International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-023-00502-9

REGULAR PAPER

Supremacy of attention-based transformer in oral cancer classification using histopathology images

Bhaswati Singha Deo[1] · Mayukha Pal[2] · Prasanta K. Panigrahi[3] · Asima Pradhan[1,4]

Received: 21 August 2023 / Accepted: 26 December 2023

Abstract

Oral cancer has emerged as one of the ubiquitous malignant tumors globally. Timely detection and treatment reduces the mortality rate of oral cancer. This study utilizes a vision transformer (ViT) framework to classify oral squamous cell carcinoma (OSCC) and healthy oral histopathology images. The proposed approach is implemented on a public database consisting of 4946 oral histopathology images. Although ViT architectures have been extensively used in the medical imaging field, they have not yet been explored in oral cancer detection. Though transformer architecture needs large dataset to attain better performance, our modified architecture accomplishes an accuracy, specificity and sensitivity of 97.78%, 96.72%, and 98.80%, respectively, on a relatively smaller medical dataset. The evaluation metrics of the proposed method have also been compared with eight pre-trained deep learning models, namely Xception, Resnet50, InceptionV3, InceptionResnetV2, Densenet121, Densenet169, Densenet201 and EfficientNetB7. It is observed that the modified ViT model performs better than the deep learning models, demonstrating the ability to extract various features from the histopathology images for the classification. The results of the proposed approach would aid the clinical community for detection of oral cancer in patients of diverse origin.

Keywords Oral cancer · Oral squamous cell carcinoma · Image classification · Deep learning · Histopathology images · Vision Transformers

1 Introduction

Oral cancer is indeed a fatal condition with a complex etiology and a high death rate. The world cancer research fund (WCRF) international claims that malignancies of the oral cavity and lip are one of the most prevalent type of cancers with more than 377,700 cases recorded globally in 2020. The malignancies of the oral cavity and lip are the 11th and 18th most frequently occurring in men and women, respectively. A well-formulated strategy is required for addressing oral cancer which includes early detection, risk factor management, and health literacy. Risk factors include contact with human papillomavirus (HPV), consuming alcohol, smoking, lack of dental hygiene, geographical location, lifestyle, and ethnicity [1].

Squamous cell carcinoma (SCC) may develop from precancerous lesions such as erythroleukoplakia, oral leukoplakia, and verrucous hyperplasia [2]. 90% of all oral cancers are SCCs [3]. The most accurate way to diagnose oral cancer is through biopsy, however, this method is painful, and in cases of extensive or many lesions, selecting the appropriate site and size for surgical treatment of the biopsy sample could be challenging [4]. Additionally, due to lesion variability, the prepared histology specimen may not accurately reflect the identification of the entire lesion. To achieve a successful cure, higher chances of survival, reduced death and morbidity rates, oral squamous cell carinoma (OSCC) must be detected early [5]. The average survival rate stands at 50% for OSCC [6, 7]. The accepted approach for diagnos-

✉ Mayukha Pal
mayukha.pal@in.abb.com

✉ Asima Pradhan
asima@iitk.ac.in

[1] Center for Lasers and Photonics, Indian Institute of Technology Kanpur, Kanpur 208016, India

[2] ABB Ability Innovation Center, Asea Brown Boveri Company, Hyderabad 500084, India

[3] Department of Physical Sciences, Indian Institute of Science Education and Research Kolkata, Mohanpur, Nadia 741246, India

[4] Department of Physics, Indian Institute of Technology Kanpur, Kanpur 208016, India

Published online: 02 February 2024

ing OSCC is tissue sample histopathological examination based on microscopy [8, 9]. However, the clinical value of this approach is constrained by the histopathologists interpretation, which is frequently laborious and prone to error [10]. Therefore, it is crucial to offer efficient diagnostic techniques to support pathologists in the evaluation and diagnosis of OSCC.

Recently, deep learning (DL) algorithms have become the state-of-the-art in field of computer vision and image processing owing to their strength in processing vast volumes of data [11–13]. As a result, numerous investigations have been conducted to aid pathologists through DL techniques specially convolutional neural networks (CNNs) in medical image classification, segmentation and localization [14–16]. Although CNNs excels at feature extraction, they are unable to encode the relative positions of distinct features. Convolution operations fails to recognize global information [17] and long-range relationships across an entire image [18]. Many researchers came up with different architectural changes for an effective solution in due course and eventually [19] proposed attention mechanism that learns the correlation between output and input patterns without relying on repetition. This enables efficient parallelization of Transformer implementations. In response to the popularity of Transformers in natural language processing (NLP) tasks, Transformer architecture was redesigned by [20], referred to as vision transformer (ViT). In the adapted version, the transformer accepts a series of fixed-size image patches as input to extricate complex features of the image. It pays global attention to the entire image overcoming the long-range dependency issue of CNNs. The potential of ViT has been explored by several researchers in diverse computer vision applications say point cloud classification, image enhancement, object detection and many more. In addition to success of ViT in NLP, it has made significant contribution in medical computer vision in a variety of medical imaging modalities.

In the realm of histopathological image classification, ViTs have demonstrated notable success in field of cancer diagnosis, i.e, renal cell carcinoma, breast cancer, cancerous esophagus tissues, glioblastoma, bladder urothelial carcinoma, lower grade glioma, and lung cancer [21, 22]. Despite the widespread utilization of ViT in various disease diagnoses, its potential in the domain of oral cancer has been underexplored. The application of ViTs to oral cancer classification introduces a novel dimension, emphasizing the distinct histopathological characteristics and clinical considerations unique to oral tissues. Oral cancer presents its own set of challenges, marked by specific cellular compositions, anatomical variations, and staining patterns that differentiate it from other cancers. The prevalence of oral cancer, often associated with risk factors like tobacco use, underscores the critical need for accurate diagnostic tools. While ViTs have been leveraged in other cancer types, the adaptation and application of ViTs to oral cancer represent a pioneering effort, addressing a notable gap in the existing literature. By recognizing the unique characteristics of oral cancer and harnessing the power of ViTs, this research contributes to advancing our understanding of oral cancer pathology and heralds a promising avenue for improved clinical outcomes. While Transformers outperform CNNs in interpreting contextual information, their computational demands and the necessity for extensive datasets present challenges in the medical imaging field. The scarcity of publicly accessible imaging datasets for oral cancer further intensifies these difficulties. Considering these constraints, the motivation emerges to employ a fine-tuned ViT for creating an automated diagnostic framework for the detection of oral cancer.

The contributions of the paper are listed as:

1. The performance of the proposed fine-tuned ViT model is either superior or comparable to that of state-of-the-art models in binary-class oral cancer classification across various publicly available oral cancer histopathology datasets.
2. We have performed a comparative analysis of the deep learning (DL) models with the fine-tuned ViT, and it is inferred that ViT model performs better in comparison to DL models for classification of oral cancer.
3. The fine-tuned ViT performs well with a smaller dataset, challenging the assumption that transformer models require large datasets for optimal performance.

The rest part of this manuscript is organized as follows: Sect. 2 discusses prior art of oral cancer classification and ViT in medical domain. Section 3 discusses the methodology utilized in the work. Section 4 presents the results of the proposed methodology and eight pre-trained deep learning models. Section 5 summarizes the work and outlines the future scope of our proposed approach.

2 Related works

Various approaches based on both machine learning and deep learning have been introduced in the literature for the diagnosis of oral cancer through the analysis of medical images. OSCC image databases involve hyperspectral imaging, autofluorescence imaging, computed tomography (CT), magnetic resonance imaging (MRI), and histopathological imaging. Tables 1 and 2 details some of the earlier recommended approaches to oral cancer classification implemented using machine learning and DL neural networks. For machine learning applications on OSCC images, [27] used SVM classifier to attain 91.64% accuracy. For CNN applications on OSCC images, [28] created a DL method that takes patient hyperspectral images into account for advanced

Table 1 Prior art related to publicly available oral cancer image databases

Years	Refs.	Method	Data type	Dataset source	Accuracy (%)
2020	[23]	Deep learning	Histopathology images	Public	91.13
2022	[24]	Customized Alexnet	Histopathology images	Public	90.06
2023	[25]	10-layer CNN	Histopathology images	Public	97.82
2023	[26]	CNN and deep belief network	Clinical photographs	Public	97.35

Table 2 Prior art related to private oral cancer image databases

Years	Refs.	Method	Data type	Dataset source	Accuracy (%)
2012	[27]	Machine learning	Histopathology images	Private	91.64
2019	[28]	Deep learning	Hyperspectral images	Private	94.50
2021	[29]	Deep learning	Clinical photographs	Private	96.20
2022	[15]	Deep learning	Clinical photographs	Private	85.00
2023	[30]	DL based Swin Transfomer	Clinical photographs	Private	98.60

computer-aided oral cancer diagnosis. The performance of the proposed regression-based partitioned DL strategy was assessed against other methods in terms of classifier accuracy, sensitivity, and specificity. [23] used CNN models to attain 91.13%. [29] developed an automated ensemble DL method that combines the benefits of Resnet-50 and VGG-16 to examine oral lesions achieving accuracy of 96.2%. [15] developed a lightweight EfficientNet-B0 DL model for classification of oral lesions images, separating benign from malignant or potentially malignant lesions. [24] explored a tailored AlexNet model designed for the detection of OSCC in histopathological images. [25] introduced a ten-layer CNN model, demonstrating superior performance in diagnosing OSCC from histopathological images compared to pre-trained CNN models. A hybrid optimization algorithm [26] was created combining particle swarm optimization (PSO) with Al-Biruni Earth Radius Optimization. This hybrid approach was employed to optimize the design parameters of Deep Belief Networks and CNNs in the context of identifying malignant oral lesions.

