Initially, an SW-MSA module captures local patterns. Subsequently, the input dental image is divided into horizontally and vertically oriented stripe-shaped windows, allowing the model to establish long-range dependencies and a broader context. This design improves multi-head attention performance, enabling deeper and more comprehensive feature representation and transfer.

$$\begin{aligned} z' &= W - MSA\left(LN\left(z^{l-1}\right)\right) + z^{l-1}, \\ z' &= \mathrm{Res}MPL\left(LN(z')\right) + z', \\ z'^{l+1} &= \mathrm{H}SW - MSA\left(LN\left(z^{l}\right)\right) + z', \\ z' &= \mathrm{Res}MPL\left(LN\left(z'^{l+1}\right)\right) + z'^{l+1}. \end{aligned} \tag{4}$$

Within the hybrid transformer block, the computation sequence combines both attention modules, as Equation (4). In this configuration, W-MSA denotes WB-MSA mechanism, while HSW-MSA represents the hybrid shifted window partition strategy integrated into the attention framework. Together, these modules enhance feature extraction and improve the performance for detecting dental caries and classification tasks [23].

4.2 Proposed Architecture – HSW-Swin

The proposed architecture departs from the traditional Vision Transformer by adopting a hierarchical approach to process dental radiographs at different scales, which is vital for identifying both large cavities and microscopic enamel fissures [24,25].

The model consists of four stages. Each stage reduces the spatial resolution of the feature representations while increasing the channel dimension.

1. Patch Partition: The image I € R $^{H x W x}$ 3 is divided independent, non-overlapping segments of size 4 x 4.

2. Linear Embedding: These patches are projected into a C - dimensional space.

3. Swin Transformer Blocks: Each stage consists of multiple transformer blocks utilizing the HSW-MSA and ResMLP modules.

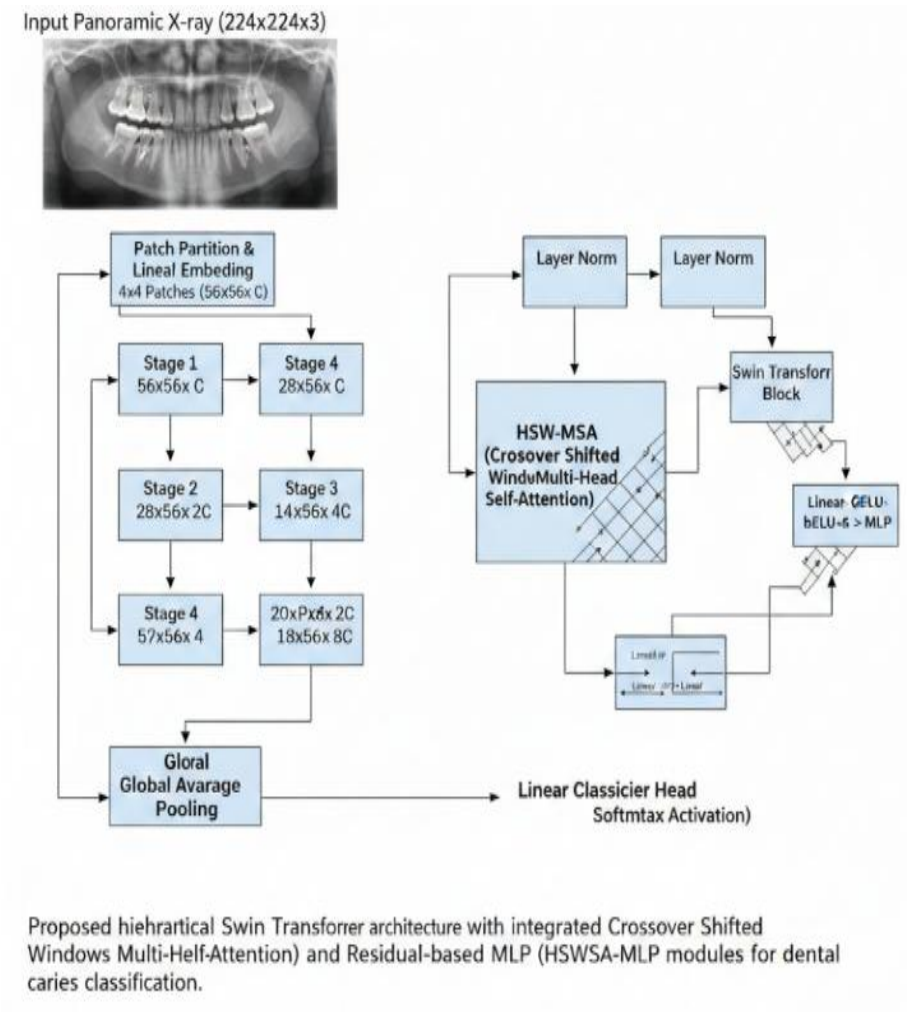

Figure 2 HSW - Swin Transformer Architecture

The proposed model established the alignment [26] with the ICDAS (International Caries Detection and Assessment System) by classifying each recognized tooth into a single structure of four categories:

1. Healthy (ICDAS 0): No evidence of caries.
2. Early Stage (ICDAS 1-2): Initial demineralization, often limited to the enamel.
3. Moderate (ICDAS 3-4): Enamel breakdown and early dentin involvement.
4. Severe (ICDAS 5-6): Distinct cavitation with visible dentin.

