BMC Cancer

RESEARCH

Open Access

Artificial intelligence based vision transformer application for grading histopathological images of oral epithelial dysplasia: a step towards AI-driven diagnosis

Mahdi Hadilou[1], Nazanin Mahdavi[2], Elham Keykha[3*], Ali Ghofrani[4], Elahe Tahmasebi[3] and Masoud Arabfard[5]

Abstract

Background This study aimed to classify dysplastic and healthy oral epithelial histopathological images, according to WHO and binary grading systems, using the Vision Transformer (ViT) deep learning algorithm—a state-of-the-art Artificial Intelligence (AI) approach and compare it with established Convolutional Neural Network models (VGG16 and ConvNet).

Methods A total of 218 histopathological slide images were collected from the Department of Oral and Maxillofacial Pathology at Tehran University of Medical Sciences archive and combined with two online databases. Two oral pathologists independently labeled the images based on the 2022 World Health Organization (WHO) grading system (mild, moderate and severe), the binary grading system (low risk and high risk), including an additional normal tissue class. After preprocessing, the images were fed to the ViT, VGG16 and ConvNet models.

Results Image preprocessing yielded 2,545 low-risk, 2,054 high-risk, 726 mild, 831 moderate, 449 severe, and 937 normal tissue patches. The proposed ViT model outperformed both CNNs with the accuracy of 94% (VGG16:86% and ConvNet: 88%) in 3-class scenario and 97% (VGG16:79% and ConvNet: 88%) in 4-class scenario.

Conclusions The ViT model successfully classified oral epithelial dysplastic tissues with a high accuracy, paving the way for AI to serve as an adjunct or independent tool alongside oral and maxillofacial pathologists for detecting and grading oral epithelial dysplasia.

Keywords Artificial Intelligence, Deep learning, Oral epithelial dysplasia, Histopathological images

*Correspondence:
Elham Keykha
Dr.keykha@chmail.ir
Full list of author information is available at the end of the article

Background

Oral epithelial dysplasia (OED), include a group of lesions characterized by various histological and clinical alterations in the oral epithelium. These alterations indicate a risk of future malignant transformation and clinically manifest as white and red mucosal lesions, namely leukoplakia and erythroplakia [1]. Although these lesions rarely transform into malignancy during an individual's lifespan, they are the origin of approximately 80% to 90% of early oral squamous cell carcinomas (OSCCs) [2]. The onset and etiology of OED are poorly understood; however, they can be associated with human papillomavirus and alcohol and tobacco consumption or can be idiopathic [3, 4]. Various architectural and cytological changes occur in oral dysplastic tissue, including increased and abnormal mitoses, drop-shaped rete ridges, loss of polarity in basal cells, abnormal variations in nuclear and cellular size and shape, and an increased nuclear-cytoplasmic ratio [1].

Evaluating histopathological images prepared from biopsies of suspected tissues is the gold-standard method for diagnosing and grading dysplasia. Several internationally accepted grading systems exist to determine the stage of oral epithelial dysplasia. The latest version introduced by the World Health Organization (WHO) in 2022 [5, 6], known as the 'thirds' grading system, along with the 'binary' grading system [1], is most commonly employed by oral and maxillofacial pathologists. These grading systems are based on the extent to which the dysplastic features extend through the thickness of the epithelium. Given the correlation between the grade of dysplasia and the risk of malignant transformation [7, 8], management strategies range from routine follow-up, cessation of risk factors, photobiomodulation, and medication to more aggressive approaches such as excision by scalpel or laser ablation [9, 10].

Current evidence has demonstrated considerable intra- and inter-observer variability when using the existing gold-standard grading systems. Criticisms regarding the reproducibility of these grading systems suggest that they are not capable of eliminating subjectivity [11–13]. This issue can be addressed by developing more robust and objective grading systems or by decreasing subjectivity in the grading process through the design of auxiliary methods that can grade OED independently or act as adjuncts to oral and maxillofacial pathologists to reduce errors.

Recently, artificial intelligence (AI) approaches have been applied in various aspects of the medical field. AI models have demonstrated successful results in the detection and classification of cancer at molecular, cellular, and clinical levels [14]. In dentistry, AI has been used for tasks such as counting teeth, grading malocclusion in orthodontics, identifying and classifying treatments or lesions on teeth in radiographs, and diagnosing OSCC through clinical manifestations and histopathological images [15, 16]. Deep learning-based algorithms, a subdivision of AI, utilize multilayered neural networks capable of learning hierarchical features from data. The data repeatedly pass through these layers, and basic features are extracted or combined in subsequent layers until the model achieves an acceptable level of accuracy in classifying each data point into its corresponding class [15]. Convolutional neural networks (CNNs), a common method used for analyzing medical images, have shown promising results in detecting cancerous and precancerous oral epithelium in a number of studies [16, 17].

CNNs are widely used in computer vision tasks, but their effectiveness can be affected by several limitations, particularly in complex domains like medical imaging. Their focus on local feature extraction is a significant limitation. While CNNs excel at identifying patterns in localized regions of an image, this can prevent them from capturing global context. In histopathological images, where relationships between distant features are crucial for accurate diagnoses, CNNs may struggle to integrate this information effectively, potentially leading to suboptimal performance [18]. Another challenge with CNNs is their fixed receptive field. Although the receptive field increases with the depth of the network, it remains constrained by the architecture. The model may not be able to capture larger-scale features without increasing computational complexity due to this limitation, which may not be feasible when working with limited data. This sensitivity can lead to decreased performance when the data is not well-aligned or standardized, a common issue in medical imaging where variations in sample preparation can occur [19].

Overfitting is another concern with CNNs, particularly when working with limited datasets. CNNs may learn noise or irrelevant features instead of generalizable patterns, resulting in poor performance on unseen data. This lack of transparency can hinder trust in clinical applications, where understanding the reasoning behind a model's predictions is critical for acceptance and implementation [20].

Deep learning approaches like state-of-the-art Vision Transformers (ViTs) have shown promising results in outperforming CNNs for medical image analysis tasks. ViTs present several advantages that address the limitations of CNNs. One of the most notable benefits is their ability to understand global context. By utilizing self-attention mechanisms, ViTs can capture relationships between all parts of the input image, enabling a more holistic understanding of the data. This is particularly beneficial in histopathological analysis, where spatial

relationships between distant features can be significant for accurate grading and diagnosis [21].