Based on the preceding discussion, it is evident that CNNs have proven to be highly effective in classifying oral cancer, showcasing remarkable accuracy, and establishing their significance in this domain. While CNNs with deep architectures excel at extracting features for numerous small objects within an image, identifying the truly critical regions may pose a challenge. To address this challenge, the utilization of the vision transformer (ViT) model has become prevalent in medical image classification which includes CT scans, X-rays, OCT/Fundus images, MRI Scans, PET, Histopathology images, Endoscopy, and Microscopy. [31] performed a multi-class colorectal cancer tissue classification using ViT and Compact Convolutional Transformer achieving accuracy of 93.3% and 95%, respectively. [32] developed an IL-MCAM framework. It employs interactive learning with attention techniques. [33] carried out a comprehensive analysis and review of the ViT framework for emphysema classification. [34] utilized ViT for Covid-19 detection using CT scans. They employed different ViTB-16, ViTB-32, ViTL-16, ViTL-32, and ViTH-14 for image classification. [35] compared the performance of pneumonia classification using ViT, CNN and VGG16 model. It was demonstrated that ViT achieved highest classification accuracy of 96.45%. [18] put forward an integrated Transformer model for multimodal image classification. The hybrid model comprised of a CNN to learn low-level features, followed by Transformers for global information. [36] classified normal and abnormal fundus images using Tranformer model achieving accuracy of 85.7%. [37] put forth a model that can interpret visual neural activities induced by natural images in form of descriptive text. In [30], a deep-learning methodology utilizing the Swin-Transformer attained a classification accuracy of 0.986 and an AUC of 0.99 in the task of classifying OSCC on clinical photographs. [22] provides an extensive overview of cutting-edge ViTs investigated in histopathological image analysis, covering applications such as segmentation, classification, and survival risk regression.

In our comprehensive review, it is evident that researchers strive to achieve promising diagnostic accuracy through diverse methods. Consequently, we have tailored the ViT framework for enhanced oral cancer detection.

3 Proposed methodology

Figure 3 shows the workflow of the classification methodology. We used the Vision Transformer architecture inspired by [20] to classify oral histopathology images into normal

Table 3 Number of oral histopathology images

Classes	Images
Normal	2435
OSCC	2511

and OSCC and named it as ViT-14. Here, 14 represents the patch size. In this study, we also compared the effectiveness of the proposed approach using 8 pre-trained DL models named Xception [38], Resnet50 [11], InceptionV3 [39], InceptionResnetV2 [40], Densenet121 [41], Densenet169, [41], Densenet201 [41], EfficientNetB7 [13].

3.1 Dataset description

An oral cancer histopathological imaging dataset is publicly available in [42]. It has three directories, namely train, test, and val. We have utilized the train directory [43], as followed by the study [24]. There are two types of subjects in the considered oral histopathology dataset: the patients having oral squamous cell carcinoma and the healthy subjects. Table 3 shows number of images present in the dataset. Figures 1 and 2 shows samples from both dataset categories.

3.2 Preprocessing and data augmentation

All the images in the dataset have been reshaped to 224 × 224 pixel resolution. Data augmentation procedures are employed to increase the image count as training the original dataset will result in overfitting the model. The Keras DL toolbox provides ImageDataGenerator function to generate images with appropriate data augmentation. The resized oral histopathology images undergoes different augmentation techniques such as normalization, randomly rotated, zoomed, horizontally flipped, varying height and width to enhance the generalizability of the model. The details of the augmentation techniques are shown in Table 4.

3.3 Description of the ViT-14 model used in our proposed work

After pre-processing and data augmentation, the images are split into non-overlapping patches inspired by the [19] architecture before being fed to the encoder section. However, non-overlapping style partly breaks the internal framework of an image [44]. Multi-headed self attention (MSA) blocks alleviate this issue by integrating information from several patches. Additionally, when non-overlapping patches are fed into Transformer, computational redundancy does not exist. In our study, an input image of size 224 × 224 × 3 (H=224, W=224, C=3) is splitted into flattened patches of size 588 (P^2C, where P=14,C=3). Thus, 256 patches (N=HW/P^2) are generated before heading into the Transformer encoder section. It is also noted that the sequence length of the Transformer and square of patch size are inversely related, hence, models having smaller patch size requires more computation.

The resulting flattened patches are used to create linear embeddings of a lower dimensional latent space (D) of size 64, known as patch embeddings. The size of the latent space remains constant through all layers of the encoder. ViT does not use convolution or recurrence in the multi-head self attention module in encoder section, hence to ensure that the images preserve their positional knowledge, position embeddings are then linearly added to the patch embeddings using Eq. 1.

$$y_j = y_j + x_j \tag{1}$$

where, y_j denotes the patch embedding of the jth patch and x_j denotes the position embedding of the jth patch and y_j, x_j

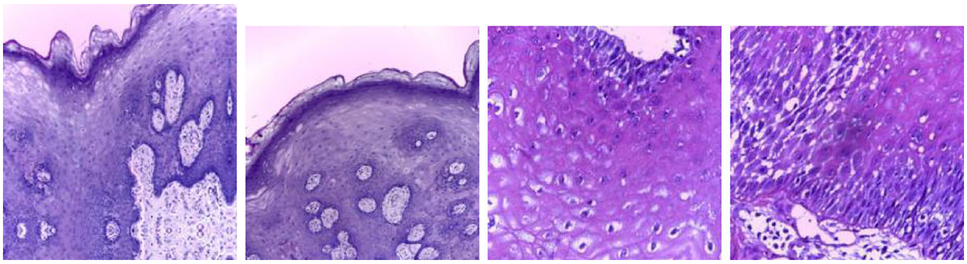


Fig. 1 Few histopathology images from Normal category

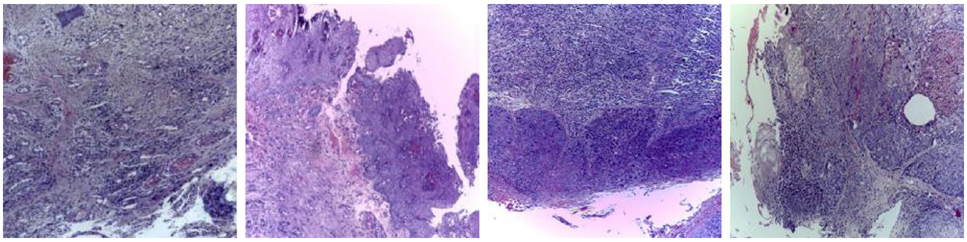


Fig. 2 Few histopathology images from OSCC category

Table 4 Details of data augmentation techniques

Techniques	Values	
	DL models	ViT-14 model
Random rotation	0.02	0.02
Random zoom	–	0.2–0.2
Random flip	Horizontally	Horizontally
Random height	0.2	–
Random width	0.2	–

$\in D_y$. D_y is the dimensionality of the jth patch embedding. An additional learnable (class) embedding is also added similar to BERT's (class) token as shown in Eq. 2. The class of the input image is predicted using this class embedding.

$$P_{\mathrm{o}} = [y_{\mathrm{class}}; y_1; y_2; ...y_N] + [x_1; x_2; ...x_N; x_{N+1}] \qquad (2)$$

where, y_{class} is the additional learnable class embedding and $P_{\mathrm{o}}^{\mathrm{o}}$ is y_{class}. The outcome of the Transformer encoder (TE) at the Lth layer ($L = 8$) is denoted as P_L^{o}. The series of patches are then passed to the TE layer. The TE module is composed of alternating layers of multi-head self attention (MSA) layer and feed forward network (FFN). The patch embedding passes through a number of layers in encoder section depicted by Eqs. 3 and 4.

$$P_l^{'} = MSA(LN(P_{l-1})) + P_{l-1} \qquad l = 1, 2, ..L \qquad (3)$$

$$P_L = FFN(LN(P_l^{'})) + P_l^{'} \qquad l = 1, 2, ..L \qquad (4)$$

where, LN is the layer normalization layer. The output of the encoder at the Lth layer, P_L^{o} is layer normalized and passed through a learnable classification network known as multi-layer perceptron (MLP) head as shown in Fig. 6.

In basic terms, a group of patches splitted from an input image is transformed into a latent vector with a specified size. Then position embeddings are added to the transformed patch embeddings and a class token is also prepended. Further, the modified input passes through a chain of encoder layers. The pictorial representation of the ViT-14 model is shown in Fig. 4.

In TE block, the embedded patches pass through MSA layer and feed forward network. A residual connection [11] is added prior to layer normalization [45] layer around MSA layer and FFN. The TE module is shown in Fig. 5. Muti-headed attention boosts the performance of the model by performing multiple self-attention mechanism simultaneously. Each self attention operation serves as a head in multi-headed attention mechanism, and each head tries to learn something unique, thus improving the representation power of the encoder module. Therefore, the model is able to capture intricate correlations of different patches present at distinct locations in a histopathology image. It focuses on local and global features encompassed within an image in contrast to conventional CNN models which emphasizes on local attention. The parameters of the adopted ViT-14 model is tabulated in Table 5. Details of the layers, output shape and number of parameters are shown in Table 6.

Table 5 Specifications of the ViT-14 model

Model	Latent dimension	Layers	Heads	Parameters
ViT-14	64	8	4	3.6M

3.4 Pre-trained deep learning models for comparison

This subsection gives a brief description of the various DL models used in our work for comparative analysis. Architecture of the DL models used in our study is shown in Fig. 7. Table 7 lists few details of the DL models.

3.4.1 Xception

The elementary theory of Inception has been pushed to an extreme in Xception architecture [38]. In Inception, 1x1 convolutions were used to extract features from the initial input, and filters of varying sizes were employed at every depth space. The reverse occurs in Xception, it uses filters at every depth space independently prior to compressing the input image at once using 1x1 convolution. The feature extraction backbone in the Xception architecture is composed of 36 convolutional layers. The Xception architecture can be summed up as a linear stack of residually connected depthwise separable convolution layers. As a result, developing and altering the architecture is relatively simple.

3.4.2 Resnet50

The Resnet50 utilizes a bottleneck framework for its building block. The residual block consists of 1×1 convolutions termed as bottleneck, which minimizes the matrix multiplications and parameter count. This makes training each layer considerably faster. Instead of using a stack of two layers, it leverages three layers [11]. It is widely known that increasing the depth of the model for deeper feature extraction reduces model performance due to exploding or vanishing gradient issue. To resolve this issue and enable the training of deeper networks, residual blocks were introduced.

3.4.3 InceptionV3

InceptionV3 is an image recognition model that achieved an accuracy higher than 77.9% on the ImageNet dataset. It is an optimized and upgraded adaptation of InceptionV1 model.