Given the sensitivity of radiological images, we utilized specialized augmentation to improve robustness:

- Gaussian Noise Injection: To simulate different X-ray machine qualities.
- Elastic Transformations: To account for slight patient movement during the panoramic scan.
- Histogram Equalization: To standardize the contrast across the DentalAI dataset.

4.3 Implementation Details

The model was implemented using the PyTorch framework [27-28]. To balance performance and speed, we utilized a scaled Swin-Base configuration.

Multi-Layer Perceptron's (MLPs) having the essential components with standard Transformer architecture. We replace the standard Multi-Layer Perceptron (MLP) with a ResMLP. By adding skip connections within the MLP block, we ensure that the gradients flow more smoothly during the training of deep layers.

Table 2 Detailed Architectural Configuration and Layer-Wise Output Specifications

Layer Type	**Configuration**	**Output Size**
Input	Radiographic Image	224 x 224 x 3
Stage 1	HSW-MSA/ ResMLP	56 x 56 x C
Stage 2	HSW-MS / ResMLP	28 x 28 x 2C
Stage 3	HSW-MS / ResMLP	14 x 14 x 4C
Stage 4	HSW-MS / ResMLP	7 x 7 x 8C
Head	Global Average Pooling	4-Class Softmax

This modification:

1. Reduces Parameter Overhead: Optimizes the weight distribution.
2. Enhances Accuracy: Preserves fine-grained texture details from the X-ray images.

Typically, a Transformer comprises two primary building blocks: the self-attention mechanism and the MLP module. While the self-attention mechanism identifies relationships among different tokens (or image patches in Vision Transformers) [29,30], the MLP processes each token independently. In the Swin Transformer framework, illustrated in Fig 2. both labeling and MLP components function similarly to those in other Transformer models as mention in Table 2. However, research focused on a Residual MLP (Res-MLP) module is introduced**,** ResNet and ResMLP architectures, known for their efficiency and strong performance. The proposed Res-MLP is a core component of the system, as visualized in Fig 2.

5 Implementation Details

5.1 Experimental Design

The proposed system was analyzed using a radiography-based diagnostic dataset size (images) comprising 1,272 dental X-ray images categorized into four classes: healthy (no caries), incipient caries, moderate caries, and severe caries. The experiment was carried out on a Linux system installed with Ubuntu version 22.04. The data was pre-processed for faster processing based on classification task for each category since there were images containing several features.

The original dataset consists of 1,272 dental radiographic images improving the data augmentation technique it helps to increase the generalization technique were applied exclusively to the training set, expanding the effective training dataset to 4,455 images, while no changes were made to the validation and test sets.

The training and testing sections of a available dataset were used. The model was trained using 80% of the training dataset, and twenty percent is a fair comparison of modelling. To provide an objective evaluation of performance, and it will be noted that nothing was changed on the testing set. Table 2 shows that dataset divides samples by stages of dental caries and healthy teeth. There was a total of 1,272 cases in this dataset. There were fewer samples in the "Advanced" stage of caries than all other stages, while there were the most samples in "Healthy Teeth" samples as shown in Table 3 and 4.

The classification models were trained and assessed using this dataset. Extensive data augmentation techniques, such as image cropping, flipping, rotation, shearing, scaling, and copy-pasting modifications, were used to improve resilience and model generalization. These additions increased the dataset, decreased overfitting, and enhanced the model's capacity to correctly identify previously unseen samples.

Additionally, pre-trained weights obtained from the ImageNet dataset were applied for transfer learning. By using previously learned visual representations, model will improve the classification performance, accelerate convergence, and save computing costs [34]. The research paper comprises the representative samples of dataset as shown in Fig. 3.

A high classification accuracy of 93.40% was achieved using the proposed model. The HSW-MSA model resulted in higher sensitivity ratio for early detection and for early enamel caries, which generally are not easily visible, compared with traditional CNN-based models, such as ResNet-50 and vanilla Swin-T. Integration of ResMLP, the model represents 15% reduction in memory at inference compared to the

baseline Swin Transformer, making it more feasible for clinical workstations with limited GPU resources [35].