ViTs are able to scale with data and show superior performance as the amount of training data increases. ViTs are more suitable for applications with abundant annotated data because they can leverage large datasets more effectively, unlike CNNs that often require extensive tuning to prevent overfitting. Additionally, ViTs offer flexibility in handling varying input sizes, as they segment images into patches and process them independently. This capability allows for greater flexibility in data preprocessing and can improve performance across diverse datasets [22].

While CNNs have been foundational in computer vision, their limitations—such as local feature extraction and sensitivity to input variations—underscore the need for alternative approaches. Vision Transformers offer significant advantages, including improved global context understanding and scalability with data, making them a compelling choice for complex tasks like histopathological image analysis. By addressing these claims with scientific rigor, researchers can better evaluate the potential of ViTs in enhancing diagnostic accuracy and interpretability in medical applications [23].

This is the first study using ViTs for classifying OED cases. Furthermore, a common issue in the current literature is the limited number of samples available to train these deep learning models. Therefore, providing more samples and combining multicenter specimens are necessary to develop a robust and reliable model. Given the subjectivity and individual-dependent nature of the current grading systems for OED, which significantly affect oral and maxillofacial pathologists' judgments, developing an AI-based algorithm to assist and support pathologists could result in faster, less costly, and less exhausting assessments for medical staff. This study aimed to develop a state-of-the-art ViT model to accurately classify healthy and dysplastic oral epithelial histological images based on WHO and binary grading systems.

Methods

Data acquisition

The Research Ethics Committee of Baqiyatallah University of Medical Sciences (IR.BMSU.REC.1402.076) approved the study design. This manuscript is written in accordance with the latest checklists presented in the AI literature (supplementary material 1) to ensure standardized research reporting. [24–26] The major part of the dataset was obtained from the archive of the Department of Oral and Maxillofacial Pathology at Tehran University of Medical Sciences, Tehran, Iran (Table 1). This archive consists of diagnostic hematoxylin and eosin (H&E)-stained histopathological slides from specimens suspected of disease, which were sent to this center for diagnosis over the past 40 years. For the current study, the slides of OED lesions (leukoplakia and erythroplakia) and normal epithelium samples (irritation fibroma and mucocele) were selected from specimens archived between 1998 and 2023. Each specimen belonged to a single individual.

Table 1 Summary of the sources of the study dataset

Number of specimens		Study dataset			
		Original dataset	Ribeiro-de-Assis et al	Rahman et al	Total
Binary system	Low risk	71	13	-	84
	High risk	85	10	-	95
WHO system	Mild	41	13	-	54
	Moderate	65	7	-	72
	Severe	43	10	-	53
Normal tissue		15	-	24	39

Each sample had a confirmed diagnosis by two oral and maxillofacial pathology residents and a professor at the time it was obtained. Further, two additional pathologists confirmed those diagnoses during the current study. They captured and labeled the images based on the 2022 WHO classification system (mild, moderate, and severe) and the binary grading system (low-risk and high-risk) [27]. To ensure a homogeneous database, the pictures were taken under the same ambient conditions, including lighting, microscope settings, and staining type. High-resolution photomicrograph images from OED slides were captured using a camera attached to an Olympus D52 microscope at × 10 magnification which enabled the authors to zoom up to greater magnifications making it possible to observe cellular alterations to label the patches.

Because a significantly large number of samples is needed to train deep learning models and develop a generally applicable model, two other free online databases (Ribeiro-de-Assis et al. [28] and Rahman et al. [29]) were used to enrich the original dataset (Table 1).

The primary objective of this study was to grade OED using advanced deep learning techniques. The authors were focused on fine-tuning a ViT model to meet the specific requirements of this medical image classification task. The ViT model was fine-tuned under two distinct scenarios. In the "3-Class Classification" scenario, the model was trained to classify images into three categories: low risk of malignancy, high risk of malignancy, and normal epithelial tissue. This approach was designed to evaluate the model's ability to differentiate between varying degrees of dysplasia, which is critical for early detection and intervention in OED. In the "4-Class

Classification" scenario, the classification task was expanded to four classes using the WHO system, assessing how the model performs with a more granular classification challenge. This added complexity allowed the authors to explore further the potential of transformers in handling nuanced differences in medical images.

Image preprocessing

Figure 1 illustrates a summary of the image preprocessing steps. First, the images were rotated to align the epithelium horizontally, positioning the basal lamina at the lowest level. Since only the epithelium was required for identifying normal tissue and grading dysplastic tissue, the surrounding structures were cropped, leaving a margin of connective tissue at the bottom and a blank margin on top. Any artifacts in the connective tissue and upper blank margin were omitted without disrupting the integrity of the epithelium. All images displayed the complete width of the epithelium. Then the images were horizontally split into patches with a width of 300 pixels, moving with a 100-pixel stride to create a 200-pixel overlap. Two oral pathologists labeled the resulting patches according to the WHO and binary dysplasia grading systems, resolving any disagreements through group discussion. If the structure of the epithelium in any patch was unsuitable for grading—due to reasons such as very narrow epithelium, the presence of ruptures, or lack of diagnostic clarity— it was excluded from further processing.

Dataset splitting methodology

A train-validation-test data splitting approach was used to ensure the reliability, generalizability and robustness of the model's performance which involved isolating 20% of the dataset as an independent test set. This test set remained unused during the training and validation phases, serving as an unbiased evaluation of performance on unseen data. The remaining 80% of the data was divided into training and validation sets, with 25% allocated for validation. The validation set monitored the model's performance during training, guided hyperparameter tuning, and informed decisions regarding early stopping and model selection. This structured approach mitigates the risks associated with small datasets and is a widely accepted practice in machine learning, enhancing the reliability and validity of the findings.