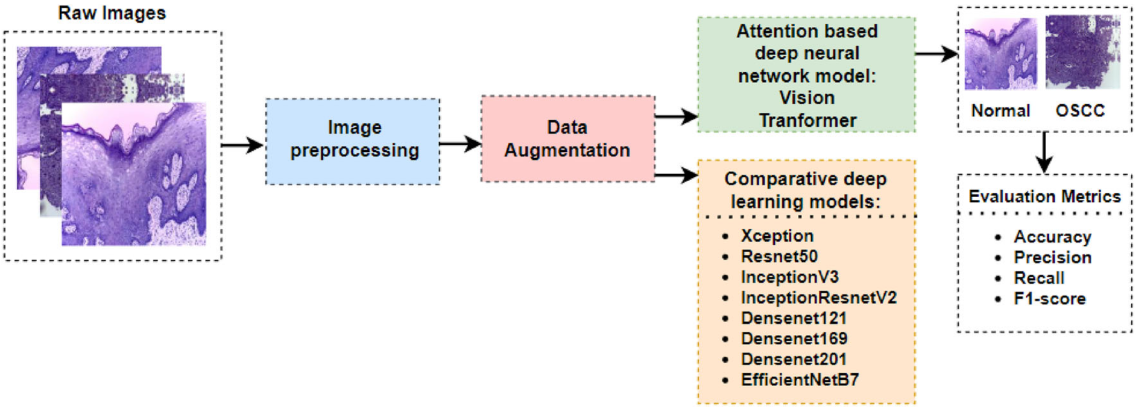


Fig. 3 Proposed classification methodology utilizing Vision Transformer and DL models

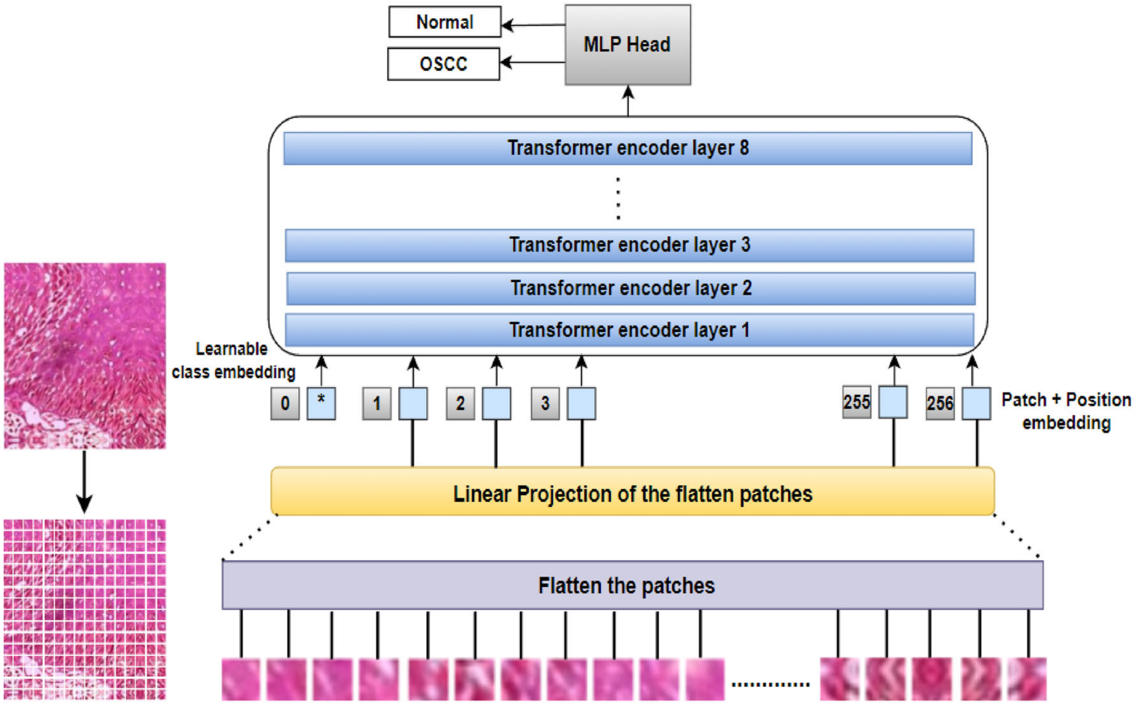

Normal
OSCC
MLP Head
Transformer encoder layer 8
Transformer encoder layer 3
Transformer encoder layer 2
Transformer encoder layer 1
Learnable class embedding
0
*
1
2
3
255
256
Patch + Position embedding
Linear Projection of the flatten patches
Flatten the patches

Fig. 4 Architecture of ViT-14 model for classifying normal and OSCC histopathology images

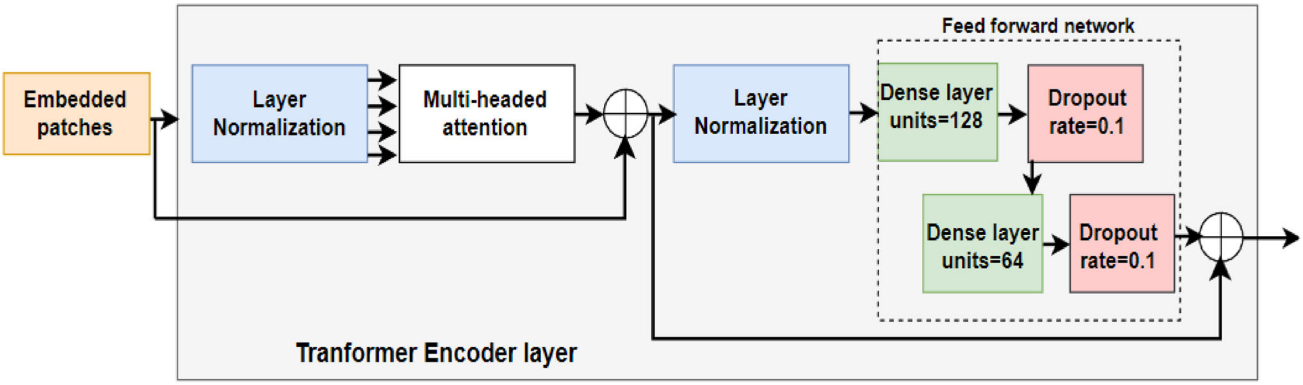


Fig. 5 Transformer encoder for processing oral cancer histopathology image patches with multi-head attention layer

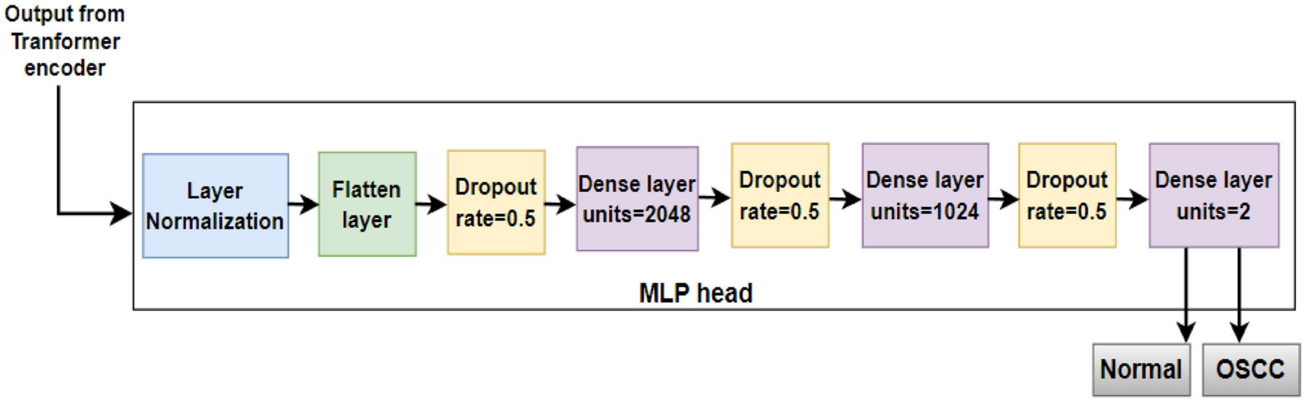


Fig. 6 MLP head for classifying oral cancer histopathology images

Table 6 Details of the names of the layers, the output shape, and the number of parameters used in each layer of the ViT-14 model(*a* to *b* layers forms the encoder section and are repeated 8 times)

Layer name	Output shape	No. of parameters
InputLayer	(None,224,224,3)	0
data_augmentation	(None,224,224,3)	7
Patches	(None,None,588)	0
PatchEncoder	(None,256,64)	54080
LayerNormalization[a]	(None,256,64)	128
MultiHeadAttention	(None,256,64)	66368
Add	(None,256,64)	0
LayerNormalization	(None,256,64)	128
Dense	(None,256,128)	8320
Dropout	(None,256,128)	0
Dense	(None,256,64)	8256
Dropout	(None,256,64)	0
Add[b]	(None,256,64)	0
LayerNormalization	(None,256,64)	128
Flatten	(None, 16384)	0
Dropout	(None, 16384)	0
Dense	(None, 2048)	33556480
Dropout	(None, 2048)	0
Dense	(None, 1024)	2098176
Dropout	(None, 1024)	0
Dense	(None, 2)	2050
Trainable params	36,376,514	
Non-trainable params	7	

Factorized convolutions, smaller convolutions, asymmetric convolutions, auxillary classifiers, and grid size reduction forms the architecture of InceptionV3 [39].

3.4.4 InceptionResnetV2

A convolutional neural network known as InceptionResNetV2 expands on the Inception group of architectures while incorporating residual connections. It replaces the filter concatenation step of the Inception model [40].