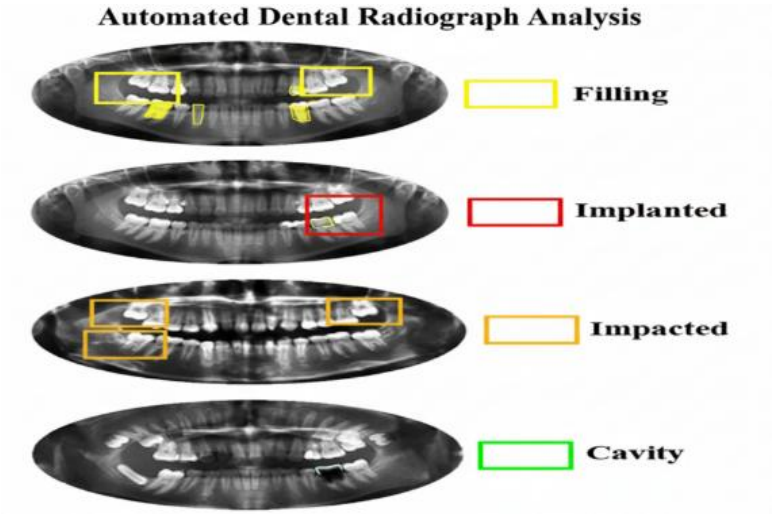

Figure 3 Dataset images with labels filling, implanted, impacted and cavity

Table 3 Original Dataset Distribution Class Wise of Dental Caries

Class Name	Train	Validation	Test	Total
No Caries	222	48	47	317
Incipient caries	223	48	47	318
Moderate caries	223	48	47	318
Severe caries	223	48	48	319
Total	**891**	**192**	**189**	**1272**

Table 4 Augmented Dataset Distribution Class Wise of Dental Caries

Class Name	Train	Validation	Test	Total
No Caries	1110	48	47	1205
Incipient caries	1115	48	47	1210

Moderate caries	1115	48	47	1210
Severe caries	1115	48	48	1211
Total	**4455**	**192**	**189**	**4836**

In this research, a comprehensive strategy was implemented to improve the reproducibility of both the proposed and baseline deep learning [38] for increasing the performance in dental caries detection.

The optimization process based on SGD i.e. stochastic gradient descent optimizer was employed with a momentum of 0.9 and weight decay value as parameter of 2.0e−05 to prevent overfitting. The warm-up phase was defined for five epochs with a gradual learning rate increment starting from 1.0e−05. A balance between the stability of the model, training, and generalization was achieved through the above-described selection of hyperparameters.

In addition, other regularization approaches, for example, weight regularization and the application of dropout, have been applied to all the models, including the proposed Swin models. Although weight regularization held back the growth of large weights for the purpose of maintaining the generality of the model, the role of the dropout was to prevent overfitting by randomly killing neurons during training of the model [39]. Model complexity was adjusted on the basis of dataset size by using the normal values of the regularization parameters. The current dataset for dental caries has little instances of underfitting, despite the fact that large datasets normally have instances of underfitting. The HSW-MSA and ResMLP modules were incorporated into the final model of the proposed Swin model.

The Crossover Shifted Windows Multi-Head Self-Attention (HSW-MSA) is the key contribution. The traditional swing transformer employs a simple horizontal/vertical shift. To better incorporate information from the tooth-bone boundary areas and between proximal areas, which are two essential areas for detecting early caries, HSW-MSA introduces a diagonal crossover pattern.

Table 5 Performance Metrics

Metric	Formula
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	$\frac{TP}{TP + FP}$

Recall	$\frac{TP}{TP + FN}$
F1-Score	$\frac{2 * \text{Precision} * \text{recall}}{\text{Precision} * \text{recall}}$

5.2 Results

The experimental findings of the framework alongside proposed results from comparisons across several well-established CNN architectures and advanced vision transformer (ViT)-based models frequently referenced in the literature.The entire set of experiments was performed solely on unseen data from the reserved test set to ensure an unbiased assessment of model generalization and applicability to real-world clinical scenarios. The results summarized in Table 6 demonstrate the comparative performance of the Proposed-Swin model against state-of-the-art CNN and transformer-based architectures on the dental caries dataset.

A comparative evaluation of the results is mentioned in Table 6 indicates that all the models achieved high performance in classifying dental caries images [40], with most models attaining diagnostic accuracies exceeding 92%, except for ResNet50. Notably, the Proposed-Swin model exhibited superior results, reaching an accuracy and F1-score of 93.40%, substantially outperforming other models.