Dataset augmentation methodology

To enhance the model's robustness and improve its generalization capabilities, data augmentation techniques were applied exclusively to the training dataset. Data augmentation is critical in training deep learning models, as it artificially expands the diversity of the training data and reduces the risk of overfitting. The training data underwent several augmentation techniques. First, varying the brightness of the images within a range of [0.75, 1.25] simulated different lighting conditions, helping the model learn to recognize features under varying illumination. This range ensures that the brightness is adjusted moderately—reducing it by up to 25% or increasing it by up to 25%—to introduce variability without distorting key features [30, 31]. Such adjustments are commonly used in image augmentation to enhance model robustness and generalization. Studies have shown that moderate brightness adjustments, such as those within the range of [0.75, 1.00], can improve model performance by simulating natural variations in lighting conditions, particularly in medical imaging datasets. In medical imaging, maintaining the visibility of critical features is paramount. Excessive brightness changes (e.g., below 0.5 or above 1.5) can obscure or distort key details, potentially degrading model performance. The selected range balances variability with feature preservation [32]. Modifying the contrast

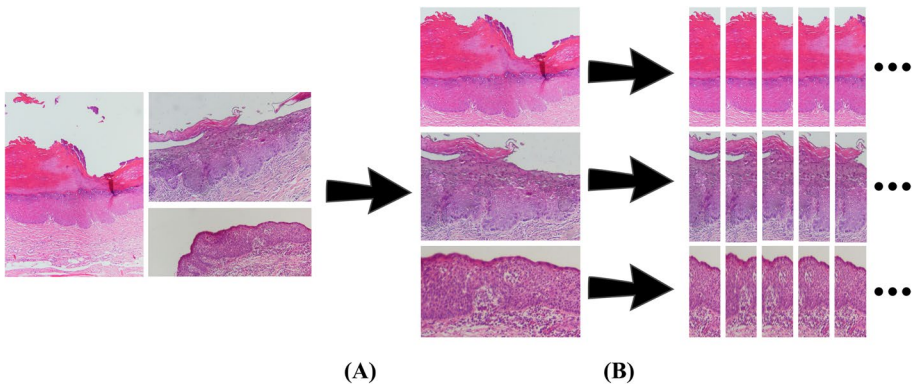

Fig. 1 Image preprocessing stages. **A** Images are cropped and rotated; artifacts are omitted without disrupting the epithelium's integrity. **B** Images are split into segments with a width of 300 pixels and a 200-pixel overlap

of the images within the same range aimed to improve the model's ability to distinguish subtle differences in image features, which is particularly important in medical imaging.

Additionally, adjusting the saturation of the images within the [0.75, 1.25] range helped the model become more resilient to variations in color intensity due to differences in imaging equipment or specimen processing settings. Horizontal flipping was also applied to the images, increasing the diversity of the training set by introducing mirrored versions and allowing the model to learn invariant features regardless of orientation. By implementing these augmentation techniques, a more diverse training dataset was created to help the model generalize unseen data better. This approach not only increased the effective size of the training set but also enhanced the model's ability to adapt to variations in real-world scenarios. Overall, combining brightness, contrast, and saturation adjustments with flipping provided a comprehensive augmentation strategy that contributed to the robustness and performance of the model during training.

Model architecture and configuration

The computational setup for this study was powered by an NVIDIA RTX 3070 GPU, 32 GB of RAM, and an Intel Core i7-7700HQ CPU. This configuration provided the necessary computational power to efficiently train and fine-tune the ViT models for both the 3-class and 4-class classification tasks. The GPU's capabilities were particularly beneficial in handling the large-scale computations required by the transformer-based architecture, enabling faster training times and more efficient experimentation.

A pre-trained ViT model, specifically the ViT-B16 variant trained on the ImageNet-21 K dataset, served as the feature extraction component. This choice leveraged the rich feature representations learned from a large and diverse dataset. The input images were resized to 224×224 pixels and divided into 8×8 pixel patches, resulting in 784 patches per image. Each patch was represented by 192 elements, enabling the model to analyse and extract features for the classification tasks effectively. The number of 8×8 patches containing OED features was 17,376,768 without data augmentation. With data augmentation, this number effectively becomes infinite, allowing for robust training and reducing the risk of overfitting.

After feature extraction using the ViT, the methodology flattened the extracted features into a one-dimensional vector. A multi-layer perceptron (MLP) stack then further processed these features. Batch normalization stabilized and accelerated the training process. The MLP consisted of three fully connected layers with 128, 64, and 32 neurons, respectively. Each of these layers employed Gaussian Error Linear Units (GELU) as the activation function, enhancing the model's ability to capture complex patterns in the data. Finally, the output layer corresponded to the number of classes in the classification task. For the 3-class experiment, the output layer consisted of three neurons; for the 4-class experiment, it contained four neurons. A softmax activation function was applied to this output layer to produce probability distributions over the classes, facilitating multi-class classification.

The model's training utilized the Rectified Adam optimizer, facilitating efficient convergence. All components of the model were set to be trainable, enabling fine-tuning on the specific medical imaging dataset to improve performance. The loss function employed was categorical cross-entropy, with label smoothing set to 0.2. This approach mitigated overfitting by preventing the model from becoming overly confident in its predictions, thereby enhancing generalization to unseen data. Figure 2 illustrates the ViT architecture used in the current study.

Comparison with VGG16 and ConvNet

In this study, the proposed ViT model was compared with VGG16 [33] and ConvNet [34] established architectures exactly the same configuration as ViT was trained with. This comparison aimed to evaluate the performance and efficiency of the ViT model relative to these well-known CNNs. VGG16 is a deep CNN that consists of 16 layers with learnable weights, primarily using small 3×3 convolutional filters. Its architecture emphasizes depth and simplicity, making it effective for image classification tasks. However, it is known for its high computational cost and large model size, which can limit its applicability in resource-constrained environments. ConvNet is a class of deep neural networks designed for processing structured grid data, particularly images. They consist of several key components, including convolutional layers that apply filters to extract features, activation functions like ReLU that introduce non-linearity, and pooling layers that reduce spatial dimensions while maintaining important information. Fully connected layers at the end of the network make final predictions based on the learned features, while dropout layers help prevent overfitting. This architecture has led to significant advancements in applications such as image classification, object detection, and semantic segmentation, establishing ConvNet as a cornerstone of modern computer vision. The models were evaluated based on several key metrics, including accuracy, training time, and model size. The proposed ViT model was assessed against VGG16 and ConvNet to