3.4.5 Densenet121/169/201

Densenets are deep CNNs that enhance the training of deeper networks by connecting the feature map of one layer with all the layers preceding it [41]. This increase the effectiveness with regard to memory utilization and computation. It can extract minute features of the input images with few channels. DenseNet further improves feature propagation, increases feature reuse, and significantly lowers the number

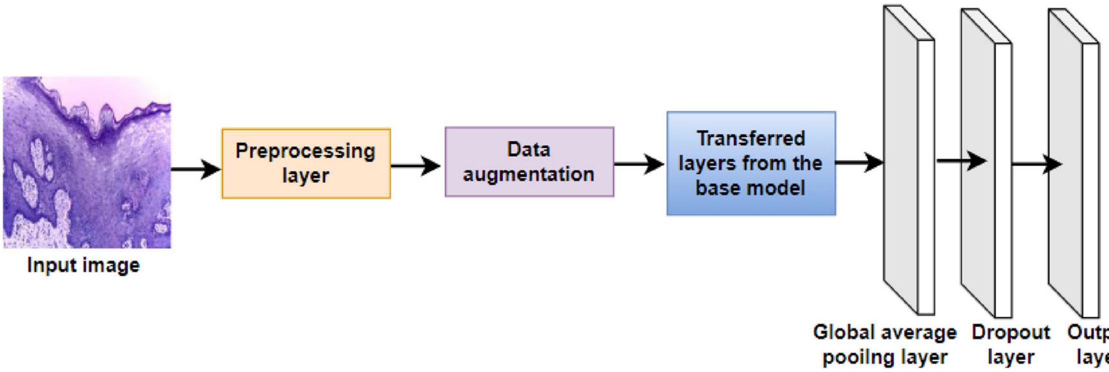


Fig. 7 Architecture of the deep neural network models for comparative analysis with ViT-14 model using oral cancer histopathology images

of parameters, obviate the vanishing gradient issue and mitigates its impacts.

3.4.6 EfficientNetB7

EfficientNetB7 is a non-repetitive, nonlinear neural network search that optimizes floating point operations per second (FLOPS) and accuracy by balancing resolution, network depth and breadth. Seven flipped residual blocks, each with its own parameters, are used in the architecture. These blocks employ swish activation, squeeze, and excitation blocks [13].

4 Experiments and analysis

All studies related to this work were carried out using Python 3.7.6, TensorFlow 2.7.0, and Keras 2.7.0 on a PC with 2.40 GHz Intel(R) Core(TM) i5-1135G7 processor, Intel(R) Iris(R) Xe graphics and 16.0 GB of RAM.

4.1 Evaluation indicators

The overall performance of our proposed approach is evaluated on the basis of the contents of the confusion matrix. There are four terms namely true positive (TP), false positive (FP), false negative (FN), and true negative (TN) included in this evaluation matrix. TP means a person has OSCC and the model accurately predicts it. TN means a person has healthy oral mucosa and the model accurately predicts it. FP means a healthy oral mucosa is inaccurately predicted as OSCC. FN means an OSCC is inaccurately predicted as healthy oral mucosa. Evaluation indicators, namely specificity, sensitivity, F1-score, precision, cohen kappa score (CKS), matthews correlation coefficient (MCC), error rate, false omission rate (FOR), false discovery rate (FDR), negative predictive value (NPV), false negative rate (FNR), and false positive rate (FPR) were evaluated to assess the performance of our proposed approach. These evaluation indicators can be calculated utilizing the formulas given as follows:

1. Precision: It represents the proportion of accurately predicted positive instances out of the total instances predicted as positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

2. Sensitivity: It denotes the proportion of accurately predicted positive instances relative to all instances in the actual positive class.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

3. Specificity: It measures the ability of the model to correctly identify true negatives.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{7}$$

4. Accuracy: It measures the ratio of accurately identified images to the total number of test images.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{8}$$

5. F1-Score: It is the harmonic mean of precision and recall, serves as a means to optimize the model either for recall or precision.

$$F1 \text{ Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

6. Cohen Kappa score: It is a metric used to measure the agreement between predicted and actual classifications while accounting for the possibility of random agreement.

$$\text{CKS} = \frac{P_\text{o} - P_\text{e}}{1 - P_\text{e}} \tag{10}$$

Table 7 A brief introduction to the DL models used along with the ViT-14 for a comparative study in oral histopathology image classification

Model	Refs.	Number of parameters	Major remarks
Xception	[38]	22.9M	(a) It is InceptionV3's improved version. (b) It makes use of separable convolutions in depth
Resnet50	[11]	25.6M	(a) There is a new block called residue (b) The issue of accuracy rapidly declining as a result of an increase in network layers has been resolved
InceptionV3	[39]	23.9M	(a) It makes use of a better Inception Module (b) Small convolutions are used to introduce the concept of factorization
InceptionResnetV2	[40]	55.9M	(a) The Resnet connection and the Inception block are combined
Densenet121	[41]	8.1M	(a) When two layers have the same feature map size,
Densenet169	[41]	14.3M	a direct connection between them is established
Densenet201	[41]	20.2M	(b) Hundreds of layers are possible at this scale
EfficientNetB7	[13]	66.7M	(a) Reduces parameter size and increases accuracy

where P_o is the observed agreement and P_e is the expected agreement.

7. MCC: It is a correlation coefficient between the observed and predicted binary classifications.

$$\text{MCC} = \frac{(\text{TP} * \text{TN} - \text{FP} * \text{FN})}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{11}$$

8. Error rate: It provides a measure of misclassification.

$$\text{Error}\quad\text{rate} = 1 - \text{Accuracy} \tag{12}$$

9. False omission rate (FOR): It is the proportion of false negatives out of the total actual negative instances.

$$\text{FOR} = \frac{\text{FN}}{\text{FN}+\text{TN}} \tag{13}$$

10. False discovery rate (FDR): It is the proportion of false positives out of the total predicted positive instances.

$$\text{FDR} = \frac{\text{FP}}{\text{FP}+\text{TP}} \tag{14}$$

11. Negative predictive value (NPV): It is the proportion of correctly predicted negative instances out of the total predicted negative instances.

$$\text{NPV} = \frac{\text{TN}}{\text{TN}+\text{FN}} \tag{15}$$

12. False negative rate (FNR): It is the proportion of false negatives out of the total actual positive instances.

$$\text{FNR} = \frac{\text{FN}}{\text{FN}+\text{TP}} \tag{16}$$

13. False positive rate (FPR): It is the proportion of false positives out of the total actual negative instances.

$$\text{FPR} = \frac{\text{FP}}{\text{FP}+\text{TN}} \tag{17}$$

4.2 Model parameters

We selected sparse categorical crossentropy as a loss function for our binary classification task. The training is done over 100 epochs with the AdamW optimizer. We have used a patch size of 14 × 14 × 3 with each image having 256 patches. In our Transformer encoder architecture, we employed a configuration with 4 heads and opted for 8 layers in the Transformer encoder. Batch size of 32, learning rate of 0.001 and weight decay of 0.0001 are chosen for model training. Table 8 lists the optimal hyperparameters used in our study.

4.3 Ablation study on model parameters

We perform an ablation study to analyze how different components and hyperparameters in our proposed model contribute to the overall performance of the model.

4.3.1 Impacts of different parameters

In the initial experimentation phase with the ViT-14 model, default hyperparameters were initially assumed. This included a learning rate of 0.001, a batch size of 8, weight decay of 0.0001, a patch size of 14 × 14 × 3, a latent dimension of 64, the number of Transformer encoder layers set to 6, and the number of heads set to 4. Subsequent exploration involved varying batch sizes to 8 and 16 while keeping other parameters constant, and it was found that a batch size of 32 yielded

Table 8 Hyperparameters used in the ViT-14 model

Hyperparameters	ViT-14 model
Optimizer	AdamW
Learning Rate	0.001
Weight decay	0.0001
Loss Function	SparseCategorical Crossentropy
Batch Size	32
Epochs	100
Input size	224 × 224 × 3
Patch size	14 × 14 × 3
Patches per image	256
Latent dimension	64
Transformer encoder layers	8
Number of heads	4

Table 9 Accuracy of the ViT-14 model using different hyperparameters

Parameters	Values	Accuracy (%)
Batch size	8	95.15
	16	95.96
	32	**96.97**
Patch size	12 × 12	96.36
	14 × 14	**96.97**
	16 × 16	96.57
Latent dimension	16	95.96
	32	96.36
	64	**96.97**
Layers	6	96.97
	8	**97.78**
	10	96.97
Heads	**4**	**97.78**
	6	97.37
	8	95.96

Values in bold indicate the best evaluation metrics

the highest accuracy, as outlined in Table 9. Further experiments focused on patch size variations while maintaining other parameters constant, confirming that the initial dimensions of 14 × 14 × 3 for the patch achieved highest accuracy. Likewise, alternative latent dimensions of 16 and 32 were explored while keeping other parameters constant, with the initial choice of 64 demonstrating the highest accuracy, as illustrated in Table 9. Once optimized values for batch size, patch size, and latent dimension were obtained, experiments on the number of layers indicated that 8 layers outperformed 6 and 10, as detailed in Table 9. Finally, experiments on the number of heads, exploring values of 6 and 8, validated the initial choice of 4 as yielding the highest accuracy, as indicated in Table 9.

Table 10 Number of training, validation and testing images in ViT-14 model

Dataset	Percentage	No. of images
Training	90%	4451
Validation	10% of the training set	445
Test	10%	495

Table 11 Evaluation metrics of the ViT-14 model

Evaluation metrics	Obtained value
Precision	96.88
Sensitivity	98.80
Specificity	96.72
Accuracy	97.78
F1-score	97.83
CKS	0.96
MCC	0.96
Error rate	2.22
FOR	0.01
FDR	0.03
NPV	0.99
FNR	0.01
FPR	0.03

4.4 Results

After obtaining the optimal model hyperparameters, we have evaluated the performance of the model. The dataset is divided into two subsets: the training set, comprising 90% of the data, with 10% of this subset allocated for validation; and the testing set, which constitutes the remaining 10% as shown in Table 10.

It is extremely important that the model does not exhibit significant overfitting to ensure the overall effectiveness of the proposed method. Figure 8a shows the training and validation accuracy and loss curves plotted over 100 epochs. It is observed that the model exhibits no major overfitting, and robustness is maintained. The confusion matrices (CM) generated have been displayed in Fig. 8b which further helps in understanding the results. It is inferred from CM that FP and FN are very less in number as compared to TN and TP, thus showing correct predictions of images into normal and OSCC classes. Table 11 shows the evaluation metrics of the proposed model. As shown in Table 11, ViT-14 model achieved an accuracy, specificity, and sensitivity of 97.78%, 96.72%, and 98.80%, respectively.