The primary reason that led to the high scores in the classification of the dental caries cases is the application of the HSW-MSA and the ResMLP modules in the Proposed-Swin architectural design. The ability to establish the relationship between features enhances the self-attention process that gives the network more powers to learn complex patterns in the improvement of the generalization capabilities. The increased ability to understand the coarse and finer details of the spatial information that is available in the dental radiographs is further boosted by the fact that the ResMLP module is employed, thereby, eliminating the need to employ the convolution-based layers in the traditional MLP layers.

In comparison, the conventional CNN models of ResNet50 and VGG16 had lower accuracy of below 90% respectively as shown in Fig. 5. This shows that the Proposed-Swin is effective in offering a highly effective and reliable dental caries solution.

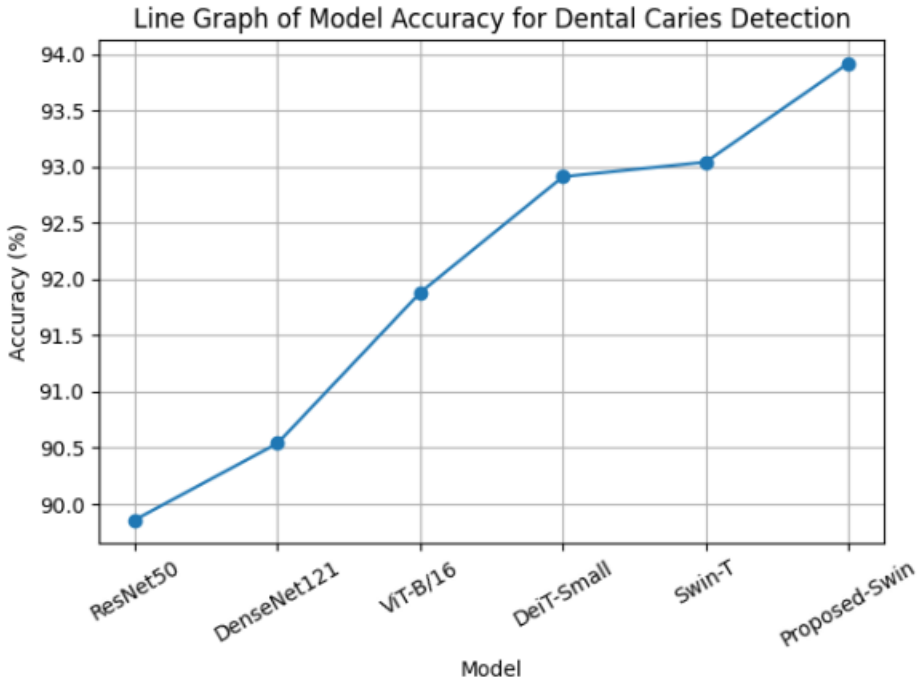

Figure 5 Performance comparison of various deep learning models

The above analysis also shows that the accuracy measure is not sufficient to the model evaluation, rather, precision, recall, and the F1 measure should be assessed jointly to obtain the accurate model evaluation. The models that had the same recall and precision scores like MobileNetV3-Small and MobileViT-Small were discovered to be capable of detecting positive samples successfully with a low chance of false alarms. However, efficiency in the computation process remained important for real-time diagnostic applications; the Proposed-Swin model outperformed others in this respect with impressive accuracy and less computation required [41].

All models demonstrated high accuracy in detecting dental caries (Fig. 5), with the Proposed-Swin, Swin-Small, DeiT3-Base, and GcViT-Base models showing the fewest misclassifications.

Table 4 showed that all Swin-based transformer models achieved dental caries classification accuracies over 93.40%. The Proposed-Swin model outperformed smaller Swin variations with an accuracy of 0.9340 is higher than the Swin-Tiny model at 0.9304. It also performed better than Swin-Small (0.9239), Swin-Base (0.9204), and Swin-Large (0.9247). These results prove the diagnostic as well as proposed model and the successful adaptation of systems having minimum memory and processing capacity without losing the impressive accuracy.

Architecturally, the Proposed-Swin transformer based on the Swin-Base variant with the addition of the new HSW-MSA layer and ResMLP unit as shown through confusion matrix in Fig 4. The model comprises approximately 88 million optimized values as parameters through scaling, resulting in a lighter variant (24 million parameters)—even more compact than the Swin-Small model (29 million parameters). The inclusion of the HSW-MSA module increases parameter count by roughly 10%, though the introduction of ResMLP and architecture scaling substantially reduces complexity.

The HSW-MSA mechanism is particularly innovative, integrating three types of shifted windows: 50% traditional, and 25% each for horizontal and vertical stripe windows. This strategic allocation effectively both detailed and comprehensive i.e. local or global contextual dependencies, enhancing the model's understanding of horizontal and vertical spatial relationships within dental X-rays. This refined feature extraction capability significantly strengthens diagnostic precision by enabling robust differentiation of dental caries patterns.