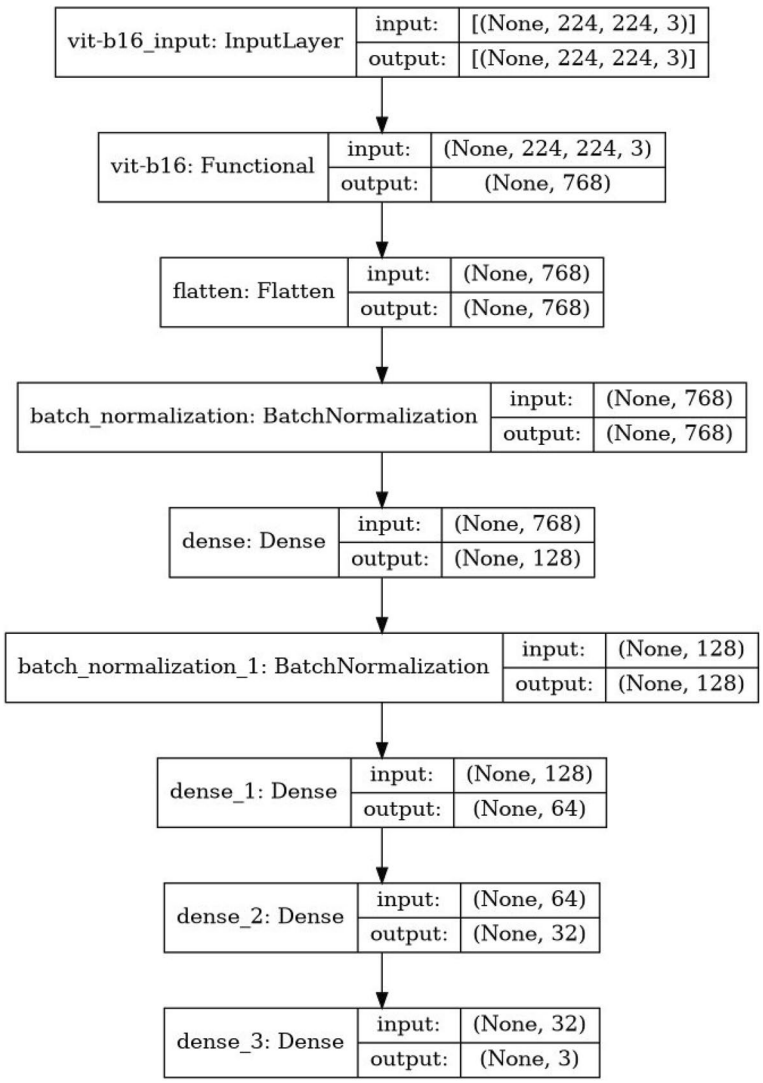

Fig. 2 The ViT architecture employed in the current study for the three-class experiment is similar to the architecture used for the four-class experiment, with the only difference being the last layer, which contains four neurons instead of three

determine its effectiveness in terms of both performance and computational efficiency.

Results

A total of 2,545 low-risk, 2,054 high-risk, 726 mild, 831 moderate, 449 severe, and 937 normal tissue patches were delivered to ViT algorithm after image preprocessing. The developed model achieved an accuracy of 94% for the binary plus normal tissue (3-class) scenario (Table 2), and 97% for the WHO plus normal tissue (4-class) scenario (Table 3). Figure 3 (left) displays the training and validation loss over 10 epochs. The training loss decreases steadily, indicating effective learning by the model. The validation loss also decreases but begins to diverge slightly from the training loss around epoch 9, suggesting the onset of overfitting. Figure 3 (right) illustrates the accuracy of training and validation over the same 10 epochs. As illustrated in Fig. 3 (left), the training loss decreases steadily, indicating effective learning by the model. However, the validation loss begins to diverge slightly from the training loss around epoch 9, suggesting the onset of overfitting. This divergence is further supported by Fig. 3 (right), which shows that while both training and validation accuracies increase rapidly during the initial epochs, the validation accuracy plateaus around epoch 9. In contrast, the training accuracy continues to rise slightly, reinforcing the indication of potential overfitting.

To assess overfitting in the ViT model, multiple quantitative measures were employed, including consistent

Table 2 ViT, VGG16 and ConvNet model evaluation for the 3-class scenario

	Precision			Recall			F1-score		
Models	**ViT**	**VGG16**	**ConvNet**	**ViT**	**VGG16**	**ConvNet**	**ViT**	**VGG16**	**ConvNet**
Low risk	0.97	0.88	0.84	0.89	0.81	0.91	0.93	0.84	0.87
High risk	0.88	0.81	0.89	0.97	0.87	0.82	0.92	0.84	0.85
Normal	1.00	0.94	0.95	1.00	0.96	0.93	1.00	0.95	0.94
Accuracy							0.94	0.86	0.88
Macro average	0.95	0.87	0.89	0.95	0.88	0.89	0.95	0.88	0.89
Weighted average	0.95	0.87	0.88	0.94	0.86	0.88	0.94	0.86	0.88

Table 3 ViT, VGG16 and ConvNet model evaluation for the 4-class scenario

	Precision			Recall			F1-score		
Models	**ViT**	**VGG16**	**ConvNet**	**ViT**	**VGG16**	**ConvNet**	**ViT**	**VGG16**	**ConvNet**
Mild	0.95	0.80	0.85	0.97	0.82	0.90	0.96	0.81	0.87
Moderate	0.96	0.75	0.87	0.96	0.71	0.80	0.96	0.73	0.84
Severe	0.97	0.70	0.81	0.93	0.72	0.90	0.95	0.71	0.86
Normal	1.00	0.95	0.98	1.00	0.95	0.95	1.00	0.95	0.97
Accuracy							0.97	0.79	0.88
Macro average	0.97	0.80	0.88		0.80	0.89	0.97	0.80	0.88
Weighted average	0.97	0.79	0.88		0.79	0.88	0.97	0.79	0.88