In our study, we implemented a fivefold cross-validation methodology, running the model five times to ensure a comprehensive and robust evaluation of its generalization to unseen data. For each iteration, the dataset was shuffled to

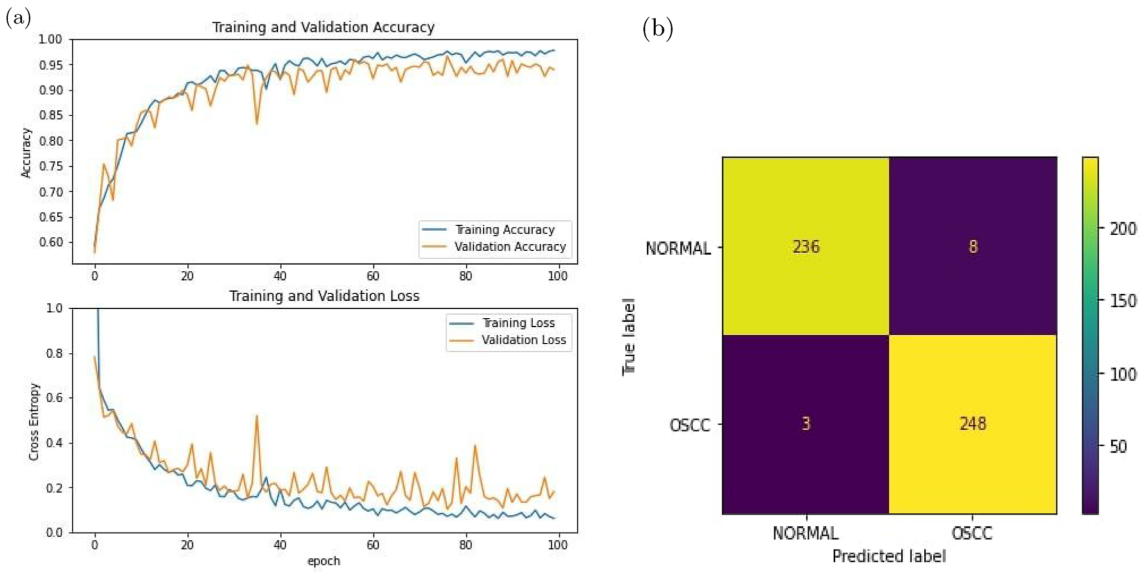


Fig. 8 **a** Training and validation accuracy vs epoch plot of ViT-14 model over 100 epochs (top); Loss vs epoch plot of ViT-14 model over 100 epochs (bottom) **b** Confusion matrix of the ViT-14 model

Table 12 Evaluation metrics for the ViT-14 model using fivefold cross-validation technique

Folds	Accuracy (%)	Sensitivity (%)	Specificity (%)
Fold 1	95.15	93.82	96.82
Fold 2	95.15	97.82	91.82
Fold 3	96.97	97.09	96.82
Fold 4	94.34	96.36	91.82
Fold 5	96.16	97.09	95.00
Average	95.56	96.44	94.46

create unique training and test sets, with the training set comprising 90% of the data and the test set representing the remaining 10%. Evaluation metrics were computed in each iteration on the assigned test set, offering a thorough assessment of performance of the model across diverse data splits. The CM generated from each of the five folds have been displayed in Fig. 9 which further helps in understanding the results. Evaluation metrics for ViT-14 model using fivefold cross validation are listed in Tables 12 and 13.

4.4.1 Impacts of split ratio

The model underwent evaluation with different combinations of training and testing ratios. This evaluation retained consistency with the same set of hyperparameters, as detailed in Table 8. This approach allowed for an assessment of its performance under different data split scenarios while keeping the experimental conditions uniform.

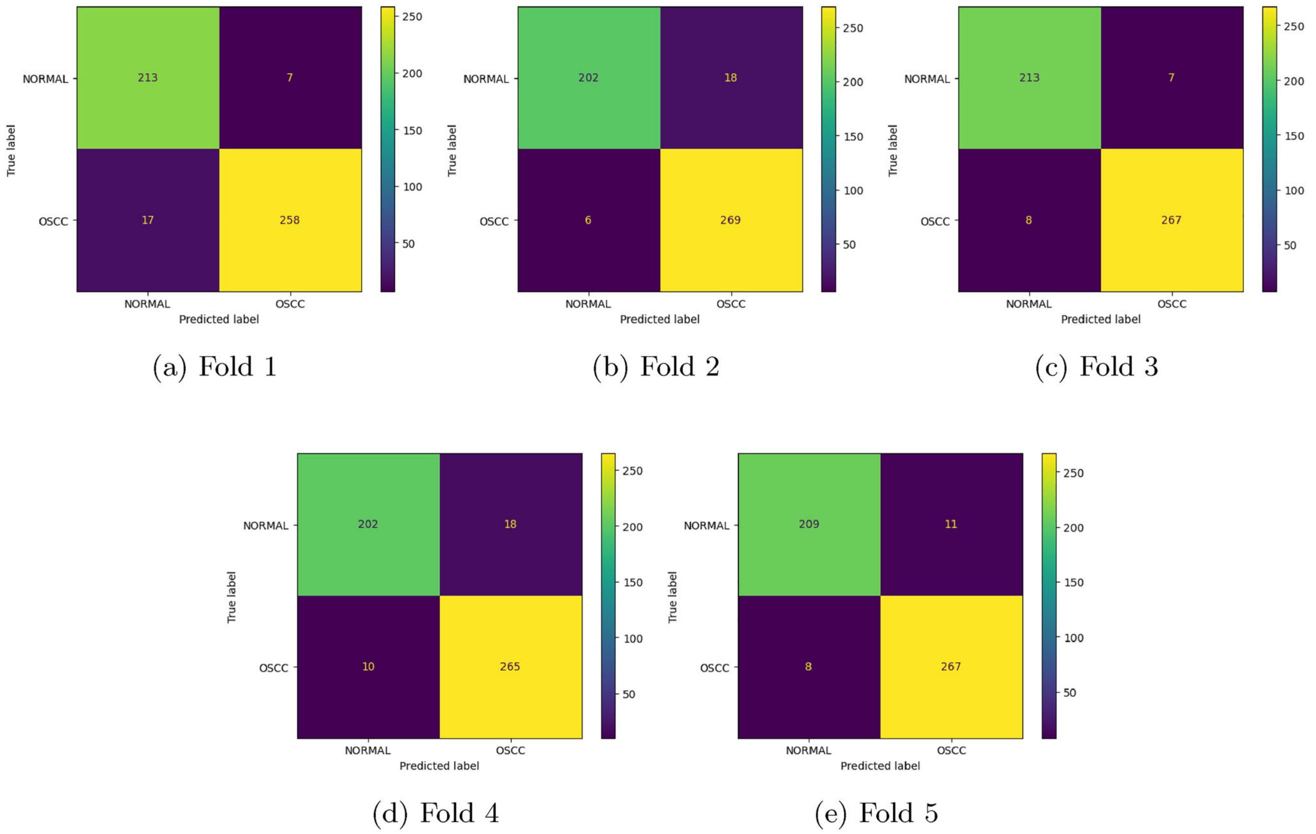
Table 13 Classification report for the ViT-14 model using fivefold cross-validation technique

Folds	Class	Precision	Recall	F1-score
Fold 1	Normal	0.93	0.97	0.95
	OSCC	0.97	0.94	0.96
Fold 2	Normal	0.97	0.92	0.94
	OSCC	0.94	0.98	0.96
Fold 3	Normal	0.96	0.97	0.97
	OSCC	0.97	0.97	0.97
Fold 4	Normal	0.95	0.92	0.94
	OSCC	0.94	0.96	0.95
Fold 5	Normal	0.96	0.95	0.96
	OSCC	0.96	0.97	0.97

Case 1 (90:10) Training images constitute 90%, and testing images make up 10% of the entire dataset. For the ViT-14 model, 4451 images were used for training, and 495 for testing. The model underwent five runs to assess generalizability, and the resulting average accuracy values are detailed in Table 14.

Case 2 (80:20) Training images constitute 80%, and testing images make up 20% of the entire dataset. For the ViT-14 model, 3956 images were used for training, and 990 for testing. The model underwent five runs to assess generalizability, and the resulting average accuracy values are detailed in Table 14.

Case 3 (70:30) Training images account for 70%, with testing images at 30% of the overall dataset. The ViT-14 model was trained on 3462 images and tested on 1484. Sim-



(a) Fold 1

(b) Fold 2

(c) Fold 3

(d) Fold 4

(e) Fold 5

Fig. 9 The confusion matrices of the five folds of the cross-validation technique for the ViT-14 proposed approach

Table 14 Average accuracy of ViT-14 model using fivefold cross validation technique for different split ratios (train:test)

	90:10	80:20	70:30
Accuracy (%)	**95.56**	95.51	94.99

Values in bold indicate the best evaluation metrics

ilar to Case 1, the model underwent five runs to evaluate generalizability, and the corresponding average accuracy values are presented in Table 14.

It is observed from Table 14 that the accuracy appears to decrease as the proportion of training data decreases relative to testing data. Thus, a higher proportion of training data (90:10 ratio) contributes to better model performance.

4.5 Comparative analysis of model performance across different datasets

After the ablation study, the optimal model configuration is used for further analysis on two publicly available oral cancer histopathological datasets.

Dataset 1 [46] was collected from a histopathological image repository of the normal epithelium of the oral cavity and OSCC images. The repository consists of 1224 total images. They are divided into two sets in two different resolutions, 100x magnification and 400x magnification. In total, there are 290 normal epithelium images and 934 OSCC images.

Dataset 2 is an oral cancer histopathological image dataset available in [42]. It comprises three directories: train, test, and val, containing a total of 5192 images. There are 2,494 normal images and 2,698 images with OSCC.

We employed a fivefold cross-validation approach, executing the model five times to ensure a thorough and robust evaluation of its ability to generalize to new data. In each iteration, the dataset was shuffled, creating distinct training and test sets. The training set constituted 90% of the data, while the test set comprised the remaining 10%. Evaluation metrics were calculated in each iteration on the assigned test set, providing a comprehensive assessment of the performance of the model across a variety of data partitions. Tables 15 and 16 present the evaluation metrics for dataset 1 and dataset 2, respectively, respectively, utilizing the fivefold cross-validation technique.