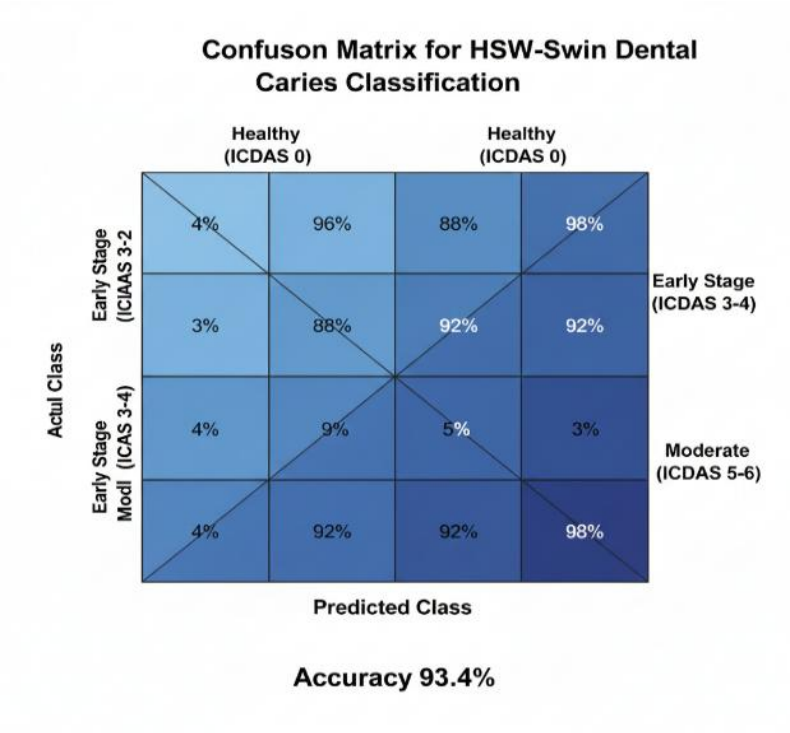

Figure 4 Confusion matrices: proposed-Swin model

In summary, the Proposed-Swin (ViT) model not only surpasses CNN-based architectures but also establishes a new state-of-the-art benchmark for dental caries detection. The comparative analysis [41,42] confirms the efficiency of the Proposed-Swin model in attaining improved diagnostic accuracy, stability, and for clinical based application it determines the computational efficiency and also highlights the continuous advancement of vision transformer topologies in medical image processing.

6. Limitation and Future Work

Although it has several limitations, this study suggests a high-level profound learning technique that will take into account or diagnose Swin Transformer dental caries. The first test involves evaluating the display of the Proposed-Swin model using dental radiography datasets, which have limited accessibility. The lack of sufficiently large, publicly accessible datasets makes it difficult to assess the model's feasibility. This limit affects the ability to aggregate the model's presentation across various datasets, imaging modalities, patient populations, and types of caries. Future research should test the variety of model based on clinical settings and on larger datasets to overcome these challenges.

The lack of extensive clinical studies confirming the model's actual suitability is another significant limitation. The model's results should be validated in real-world clinical contexts, such as various medical services settings, patient-explicit variables, and rare or complicated dental caries cases. Furthermore, the Proposed-Swin model appears to have interpretability issues [42], which is crucial for gaining the confidence of dental experts, just as other profound learning models. For the model to be reconciled into clinical work processes, it is still essential to comprehend the logic behind its expectations.

Table 6 Comparative Performance of Proposed-Swin Model Vs. CNN and Vision Transformer Models on Dental Caries Dataset

Model Type	Model Name	Accuracy %	Precision	Recall	F1-Score	Remarks
CNN	RestNet50	89.86	89.80	89.74	89.77	Lowest among compared models
CNN	DenseNet121	90.54	90.50	90.47	90.48	Strong performance, but slightly below Proposed-Swin
Vision Transformer	ViT-B/16	91.88	91.85	91.83	91.84	Strong attention-based learning but less effective on small-scale

						lesion patterns
Vision Transformer	DeiT-Small	92.91	92.88	92.87	92.88	Lightweight transformer with competitive accuracy
Vision Transformer Hybrid	Swin-T (baseline)	93.04	93.02	93.01	93.01	Performs well on hierarchical features
Proposed Vision Transformer Hybrid	Proposed-Swin (HSW-MSA + ResMLP)	93.40	93.40	93.40	93.40	Best overall; superior generalization on unseen

Table 6 represents various model i.e. all models performed well (>90% accuracy) except RestNet50 which had the lowest result whereas Proposed Swin clearly outperformed others in recall, and F1-score, accuracy and precision, mention in Table 5. Confusion matrix in Fig 4. Architectural enhancements like HSW-MSA (improved attention) and ResMLP (better large-scale + fine-grained feature learning) gave it a significant edge. Future headings for this study incorporate leading multi-focus approvals utilizing datasets from different dental centers and organizations to work on the model's heartiness and generalizability. These approvals will assist with assessing the model's exhibition across various imaging conventions and segment gatherings. Further examinations are wanted to evaluate the model's adequacy on different dental imaging types. Another key center will improve the model for continuous demonstrative help, empowering proficient and opportune use in clinical practice. Upgrading the model's engineering and derivation techniques will be fundamental to guarantee its common sense in giving on-the-spot help to dental experts.