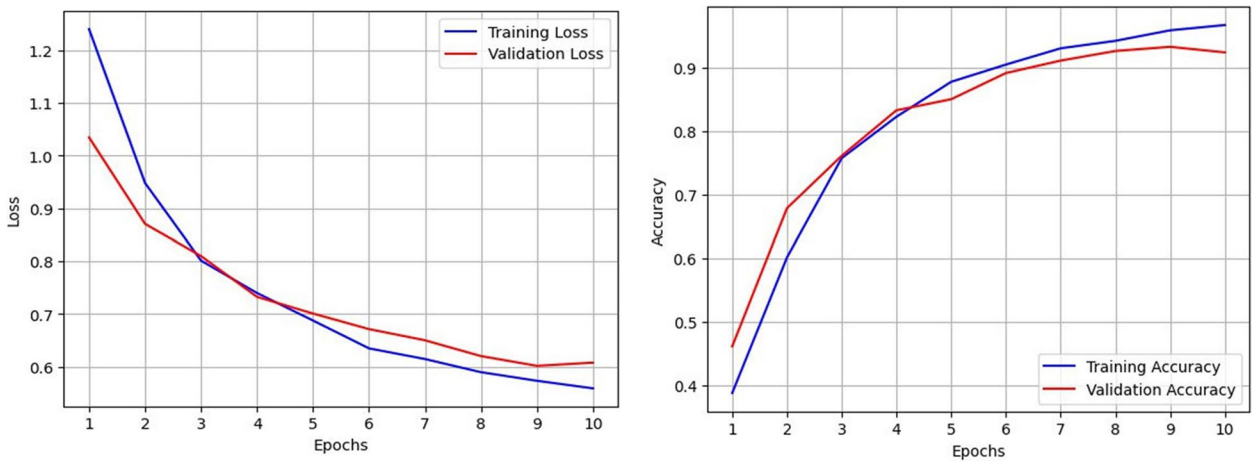

Fig. 3 Training-validation loss and accuracy for the 3-class scenario

accuracy, precision, recall, and F1-scores across cross-validation folds. In addition to these quantitative metrics, qualitative measures were utilized such as the inspection of ViT attention maps to confirm that the model focuses on anatomically relevant features rather than background artifacts. The stability of performance metrics across different cross-validation folds and the external test set indicates that the learned representations are capturing pathology-relevant variations rather than overfitting to a limited number of samples. Notably, it was observed that performance metrics remained stable or improved across folds, reinforcing our confidence in the model's generalizability. Although the total number of samples was modest, our patch-based strategy effectively increased the diversity of the training data. Furthermore, standard deep learning regularization techniques were implemented, which further mitigate the overfitting risk inherent in deep learning models, particularly in settings with

limited data. The observed stability of performance metrics and the alignment of attention maps with known histopathological features collectively support the validity of the study's approach. Both accuracies increase rapidly during the initial epochs and reach high levels. However, similar to the loss, the validation accuracy plateaus around epoch 9, while the training accuracy continues to rise slightly, further indicating potential overfitting. Figure 4 demonstrates similar behavior observed during training in the 4-class scenario.

Two samples are shown in Fig. 5, representing normal, low risk, and high risk in 3-class scenario, and normal, mild, moderate, and severe in 4-class scenario. The left images illustrate the original samples fed to ViT, highlighting cellular morphology. The right images display ViT-generated attention maps, where brighter regions indicate stronger model focus. Notably, the model appears to emphasize alterations in epithelial architecture that underlie the grading of oral epithelial dysplasia.

In this study, the performance of the proposed ViT model was evaluated against two established CNN architectures, VGG16 and ConvNet, in both the "3-Class Classification" and "4-Class Classification" scenarios. In the 3-Class Classification scenario, VGG16 achieved an accuracy of 0.86, while ConvNet demonstrated a slightly higher accuracy of 0.88. whereas the ViT model outperformed both, achieving an impressive accuracy of 0.94. In the more complex 4-Class Classification scenario, VGG16 attained an accuracy of 0.79, whereas ConvNet significantly outperformed it with an accuracy of 0.88. Despite this, our model excelled, achieving a remarkable accuracy of 0.97. (Tables 2 and 3) These results highlight the advantages of our model in managing increased class variability and underscore the effectiveness of modern architectures in enhancing classification performance. Overall, our model consistently outperformed both VGG16 and ConvNet across both scenarios, demonstrating its superior capability in classification tasks.

Discussion

Examining histopathological images obtained from biopsies by oral and maxillofacial pathologists remains the gold-standard method for diagnosing and grading the severity of OED. This approach is crucial for timely diagnosis and selecting appropriate treatment [1]. However, current grading systems suffer from subjectivity and poor reproducibility [11–13]. Given the promising results of AI-based approaches in diagnostic dentistry—aimed at assisting or even replacing clinicians—this study prepared a rich database of dysplastic and normal oral epithelial histopathological images and established a ViT model [35] to classify images based on the WHO 2022 and binary grading systems and compare it with the VGG16 and ConvNet CNNs.

This was the first study evaluating the efficiency of ViT models for classification of the OED samples. Recent studies about the classification of OED mostly have used CNNs for this task. Peng et al. [36] evaluated four CNNs (ResNet-50, Inception-V4, ShuffleNet-V2, and EfficientNet-B0) to classify a dataset of 56 whole-slide images of OED into four classes (mild, moderate, severe dysplasia, and hyperplasia), achieving a slide-level accuracy of 63.5% with the optimal model (EfficientNet-B0). Similarly, Nguyen et al. [16] used a 4-class dataset (normal epithelium, low-grade dysplasia, high-grade dysplasia, and OSCC) consisting of 203 whole-slide images at 200× magnification. Using the Inception-v3 CNN, they achieved a high area under the curve (AUC) of 0.996. Liu et al. [37] performed a classification scenario on 112 whole slide images of normal and dysplastic tissue