4.6 Comparison with deep learning models

The proposed approach is compared with eight pre-trained DL models to demonstrate its effectiveness. The details of

Table 15 Evaluation metrics of ViT-14 model using fivefold cross validation technique for Dataset 1 (Acc:Accuracy, Prec:Precision, Sens:Sensitivity, F1:F1-score, Spec:Specificity)

Folds	Acc. (%)	Prec. (%)	Sens. (%)	F1 (%)	Spec. (%)
Fold 1	91.87	95.35	93.18	94.25	88.57
Fold 2	94.31	94.12	98.97	96.48	76.92
Fold 3	93.49	96.08	96.08	96.08	80.95
Fold 4	**95.12**	**95.05**	**98.97**	**96.09**	**80.76**
Fold 5	91.05	92.78	95.74	94.24	70.96
Average	93.17	94.68	96.59	95.43	79.63

Values in bold indicate the best evaluation metrics

Table 16 Evaluation metrics of ViT-14 model using fivefold cross-validation technique for Dataset 2 (Acc:Accuracy, Prec:Precision, Sens:Sensitivity, F1:F1-score, Spec:Specificity)

Folds	Acc. (%)	Prec. (%)	Sens. (%)	F1. (%)	Spec. (%)
Fold 1	**97.69**	**98.21**	**97.51**	**97.86**	**97.91**
Fold 2	97.69	97.74	97.74	97.74	97.64
Fold 3	97.12	96.22	98.08	97.14	96.15
Fold 4	96.35	95.24	98.25	96.73	94.04
Fold 5	96.92	95.86	98.08	96.96	95.77
Average	97.15	96.65	97.33	97.29	96.30

Values in bold indicate the best evaluation metrics

the hyperparameters are listed in Table 18. The dataset is divided into two subsets: the training set, comprising 90% of the data, with 10% of this subset allocated for validation; and the testing set, which constitutes the remaining 10% as shown in Table 17. The training-to-testing split ratio of 9:1 was maintained, consistent with the proposed ViT-14 method. We selected binary cross-entropy as a loss function for our binary classification task. Adam optimizer is used for training over a course of 100 epochs. The reduction of the generalization gap between training loss and validation loss was our main objective during model training. A batch size of 32 and a learning rate of 0.001 is used. Additionally, a dropout rate of 0.2 is used to address overfitting during training time [29]. Then, we saved the weights of the model having lowest validation loss for evaluation purposes. We adhered to the original architectural descriptions of convolutional filters, padding, pooling and strides in Xception, Resnet50, InceptionV3, InceptionResnetV2, Densenet121/169/201, EfficientNetB7 models.

We have utilized a model pre-trained on the ImageNet dataset for the analytical results of DL models on the oral histopathology dataset. The training and validation accuracy and loss curves are plotted over 100 epochs as displayed in Figs. 10 and 11. The confusion matrix (CM) for DL models were also calculated to help in understanding the results as shown in Fig. 12. It is inferred from the CM that FP and FN have increased in comparison to the ViT-14 model, thus showing lesser correct predictions of images into normal and OSCC classes. Table 19 lists the various evaluation measures of the compared DL models and ViT-14 model. Table 20 shows the superior performance of ViT-14 model in terms of accuracy, specificity and sensitivity in comparison to the DL models.

Table 17 Number of training, validation and testing images in DL models

Dataset	Percentage	No. of images
Training	90%	4450
Validation	10% of the training set	445
Test	10%	496

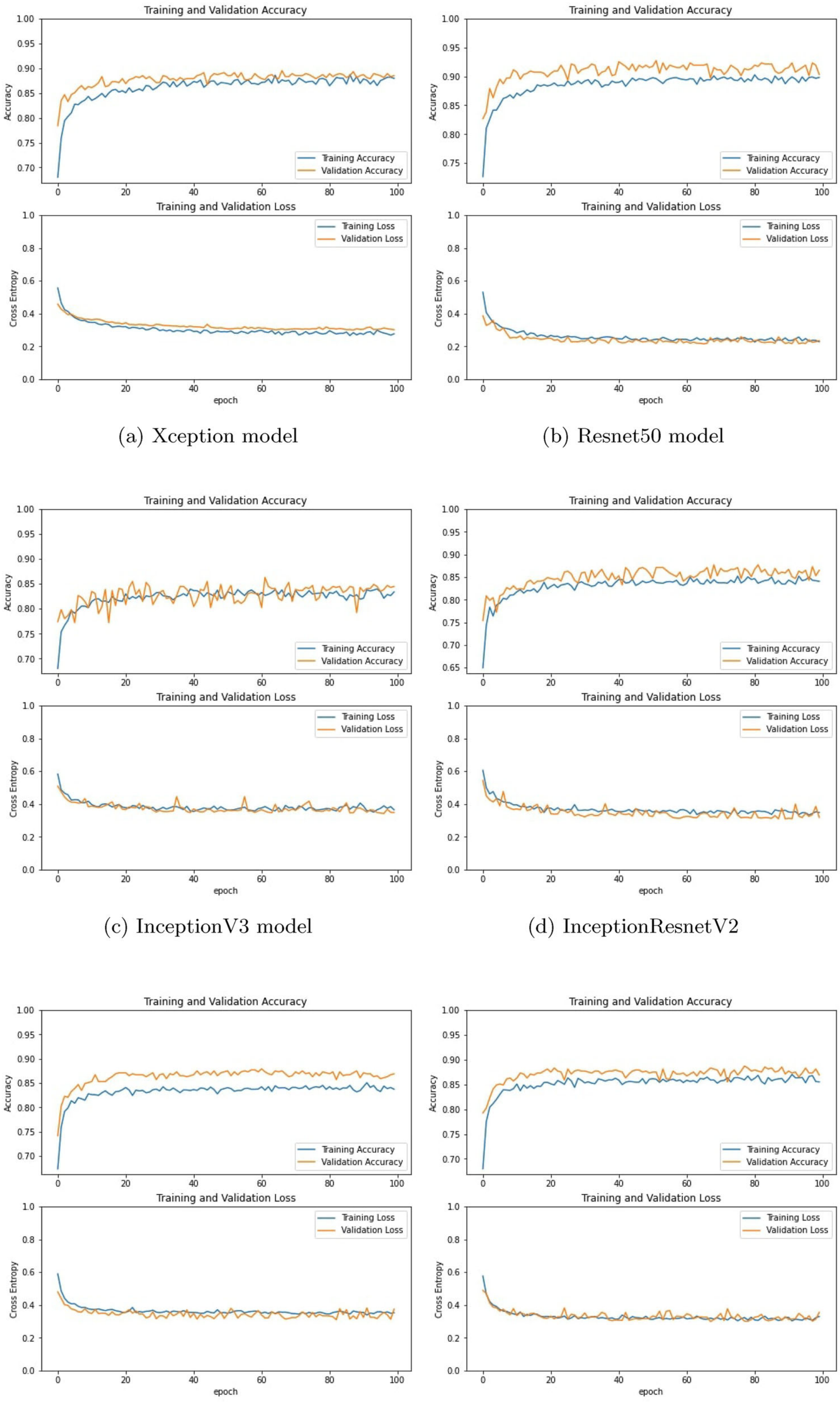
Table 18 Hyperparameters used in the considered deep learning models

Hyperparameters	DL models
Optimizer	Adam
Learning Rate	0.001
Loss Function	Binary Cross-entropy
Batch Size	32
Epochs	100
Dropout	0.2
Input size	224 × 224 × 3

4.7 Comparison with previous works

Table 21 provides a comprehensive comparative analysis of diverse methods and models applied to various publicly available oral cancer datasets. In previous research [23], transfer learning methods using Resnet50, MobileNet, and InceptionV3 achieved accuracies ranging from 76.61% to 91.13%

Fig. 10 The convergence behavior of the Deep learning models used for comparative analysis