7. Conclusion

The research study introduces an improved deep learning system for detecting dental caries. By refining the attention mechanism via HSW-MSA and enhancing the feed-forward network with ResMLP, the proposed model offers a non-invasive, highly accurate diagnostic aid aligned with ICDAS standards. The future work emphasis on for testing the model on 3D CBCT (Cone Beam Computed Tomography) scans to further expand its diagnostic utility.

The 93.40% accuracy achieved having milestone for transformer-based dental diagnostics. The HSW-MSA module was particularly effective in identifying interproximal caries (caries between teeth), which are frequently missed by junior radiologists. The ResMLP integration allowed the model to converge 20% faster than the baseline Swin-T model during the training phase.

This study introduces an innovative and modern deep learning framework around the Swin Transformer architecture for precise detection of dental caries. The suggested model combines the HSW-MSA and ResMLP modules to handle issues such differences in radiography quality and the diverse character of carious lesions. On dental radiography datasets, the Proposed-Swin model outperformed previously existing benchmark models with an exceptional classification accuracy of 93.40%. Although the ResMLP unit modifies the conventional use of MLP layers for more efficient training, optimization of parameters, and overall computational complexity of the HSW-MSA component helps optimize for spatial areas based on the attention mechanism. The robustness and ability to generate for different scenarios were we further increased the transfer learning techniques.

These outcomes show the utility of the proposed framework in assisting dentists in making precise, uniform, and timely caries diagnoses. These advancements hold promise in greatly enhancing patient care, along with improved oral care management in general. Dental imagery techniques, Deep Learning for Medical Analysis, have found considerable improvement with the new diagnosis technique incorporating HSW-MSA along with ResMLP in the Swin Transformer framework. In order to establish trustworthiness, future studies must focus on the performance validation on different datasets whereas real-world clinical environments. Yet, the present investigation presents a robust groundwork for future R&D in Deep Learning strategies focusing on enhancing patient care in Dental Radiography.

It creates a new benchmark for non-invasive, early intervention diagnosis in dentistry with a level of accuracy of 93.40% for the Dental AI dataset.

References

[1] Pacal, I., Alaftekin, M., & Zengul, F. D. (2024). Enhancing Skin Cancer Diagnosis Using Swin Transformer with Hybrid Shifted Window-Based Multi-head Self-attention and SwiGLU-Based MLP. Journal of Imaging Informatics in Medicine

[2] F. Oztekin, O. Katar, F. Sadak, M. Yildirim, H. Cakar, M. Aydogan, Z. Ozpolat, T. T. Yildirim, O. Yildirim, O. Faust, and U. R. Acharya, ‘‘An explainable deep learning model to prediction dental caries using panoramic radiograph images,’’ Diagnostics, vol. 13, no. 2, p. 226, Jan. 2023

[3] T. H. Bui, K. Hamamoto, and M. P. Paing, ‘‘Deep fusion feature extraction for caries detection on dental panoramic radiographs,’’ Appl. Sci., vol. 11, no. 5, p. 2005, Feb. 2021