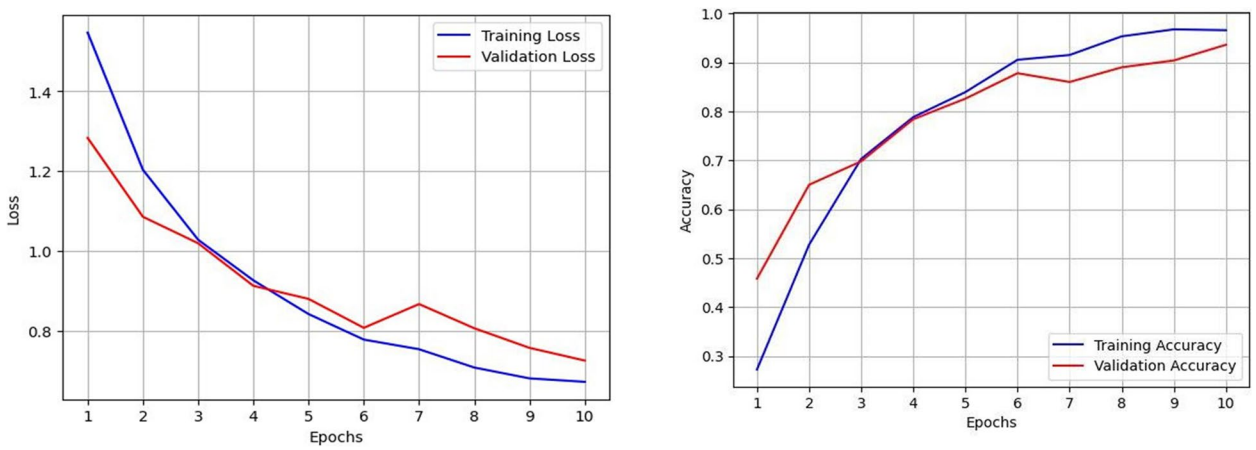

Fig. 4 Training-validation loss and accuracy for the 4-class scenario

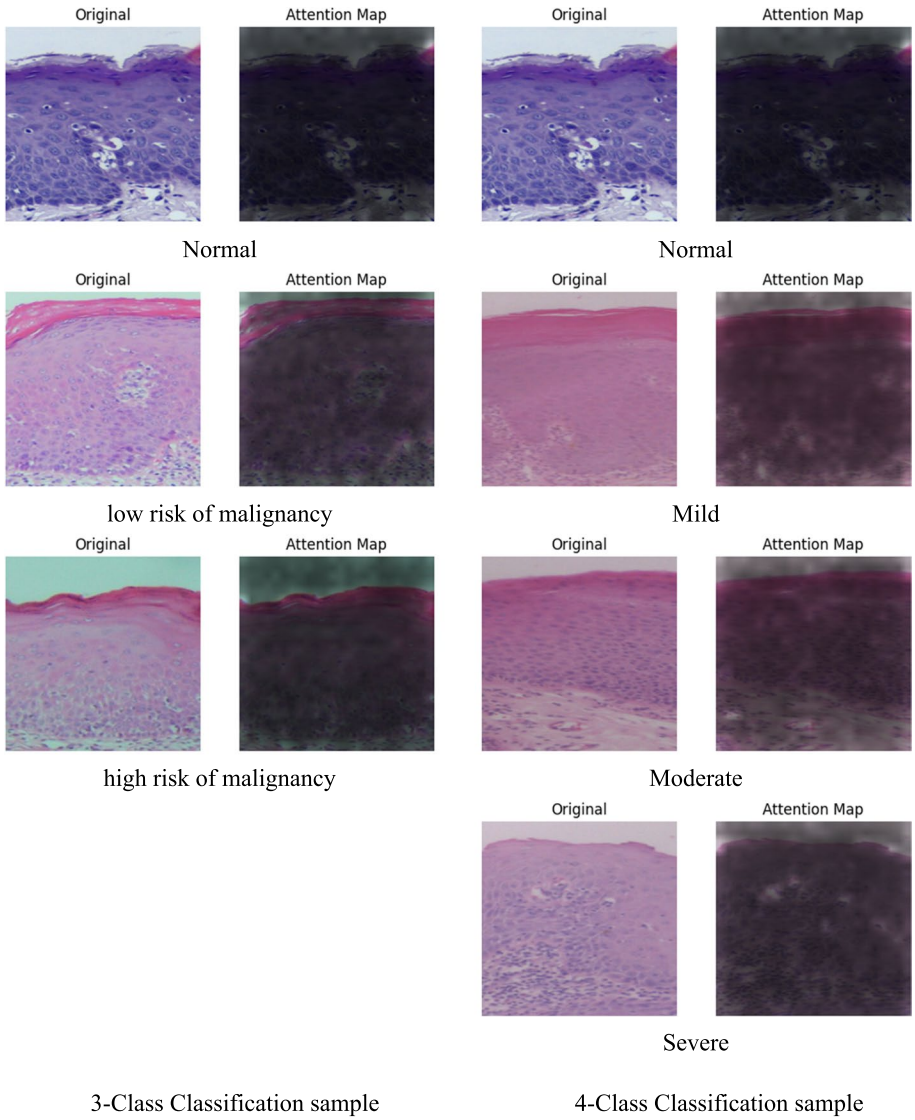

Fig. 5 Representative histopathological images of oral epithelial dysplasia and corresponding ViT attention maps

classes using DeepLabv3 CNN model which resulted a 93.3% accuracy. The results of the current study showed that the proposed ViT model successfully classified OED in both of the 4-class and 3-class scenarios with accuracies of 97% and 94%, respectively. It outperformed both CNNs with a significant difference.

Recently, several studies have aimed to develop deep learning algorithms, mostly using CNNs, to grade OED histological images. They typically follow three strategies: feature-based differentiation [38, 39], texture-based differentiation [16], and a combination of both [36]. The feature-based approach detects different cellular components and their distribution throughout the epithelium, while the texture-based approach differentiates an area's texture from surrounding structures. In this study, the texture-based differentiation strategy was employed due to its advantages of speed and reduced risk of errors in feature extraction as ViTs classify OED images through differentiating the tissue textures (texture-based) rather than individual objects in the images (feature-based) [40–42].

Deep learning approaches like ViTs have shown promising results in outperforming CNNs in medical image analysis tasks. While CNNs excel at extracting local features, they struggle to capture global context and long-range dependencies. ViTs, on the other hand, have demonstrated strong capabilities in handling global information and modeling long-range dependencies [23].

This is particularly advantageous for medical images, which require integrating both local features and global dependencies for accurate analysis. However, ViTs have limitations when dealing with high-resolution medical images common in modalities like whole-slide pathology, MRI, and CT scans. Their attention mechanism suffers from high quadratic complexity for long sequences, making them computationally expensive [43].

Additionally, ViTs tend to overfit when trained on limited medical datasets, which is often the case in the healthcare domain. While existing models are mostly trained from scratch, the impact of pretraining on their performance in the medical domain remains unclear. Pretraining has proven effective for data-efficient medical image analysis with CNNs, and understanding its effectiveness with ViT-based models could provide valuable insights for enhancing deep learning in medical imaging applications [41].

ViTs offer distinct advantages over CNNs in medical imaging. A major strength of ViTs is their ability to capture long-range dependencies within images. While CNNs focus on local features through convolutional layers, ViTs partition images into non-overlapping patches and process these patches as sequential data. This approach enables ViTs to model relationships between different regions of an image effectively, which is particularly beneficial in medical diagnostics, where spatial relationships are crucial for accurate diagnosis.

Moreover, ViTs demonstrate enhanced performance when scaled to larger datasets. Although CNNs can be effective with smaller datasets using transfer learning, ViTs pre-trained on extensive datasets like ImageNet can utilize these comprehensive learned representations when fine-tuned on medical imaging tasks. This capability is especially important in medical imaging, where labeled data is often scarce, allowing ViTs to achieve superior performance in tasks such as classification and detection [41, 44].