(a) Xception model

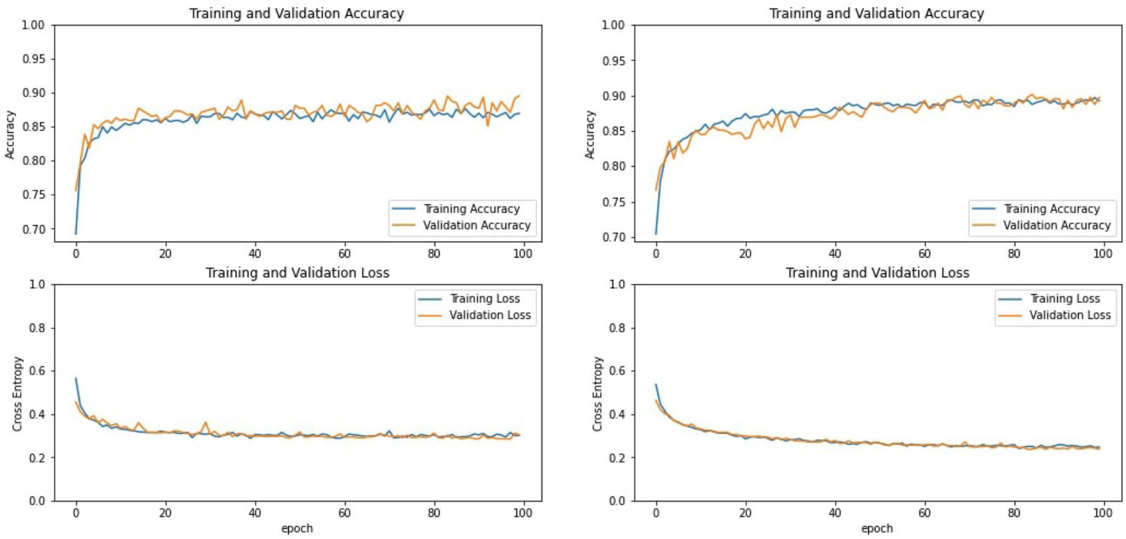
(b) Resnet50 model

(c) InceptionV3 model

(d) InceptionResnetV2

(e) Densenet121 model

(f) Densenet169 model



(a) Densenet201 model

(b) EfficientNetB7 model

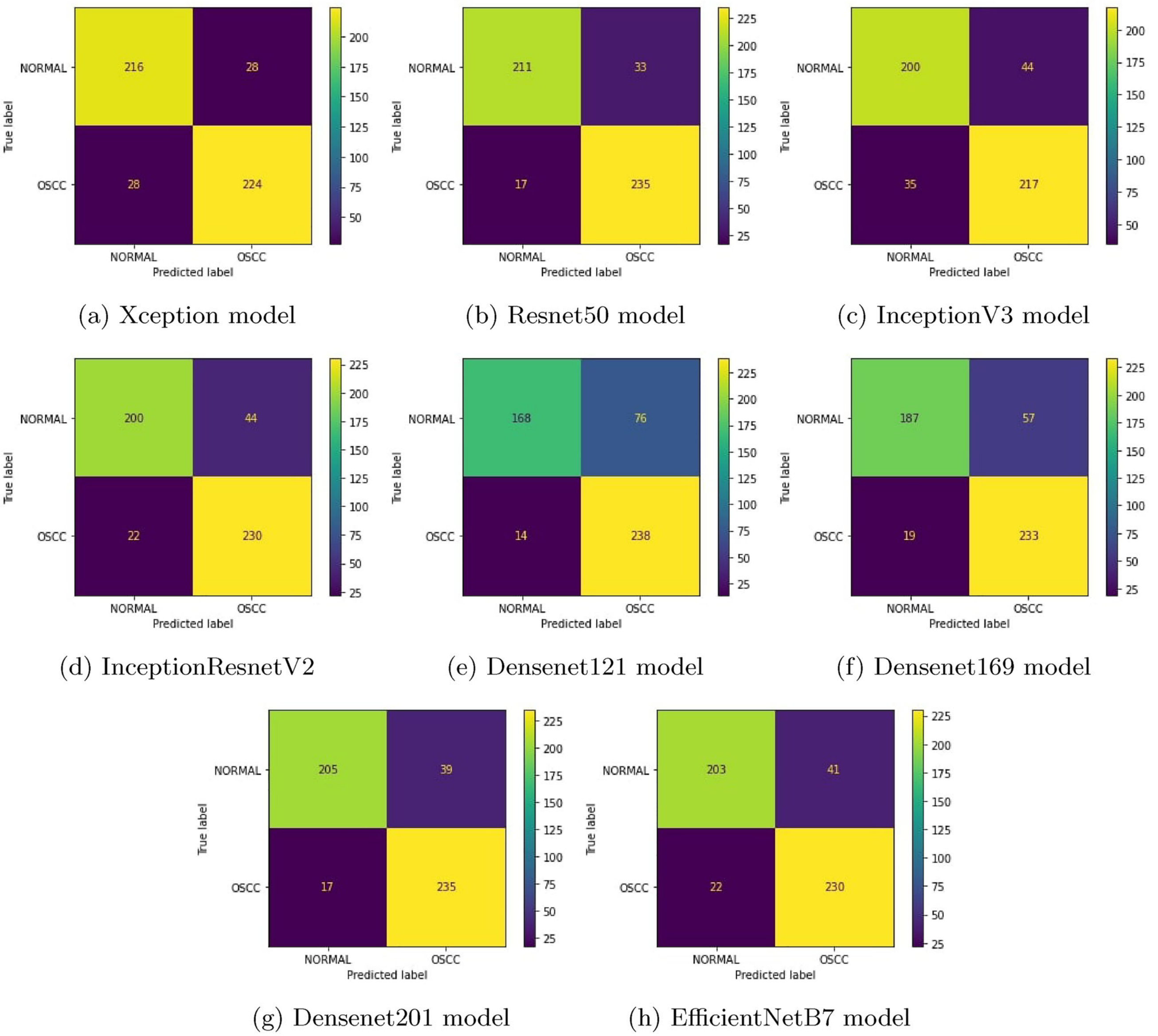
Fig. 11 The convergence behavior of the DL models used for comparative study

on a dataset containing 290 normal and 934 OSCC images [46]. A customized 10-layer CNN [25] attained a higher accuracy of 97.82% on the same dataset [46]. A hybrid approach involving both CNNs and SVM, and the integration of deep and texture-based features, the study [47] demonstrated an accuracy of 97.00% on 2698 OSCC images and 2494 healthy tissue images [42]. Additionally, Gabor filter combined with a Catboost classifier [48] achieved 94.92% accuracy on the same dataset [42]. A transformer with external attention [49] attained an accuracy of 96.97% on 2511 OSCC images and 2435 healthy tissue images [43]. Transfer learning using Alexnet [24] achieved 90.06% accuracy on same set [43]. While the proposed method demonstrated an accuracy of 95.12% on the dataset [46], it is noteworthy that the 10-layer CNN model [25] achieved a higher accuracy of 97.82%. However, it is important to highlight that the proposed method showcased competitive performance on other datasets, achieving accuracies of 97.69% and 97.78% on datasets [42] and [43] respectively. It is evident that a reduction in performance on the dataset [46] is significantly due to class imbalance, potentially impacting the model's ability to effectively learn and generalize across both classes. While the presently employed data augmentation techniques, such as rotation, zoom, flip, height, and width variations, contribute to model resilience, addressing the class imbalance may require additional augmentation strategies. This could involve applying techniques like synthetic minority over-sampling technique (SMOTE) and employing generative adversial networks (GANs) for the creation of realistic synthetic samples, particularly for the minority class. By implementing such additional data augmentation techniques tailored to address class imbalances, the proposed model is likely to achieve improved generalization and classification accuracy across all datasets, ensuring consistent performance in the presence of varied class distributions.

Table 19 Oral cancer image classification summary report without cross-validation technique

Models	Class	Prec.	Rec.	F1
Xception	Normal	0.89	0.89	0.89
	OSCC	0.89	0.89	0.89
Resnet50	Normal	0.93	0.86	0.89
	OSCC	0.88	0.93	0.90
InceptionV3	Normal	0.85	0.82	0.84
	OSCC	0.83	0.86	0.85
Inception ResnetV2	Normal	0.90	0.82	0.86
	OSCC	0.84	0.91	0.87
Densenet121	Normal	0.92	0.69	0.79
	OSCC	0.76	0.94	0.84
Densenet169	Normal	0.91	0.77	0.83
	OSCC	0.80	0.92	0.86
Densenet201	Normal	0.92	0.84	0.88
	OSCC	0.86	0.93	0.89
EfficientNetB7	Normal	0.90	0.83	0.87
	OSCC	0.85	0.91	0.88
ViT-14	**Normal**	**0.99**	**0.97**	**0.98**
	OSCC	**0.97**	**0.99**	**0.98**

Values in bold indicate the best evaluation metrics



(a) Xception model (b) Resnet50 model (c) InceptionV3 model

(d) InceptionResnetV2 (e) Densenet121 model (f) Densenet169 model

(g) Densenet201 model (h) EfficientNetB7 model

Fig. 12 The confusion matrices of the DL models used for comparative study

5 Conclusion

Histopathological assessment by pathologists stands as the gold standard for detecting oral squamous cell carcinoma (OSCC). However, the intricate morphological variations in cancerous conditions pose a significant challenge for human evaluation. This study is a dedicated effort to aid clinicians in early OSCC identification. While deep learning (DL) models have advanced to enhance various applications for effective medical assessments, the incorporation of attention mechanisms into Vision Transformers (ViTs) introduces a level of precision that is essential in the medical industry, where inaccuracies could have profound consequences. The study introduces ViT-14, a fine-tuned ViT framework, specifically designed for classifying oral histopathology images into normal and OSCC categories across diverse publicly available datasets. The ViT-14 model demonstrates performance on par with or exceeding that of state-of-the-art models, emphasizing its effectiveness in early oral cancer detection using histopathological images. This study not only underscores the capabilities of ViTs in the field of medical imaging but also establishes ViT-14 as a promising instrument to assist clinicians in achieving more precise and timely diagnoses in cases of oral cancer.

The potential for enhancing oral cancer classification with fine-tuned ViT models is promising, but it is crucial to recognize certain limitations. Limited and imbalanced datasets may hinder generalization, and interpreting complex

Table 20 Comparison of ViT-14 model with other DL models for oral cancer histopathology image classification without cross-validation

Models	Acc. (%)	Spec. (%)	Sens. (%)	CKS	MCC
Xception	88.70	88.87	88.52	0.77	0.77
Resnet50	89.92	87.69	92.54	0.80	0.80
InceptionV3	84.07	83.14	85.10	0.68	0.68
Inception ResnetV2	86.69	83.94	90.09	0.73	0.74
Densenet121	81.85	75.80	92.31	0.64	0.66
Densenet169	84.68	80.34	90.78	0.69	0.70
Densenet201	88.71	85.77	92.34	0.77	0.78
EfficientNetB7	87.30	84.87	90.22	0.75	0.75
ViT-14	**97.78**	**96.72**	**98.80**	**0.96**	**0.96**

Values in bold indicate the best evaluation metrics

Table 21 Comparative analysis with previous research on different publicly available oral cancer histopathology datasets (Acc: Accuracy, Prec: Precision, Sens: Sensitivity)

Method	Year	Model/Classifier	Samples	Acc	Prec	Sens
Classification by Transfer learning Method [23]	2021	Resnet50 MobileNet Large CNN InceptionV3	290 normal epithelium images and 934 OSCC images [46]	91.13 85.48 76.61 89.52	88.00 81.00 58.00 89.00	87.00 79.00 52.00 82.00
10 layer CNN model [25]	2023	Customized CNN	290 normal epithelium images and 934 OSCC images [46]	**97.82**	97.00	98.00
Deep learning and hybrid techniques [47]	2023	Support vector machine	2698 OSCC images 2494 healthy tissue images [42]	97.00	96.77	90.90
Gabor filter and Resnet50 [48]	2023	Catboost classifier	2698 OSCC images 2494 healthy tissue images [42]	94.92	95.51	84.30
Transformer [49]	2023	External attention	2511 OSCC images 2435 healthy tissue images [43]	96.97	–	97.61
Transfer learning [24]	2022	Alexnet	2511 OSCC images 2435 healthy tissue images [43]	90.06	–	92.74
Proposed method	2023	ViT-14	290 normal epithelium images and 934 OSCC images [46]	95.12	95.05	98.97
Proposed method	2023	ViT-14	2698 OSCC images 2494 healthy tissue images [42]	**97.69**	98.21	97.51
Proposed method	2023	ViT-14	2511 OSCC images 2435 healthy tissue images [43]	**97.78**	96.88	98.80

Values in bold indicate the best evaluation metrics

models like ViT remains difficult. Class imbalance and the "black-box" nature of these models can introduce bias and limit explainability. Computational demands pose challenges for resource-limited institutions, and integrating these models into clinical workflows requires addressing privacy and regulatory issues. Despite these challenges, the future outlook is promising, with ongoing efforts to overcome these limitations through the accumulation of more diverse and expansive datasets, advancements in model interpretability, and optimization of computational efficiency for broader applicability in clinical settings.