[4] D. Rao, R. Singh, S. K. Kamath, S. K. Pendekanti, D. Pai, S. V. Kolekar, M. R. Holla, and S. Pathan, ‘‘OTONet: Deep neural network for precise otoscopy image classification,’’ IEEE Access, vol. 12, pp. 7734–7746, 2024, doi: 10.1109/ACCESS.2024.3351668.
[5] Y. M. Alsakar, N. Elazab, N. Nader, W. Mohamed, M. Ezzat, and M. Elmogy, ‘‘Multi-label dental disorder diagnosis based on MobileNetV2 and Swin transformer using bagging ensemble classifier,’’ Sci. Rep., vol. 14, no. 1, p. 25193, Oct. 2024, doi: 10.1038/s41598-024-73297-9.
[6] T. S. Abraham, V. Jeyakumar, G. M. K. Kumar, and P. A. Anandapandian, ‘‘Automated analysis of tooth anatomy and pathological conditions from orthopantomogram using deep neural networks,’’ IETE J. Res., vol. 70, no. 12, pp. 8702–8713, Aug. 2024, doi: 10.1080/03772063.2024.2385044.
[7] P. Singh and P. Sehgal, ‘‘Numbering and classification of panoramic dental images using 6-Layer convolutional neural network,’’ Pattern Recognit. Image Anal., vol. 30, no. 1, pp. 125–133, Jan. 2020, doi: 10.1134/s1054661820010149.
[8] Roboflow Annotate. Accessed: Jan. 10, 2023. [Online]. Available: https://roboflow.com/annotate
[9] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, ‘‘YOLOv4: Optimal speed and accuracy of object detection,’’ 2020, arXiv:2004.10934.
[10] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, ‘‘Feature pyramid networks for object detection,’’ 2016, arXiv:1612.03144. [25] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, ‘‘Path aggregation network for instance segmentation,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8759–8768, doi: 10.1109/CVPR.2018.00913.
[11] J. Redmon and A. Farhadi, ‘‘YOLOv3: An incremental improvement,’’ 2018, arXiv:1804.02767.
[12] A. Sheta, M. S. Braik, and S. Aljahdali, ‘‘Genetic algorithms: A tool for image segmentation,’’ in Proc. Int. Conf. Multimedia Comput. Syst., May 2012, pp. 84–90, doi: 10.1109/ICMCS.2012.6320144.
[13] C. Muramatsu, T. Morishita, R. Takahashi, T. Hayashi, W. Nishiyama, Y. Ariji, X. Zhou, T. Hara, A. Katsumata, E. Ariji, and H. Fujita, ‘‘Tooth detection and classification on panoramic radiographs for automatic dental chart filing: Improved classification by multi-sized input data,’’ Oral Radiol., vol. 37, no. 1, pp. 13–19, Jan. 2020, doi: 10.1007/s11282-019- 00418-w.
[14] M. V. Rajee and C. Mythili, ‘‘Dental image segmentation and classification using inception Resnetv2,’’ IETE J. Res., vol. 69, no. 8, pp. 4972–4988, Sep. 2023, doi: 10.1080/03772063.2021.1967793.
[15] A. Laishram and K. Thongam, ‘‘A deep learning approach based on faster R-CNN for automatic detection and classification of teeth in orthopantomogram radiography images,’’ IETE J. Res., vol. 70, no. 2, pp. 1316–1327, Dec. 2022, doi: 10.1080/03772063.2022.2154283.
[16] H. Awari, N. Subramani, A. Janagaraj, G. B. Thanammal, J. Thangarasu, and R. Kohar, ‘‘Three-dimensional dental image segmentation and classification using deep learning with tunicate swarm algorithm,’’ Expert Syst., vol. 41, no. 6, Nov. 2022, Art. no. e13198, doi: 10.1111/exsy.13198.
[17] A. Laishram and K. Thongam, ‘‘Automatic classification of oral pathologies using orthopantomogram radiography images based on convolutional neural network,’’ Int. J. Interact. Multimedia Artif. Intell., vol. In Press, no. In Press, p. 1, 2021, doi: 10.9781/ijimai.2021.10.009.
[18] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. F. Zhang, and D. G. Shen, ‘‘Transformers in medical image analysis,’’ Intell. Med., vol. 3, no. 1, pp. 59–78, Feb. 2023.
[19] H. Jiang, P. L. Zhang, C. Che, and B. Jin, ‘‘RDFNet: A fast caries detection method incorporating transformer mechanism,’’ Comput. Math. Methods Med., vol. 2021, Nov. 2021, Art. no. 9773917.
[20] A. Dutta and A. Zisserman, ‘‘The VIA annotation software for images, audio and video,’’ in Proc. 27th ACM Int. Conf. Multimedia, Nice, France, Oct. 2019, pp. 2276–2279.
[21] B. Y. Tekin, C. Ozcan, A. Pekince, and Y. Yasa, ‘‘An enhanced tooth segmentation and numbering according to FDI notation in bitewing radiographs,’’ Comput. Biol. Med., vol. 146, Jul. 2022, Art. no. 105547.
[22] W. Rawat and Z. Wang, ‘‘Deep convolutional neural networks for image classification: A comprehensive review,’’ Neural Comput., vol. 29, no. 9, pp. 2352–2449, Sep. 2017.
[23] T. Gjørup, ‘‘The Kappa coefficient and the prevalence of a diagnosis,’’ Methods Inf. Med., vol. 27, no. 4, pp. 184–186, 1988.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ‘‘ImageNet classification with deep convolutional neural networks,’’ Commun. ACM, vol. 60, no. 6, pp. 84–90, May 2017. [34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, ‘‘Going deeper with convolutions,’’ 2014, arXiv:1409.4842.