Recent advancements in self-supervised learning have further amplified the potential of ViTs. By pre-training on unlabeled data, ViTs can develop robust feature representations without the need for extensive labeled datasets—a significant advantage in medical imaging, where acquiring labeled data is both challenging and costly. This approach enhances the model's performance on downstream tasks. The flexible architecture of ViTs also permits adaptation to various tasks, including segmentation and classification. Hybrid models like TransUNet combine the strengths of CNNs for feature extraction with ViTs for capturing global context, leading to improved performance in medical image segmentation tasks.

Empirical studies have demonstrated that ViTs can outperform CNNs in specific medical imaging applications. For example, research indicates that ViTs achieve higher accuracy in detecting bone fractures and identifying tumors compared to traditional CNN models. These performance gains highlight the potential of ViTs in clinical settings where accuracy is paramount. Additionally, ViTs may exhibit reduced overfitting when trained on limited datasets, especially when combined with self-supervised learning techniques. This advantage contributes to developing more robust models that generalize better to unseen data—a critical requirement in medical imaging applications characterized by data variability [45, 46].

As AI systems, particularly in medical imaging, become increasingly integrated into diagnostic processes, understanding the features that contribute to predictive results is crucial. One promising approach is the integration of Explainable AI (XAI) techniques. Methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) can be employed to analyze feature importance, helping to identify which specific features or pixels in histopathological images are most influential in the model's predictions [47]. Another vital direction is enhancing model interpretability through the use of simpler, inherently interpretable models, such as decision trees or rule-based systems. These models can serve as a baseline for understanding feature importance in comparison to more complex architectures. Techniques like Layer-wise Relevance Propagation (LRP) can also be utilized to trace back the contributions of individual pixels to the final prediction, thereby improving our understanding of how the model processes the input data.

Collaboration with domain experts, such as pathologists, is essential for validating the identified features. Their insights can guide the interpretation of model outputs and ensure that the features align with biological or clinical relevance. Moreover, combining AI-derived features with domain-specific features—such as histological characteristics—can enhance both model interpretability and accuracy. This integration fosters a more comprehensive understanding of the underlying mechanisms driving predictive results. Finally, ensuring transparency in AI systems is critical for fostering trust among clinicians. Developing standardized reporting frameworks for AI systems in medical diagnostics will help ensure that the decision-making process is transparent and understandable. User-friendly interfaces that allow clinicians to interact with the model's predictions and explanations can further facilitate better clinical decision-making.

In conclusion, while CNNs have been the conventional choice for medical imaging tasks, Vision Transformers are emerging as a promising alternative. Their ability to capture complex relationships within images, adapt to

various tasks, and leverage large datasets makes them a compelling option for advancing medical image analysis.

Strengths and limitations

According to the risk of bias assessment checklist used in the systematic review by Mahmood et al. [17] in 2020—which evaluated the application of AI in diagnosing head and neck premalignant and malignant lesions—the current study achieved a high score of 10 out of 13. The main strengths of this study include (1) the use of state-of-the-art ViT algorithms, (2) the incorporation of a multicenter dataset by merging available online databases with our primary dataset to enhance generalizability, and (3) the involvement of two pathologists in annotating the image patches. However, the study is limited by the relatively small sample size, which is critical for deep learning algorithms to produce more robust results.

Conclusions

The results from both classification scenarios indicate that the transformer-based ViT model is highly effective for our dataset's unique modality. The model achieved strong performance metrics across all classes, highlighting its potential for clinical application in classifying OED. Both the 3-class and 4-class models demonstrated high accuracy, with slight variations depending on the granularity of the classification task. These findings pave the way for utilizing AI-based tools to assist clinicians and enhance the accuracy and objectivity of grading OED.

Abbreviations

OED	Oral epithelial dysplasia
OSCC	Oral squamous cell carcinomas
WHO	World Health Organization
AI	Artificial intelligence
CNN	Convolutional neural network
ViT	Vision transformers
MLP	Multi-layer perceptron
GELU	Gaussian Error Linear Units
AUC	Area under the curve

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12885-025-14193-x.

Supplementary Material 1.

Acknowledgements

Special thanks to the Department of Oral and Maxillofacial Pathology at Tehran University of Medical Sciences staff. Their extraordinary contributions have greatly improved the quality of this research.

Authors' contributions

MH: Writing – review & editing, Writing – original draft, Investigation. NM: Methodology, Data curation. EK: Conceptualization, Project administration, Writing – review & editing, Writing – original draft. AG: Formal analysis, Methodology, Writing – review & editing. ET: Validation and Supervision. MA: Validation, Supervision, Writing – review & editing, Formal analysis. All authors read and approved the final manuscript.

Funding

Not applicable.

Data availability

The data used to support the findings of this study are available from the corresponding author upon request.

Declarations

Ethics approval and consent to participate

The Research Ethics Committee of Baqiyatallah University of Medical Sciences (IR.BMSU.REC.1402.076) approved the study design.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Research Center for Prevention of Oral and Dental Diseases, Baqiyatallah University of Medical Sciences, Tehran, Iran. [2]Department of Oral and Maxillofacial Pathology, School of Dentistry, Tehran University of Medical Sciences, Tehran, Iran. [3]Research Center for Prevention of Oral and Dental Diseases, School of Dentistry, Baqiyatallah University of Medical Sciences, Tehran, Iran. [4]Student Research Committee, Baqiyatallah University of Medical Sciences, Tehran, Iran. [5]Chemical Injuries Research Center, Systems Biology and Poisonings Institute, Baqiyatallah University of Medical Sciences, Tehran, Iran.