Acknowledgements Bhaswati Singha Deo is thankful to IIT Kanpur for the Institute fellowship. The author MPal wishes to thank ABB Ability Innovation Centre, India, for their support in this work.

Author Contributions Bhaswati Singha Deo: Developed the model from the concept, Developed the code, Generated the results, Wrote the manuscript. Mayukha Pal: Conceived the idea and conceptualized it, Developed the methodology from the concept, Reviewed the manuscript and mentored the work. Prasanta K. Panigrahi: Reviewed the manuscript and mentored the work. Asima Pradhan: Developed the analysis methodology, Reviewed the manuscript and mentored the work.

Funding This research received no specific support from governmental, private, or non-profit funding agencies.

Data Availability The obtained datasets for our analysis are from free public database.

Code Availability The codes that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics approval Datasets are publicly available for research purpose, hence not applicable for this paper.

Consent to participate Datasets are publicly available for research purpose, hence not applicable for this paper.

Consent for publication Datasets are publicly available for research purpose, hence not applicable for this paper.

References

1. Scully, C., Bedi, R.: Ethnicity and oral cancer. Lancet Oncol. **1**(1), 37–42 (2000)
2. Tsai, M.-T., Lee, H.-C., Lee, C.-K., Yu, C.-H., Chen, H.-M., Chiang, C.-P., Chang, C.-C., Wang, Y.-M., Yang, C.: Effective indicators for diagnosis of oral cancer using optical coherence tomography. Opt. Express **16**(20), 15847–15862 (2008)
3. Montero, P.H., Patel, S.G.: Cancer of the oral cavity. Surg. Oncol. Clin. **24**(3), 491–508 (2015)
4. Albrecht, M., Schnabel, C., Mueller, J., Golde, J., Koch, E., Walther, J.: In vivo endoscopic optical coherence tomography of the healthy human oral mucosa: qualitative and quantitative image analysis. Diagnostics **10**(10), 827 (2020)
5. Chakraborty, D., Natarajan, C., Mukherjee, A.: Advances in oral cancer detection. Adv. Clin. Chem. **91**, 181–200 (2019)
6. Eckert, A.W., Kappler, M., Große, I., Wickenhauser, C., Seliger, B.: Current understanding of the hif-1-dependent metabolism in oral squamous cell carcinoma. Int. J. Mol. Sci. **21**(17), 6083 (2020)
7. Ghosh, A., Chaudhuri, D., Adhikary, S., Chatterjee, K., Roychowdhury, A., Das, A.K., Barui, A.: Deep reinforced neural network model for cyto-spectroscopic analysis of epigenetic markers for automated oral cancer risk prediction. Chemometrics Intell. Lab. Syst. **224**, 104548 (2022)
8. Kong, J., Sertel, O., Shimada, H., Boyer, K.L., Saltz, J.H., Gurcan, M.N.: Computer-aided evaluation of neuroblastoma on whole-slide histology images: classifying grade of neuroblastic differentiation. Pattern Recognit. **42**(6), 1080–1092 (2009)
9. Deif, M.A., Hammam, R.E.: Skin lesions classification based on deep learning approach. J. Clin. Eng. **45**(3), 155–161 (2020)
10. Santana, M.F., Ferreira, L.C.L.: Diagnostic errors in surgical pathology. J. Brasi. Patol. Med. Lab. **53**, 124–129 (2017)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
13. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019). PMLR
14. Ariji, Y., Kise, Y., Fukuda, M., Kuwada, C., Ariji, E.: Segmentation of metastatic cervical lymph nodes from ct images of oral cancers using deep-learning technology. Dentomaxillofac. Radiol. **51**(4), 20210515 (2022)
15. Jubair, F., Al-karadsheh, O., Malamos, D., Al Mahdi, S., Saad, Y., Hassona, Y.: A novel lightweight deep convolutional neural network for early detection of oral cancer. Oral Dis. **28**(4), 1123–1130 (2022)
16. Zhang, X., Liang, Y., Li, W., Liu, C., Gu, D., Sun, W., Miao, L.: Development and evaluation of deep learning for screening dental caries from oral photographs. Oral Dis. **28**(1), 173–181 (2022)
17. Park, J., Kim, Y.: Styleformer: Transformer based generative adversarial networks with style vector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8983–8992 (2022)
18. Dai, Y., Gao, Y., Liu, F.: Transmed: Transformers advance multimodal medical image classification. Diagnostics **11**(8), 1384 (2021)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems, vol. 30 (2017)
20. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
21. Parvaiz, A., Khalid, M.A., Zafar, R., Ameer, H., Ali, M., Fraz, M.M.: Vision transformers in medical computer vision-a contemplative retrospection. Eng. Appl. Artif. Intell. **122**, 106126 (2023)
22. Xu, H., Xu, Q., Cong, F., Kang, J., Han, C., Liu, Z., Madabhushi, A., Lu, C.: Vision transformers for computational histopathology. IEEE Rev. Biomed. Eng. (2023)
23. Palaskar, R., Vyas, R., Khedekar, V., Palaskar, S., Sahu, P.: Transfer learning for oral cancer detection using microscopic images. arXiv preprint arXiv:2011.11610 (2020)
24. Rahman, A.-U., Alqahtani, A., Aldhafferi, N., Nasir, M.U., Khan, M.F., Khan, M.A., Mosavi, A.: Histopathologic oral cancer prediction using oral squamous cell carcinoma biopsy empowered with transfer learning. Sensors **22**(10), 3833 (2022)
25. Das, M., Dash, R., Mishra, S.K.: Automatic detection of oral squamous cell carcinoma from histopathological images of oral mucosa using deep convolutional neural network. Int. J. Environ. Res. Public Health **20**(3), 2131 (2023)
26. Myriam, H., Abdelhamid, A.A., El-Kenawy, E.-S.M., Ibrahim, A., Eid, M.M., Jamjoom, M.M., Khafaga, D.S.: Advanced meta-heuristic algorithm based on particle swarm and al-biruni earth radius optimization methods for oral cancer detection. IEEE Access **11**, 23681–23700 (2023)
27. Muthu Rama Krishnan, M., Shah, P., Chakraborty, C., Ray, A.K.: Statistical analysis of textural features for improved classification of oral histopathological images. J. Med. Syst. **36**, 865–881 (2012)
28. Jeyaraj, P.R., Samuel Nadar, E.R.: Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm. J. Cancer Res. Clin. Oncol. **145**(4), 829–837 (2019)
29. Nanditha, B., Geetha, A., Chandrashekar, H., Dinesh, M., Murali, S.: An ensemble deep neural network approach for oral cancer screening (2021)

30. Flügge, T., Gaudin, R., Sabatakakis, A., Tröltzsch, D., Heiland, M., Nistelrooij, N., Vinayahalingam, S.: Detection of oral squamous cell carcinoma in clinical photographs using a vision transformer. Sci. Rep. **13**(1), 2296 (2023)
31. Zeid, M.A.-E., El-Bahnasy, K., Abo-Youssef, S.: Multiclass colorectal cancer histology images classification using vision transformers. In: 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS), pp. 224–230 (2021). IEEE
32. Chen, H., Li, C., Li, X., Rahaman, M.M., Hu, W., Li, Y., Liu, W., Sun, C., Sun, H., Huang, X., et al.: Il-mcam: An interactive learning and multi-channel attention mechanism-based weakly supervised colorectal histopathology image classification approach. Comput. Biol. Med. **143**, 105265 (2022)
33. Wu, Y., Qi, S., Sun, Y., Xia, S., Yao, Y., Qian, W.: A vision transformer for emphysema classification using ct images. Phys. Med. Biol. **66**(24), 245016 (2021)
34. Ambita, A.A.E., Boquio, E.N.V., Naval, P.C.: Covit-gan: vision transformer forcovid-19 detection in ct scan imageswith self-attention gan fordataaugmentation. In: Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part II 30, pp. 587–598 (2021). Springer
35. Tyagi, K., Pathak, G., Nijhawan, R., Mittal, A.: Detecting pneumonia using vision transformer and comparing with other techniques. In: 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 12–16 (2021). IEEE
36. Kamran, S.A., Hossain, K.F., Tavakkoli, A., Zuckerbrod, S.L., Baker, S.A.: Vtgan: Semi-supervised retinal image synthesis and disease prediction using vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3235–3245 (2021)
37. Zhang, J., Li, C., Liu, G., Min, M., Wang, C., Li, J., Wang, Y., Yan, H., Zuo, Z., Huang, W., et al.: A cnn-transformer hybrid approach for decoding visual neural activity into text. Comput. Methods Programs Biomed. **214**, 106586 (2022)
38. Chollet, F.: Xception: deep learning with depthwise separable convolutions In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
39. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
40. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI Conference on Artificial Intelligence (2017)
41. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
42. Oral cancer histopathology dataset. https://www.kaggle.com/datasets/ashenafifasilkebede/dataset. Accessed 17 Nov 2023
43. Oral cancer histopathology dataset. https://www.kaggle.com/datasets/ashenafifasilkebede/dataset?select=train. Accessed 17 Nov 2023
44. Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y.: Transformer in transformer. Adv. Neural. Inf. Process. Syst. **34**, 15908–15919 (2021)
45. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
46. Rahman, T.Y., Mahanta, L.B., Das, A.K., Sarma, J.D.: Histopathological imaging database for oral cancer analysis. Data Brief **29**, 105114 (2020)
47. Ahmad, M., Irfan, M.A., Sadique, U., Haq, I., Jan, A., Khattak, M.I., Ghadi, Y.Y., Aljuaid, H.: Multi-method analysis of histopathological image for early diagnosis of oral squamous cell carcinoma using deep learning and hybrid techniques. Cancers **15**(21), 5247 (2023)
48. Haq, I.U., Ahmad, M., Assam, M., Ghadi, Y.Y., Algarni, A.: Unveiling the future of oral squamous cell carcinoma diagnosis: an innovative hybrid ai approach for accurate histopathological image analysis. IEEE Access (2023)
49. Deo, B.S., Pal, M., Pradhan, A.: External-attention-based deep neural network model for reliable detection of oral cancer from histopathological images. In: Women in Optics and Photonics in India 2022, vol. 12638, pp. 25–28. SPIE (2023)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.