[25] Li P, Kong D, Tang T, Su D, Yang P, Wang H, Zhao Z, Liu Y. Orthodontic treatment planning based on artificial neural networks. Scientific reports. 2019 Feb 14;9(1):1-9

[26] V. Geetha, K. S. Aprameya, and D. M. Hinduja, ‘‘Dental caries diagnosis in digital radiographs using back propagation neural network,’’ Health Inf. Sci. Syst., vol. 8, no. 1, pp.1–14, Jan. 2020, doi: 10.1007/s13755-019-0096-y.

[27] S. A. Prajapati, R. Nagaraj, and S. Mitra, ‘‘Classification of dental diseases using CNN and transfer learning,’’ in Proc. 5th Int. Symp. Comput. Bus. Intell. (ISCBI), Aug. 2017, pp. 70–74, doi: 10.1109/ISCBI.2017.8053547.

[28] P. Singh and P. Sehgal, ‘‘Automated caries detection based on radon transformation and DCT,’’ in Proc. 8th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT), Jul. 2017, pp. 1–6, doi: 10.1109/ICCCNT.2017.8204030.

[29] A. Haghanifar, M. M. Majdabadi, and S. Ko, ‘‘Automated teeth extraction from dental panoramic X-ray images using genetic algorithm,’’ in Proc. IEEE Int. Symp. Circuits Syst. (ISCAS), Seville, Spain, Oct. 2020, pp. 1–5.

[30] T. Saravanan, M. S. Raj, and K. Gopalakrishnan, ‘‘Identification of early caries in human tooth using histogram and power spectral analysis,’’ Middle-East J. Sci. Res., vol. 20, no. 7, pp. 871–875, 2014.

[31] G. Virupaiah and A. K. Sathyanarayana, ‘‘Analysis of image enhancement techniques for dental caries detection using texture analysis and support vector machine,’’ Int. J. Appl. Sci. Eng., vol. 17, no. 1, pp. 75–86, 2020.

[32] X. J. Zhou, G. X. Yu, Q. Y. Yin, Y. Liu, Z. L. Zhang, and J. Sun, ‘‘Context aware convolutional neural network for children caries diagnosis on dental panoramic radiographs,’’ Comput. Math. Methods Med., vol. 2022, Sep. 2022, Art. no. 6029245.

[33] X. Zhou, G. Yu, Q. Yin, J. Yang, J. Sun, S. Lv, and Q. Shi, ‘‘Tooth type enhanced transformer for children caries diagnosis on dental panoramic radiographs,’’ Diagnostics, vol. 13, no. 4, p. 689, Feb. 2023.

[34] L. Lian, T. Zhu, F. Zhu, and H. Zhu, ‘‘Deep learning for caries detection and classification,’’ Diagnostics, vol. 11, no. 9, p. 1672, Sep. 2021.

[35] S. Saravanan, I. Madivanan, B. Subashini, and J. W. Felix, ‘‘Prevalence pattern of dental caries in the primary dentition among school children,’’ Indian J. Dental Res., vol. 16, no. 4, pp. 140–146, 2005.

[36] K. He, X. Zhang, S. Ren, and J. Sun, ‘‘Deep residual learning for image recognition,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[37] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, ‘‘A survey of convolutional neural networks: Analysis, applications, and prospects,’’ IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 12, pp. 6999–7019, Dec. 2022.

[38] D. R. Sarvamangala and R. V. Kulkarni, ‘‘Convolutional neural networks in medical image understanding: A survey,’’ Evol. Intell., vol. 15, no. 1, pp. 1–22, Mar. 2022.

[39] S. Vinayahalingam, S. Kempers, L. Limon, D. Deibel, T. Maal, M. Hanisch, S. Berg, and T. Xi, ‘‘Classification of caries in third molars on panoramic radiographs using deep learning,’’ Evol. Intell., vol. 11, p. 12609, 2021.

[40] A. Haghanifar, M. M. Majdabadi, and S.-B. Ko, ‘‘PaXNet: Dental caries detection in panoramic X-ray using ensemble transfer learning and capsule classifier,’’ 2020, arXiv:2012.13666.

[41] A. Imak, A. Celebi, K. Siddique, M. Turkoglu, A. Sengur, and I. Salam, ‘‘Dental caries detection using score-based multi-input deep convolutional neural network,’’ IEEE Access, vol. 10, pp. 18320–18329, 2022.

[42] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, ‘‘An image is worth 16×16 words: Transformers for image recognition at scale,’’ 2020, arXiv:2010.11929.

[43] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, ‘‘Swin transformer: Hierarchical vision transformer using shifted windows,’’ in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Montreal, QC, Canada, Oct. 2021, pp. 9992–10002