Received: 29 November 2024 Accepted: 21 April 2025
Published online: 25 April 2025

References

1. Odell E, Kujan O, Warnakulasuriya S, Sloan P. Oral epithelial dysplasia: Recognition, grading and clinical significance. Oral Dis. 2021;27(8):1947–76.
2. Maymone MBC, Greer RO, Kesecker J, Sahitya PC, Burdine LK, Cheng AD, Maymone AC, Vashi NA. Premalignant and malignant oral mucosal lesions: Clinical and pathological findings. J Am Acad Dermatol. 2019;81(1):59–71.
3. Shanbhag VKL. New definition proposed for oral leukoplakia. Dent Res J (Isfahan). 2017;14(4):297–8.
4. Greer RO, Eversole LR, Crosby LK. Detection of human papillomavirus-genomic DNA in oral epithelial dysplasias, oral smokeless tobacco-associated leukoplakias, and epithelial malignancies. J Oral Maxillofac Surg. 1990;48(11):1201–5.
5. Reibel J, Gale N, Hille J, Hunt JL, Lingen M, Muller S, Sloan P, Tilakarante WM, Westra WH, Williams MD, et al. Oral potentially malignant disorders and oral epithelial dysplasia : oral potentially malignant disorders. WHO classification of head and neck tumours. 2017;9:112.
6. Muller S, Tilakaratne WM. Update from the 5th Edition of the World Health Organization Classification of Head and Neck Tumors: Tumours of the Oral Cavity and Mobile Tongue. Head Neck Pathol. 2022;16(1):54–62.
7. Iocca O, Sollecito TP, Alawi F, Weinstein GS, Newman JG, De Virgilio A, Di Maio P, Spriano G, Pardiñas López S, Shanti RM. Potentially malignant disorders of the oral cavity and oral dysplasia: A systematic review and meta-analysis of malignant transformation rate by subtype. Head Neck. 2020;42(3):539–55.
8. Yan F, Reddy PD, Nguyen SA, Chi AC, Neville BW, Day TA. Grading systems of oral cavity pre-malignancy: a systematic review and meta-analysis. Eur Arch Otorhinolaryngol. 2020;277(11):2967–76.
9. Monteiro L, Barbieri C, Warnakulasuriya S, Martins M, Salazar F, Pacheco JJ, Vescovi P, Meleti M. Type of surgical treatment and recurrence of oral leukoplakia: A retrospective clinical study. Med Oral Patol Oral Cir Bucal. 2017;22(5):e520–6.

10. Nadeau C, Kerr AR. Evaluation and Management of Oral Potentially Malignant Disorders. Dent Clin North Am. 2018;62(1):1–27.
11. Geetha K, Leeky M, Narayan T, Sadhana S, Saleha J. Grading of oral epithelial dysplasia: Points to ponder. Journal of Oral and Maxillofacial Pathology. 2015;19(2):198–204.
12. Manchanda A, Shetty DC. Reproducibility of grading systems in oral epithelial dysplasia. Med Oral Patol Oral Cir Bucal. 2012;17(6):e935-942.
13. Sa R, Np B. Hegde U, K U, G S, G K, Sil S: Inter- and Intra-Observer Variability in Diagnosis of Oral Dysplasia. Asian Pac J Cancer Prev. 2017;18(12):3251–4.
14. Bhinder B, Gilvary C, Madhukar NS, Elemento O. Artificial Intelligence in Cancer Research and Precision Medicine. Cancer Discov. 2021;11(4):900–15.
15. Ahmed N, Abbasi MS, Zuberi F, Qamar W, Halim MSB, Maqsood A, Alam MK. Artificial Intelligence Techniques: Analysis, Application, and Outcome in Dentistry-A Systematic Review. Biomed Res Int. 2021;2021:9751564.
16. Nguyen PTH, Sakamoto K, Ikeda T. Deep-learning application for identifying histological features of epithelial dysplasia of tongue. Journal of Oral and Maxillofacial Surgery Medicine and Pathology. 2022;34(4):514–22.
17. Mahmood H, Shaban M, Indave BI, Santos-Silva AR, Rajpoot N, Khurram SA. Use of artificial intelligence in diagnosis of head and neck precancerous and cancerous lesions: A systematic review. Oral Oncol. 2020;110:104885.
18. Abdou MA. Literature review: Efficient deep neural networks techniques for medical image analysis. Neural Comput Appl. 2022;34(8):5791–812.
19. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaría J, Fadhel MA, Al-Amidie M, Farhan L. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of big Data. 2021;8:1–74.
20. Salehi AW, Khan S, Gupta G, Alabduallah BI, Almjally A, Alsolai H, Siddiqui T, Mellit A. A study of CNN and transfer learning in medical imaging: Advantages, challenges, future scope. Sustainability. 2023;15(7):5930.
21. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in vision: A survey. ACM computing surveys (CSUR). 2022;54(10s):1–41.
22. Maurício J, Domingues I, Bernardino JJAS. Comparing vision transformers and convolutional neural networks for image classification: A literature review. 2023;13(9):5521.
23. Liu J, Yang H, Zhou HY, Xi Y, Yu L, Li C, Liang Y, Shi G, Yu Y, Zhang S, Zheng H. Swin-umamba: mamba-based unet with imagenet-based pretraining. In: International conference on medical image computing and computer-assisted intervention. Cham: Springer Nature Switzerland; 2024. pp. 615–625
24. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, Kahn CE: Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. Radiology: Artificial Intelligence. 2024; 6(4):e240300.
25. Cerdá-Alberich L, Solana J, Mallol P, Ribas G, García-Junco M, Alberich-Bayarri A, Marti-Bonmati L. MAIC–10 brief quality checklist for publications using artificial intelligence and medical images. Insights Imaging. 2023;14(1):11.
26. Kline TL, Kitamura F, Pan I, Korchi AM, Tenenholtz N, Moy L, Gichoya JW, Santos I, Blumer S, Hwang MY, Git KA. Best practices and scoring system on reviewing AI based medical imaging papers: part 1 classification. arXiv preprint arXiv:2202.01863. 2022.
27. Nankivell P, Williams H, Matthews P, Suortamo S, Snead D, McConkey C, Mehanna H. The binary oral dysplasia grading system: validity testing and suggested improvement. Oral Surg Oral Med Oral Pathol Oral Radiol. 2013;115(1):87–94.
28. Ribeiro-de-Assis MCF, Soares JP, de Lima LM, de Barros LAP, Grão-Velloso TR, Krohling RA, Camisasca DR. NDB-UFES: An oral cancer and leukoplakia dataset composed of histopathological images and patient data. Data Brief. 2023;48:109128.
29. Rahman TY, Mahanta LB, Chakraborti C, Das AK, Sarma JD. Textural pattern classification for oral squamous cell carcinoma. J Microsc. 2018;269(1):85–93.
30. Kandel I, Castelli M, Manzoni L. Brightness as an augmentation technique for image classification. Emerging Science Journal. 2022;6(4):881–92.
31. Wang J, Perez L. The effectiveness of data augmentation in image classification using deep learning. Convolutional Neural Networks Vis Recognit. 2017;2017(11):1–8.
32. Cossio M. Augmenting medical imaging: a comprehensive catalogue of 65 techniques for enhanced data analysis. arXiv preprint arXiv:2303.01178. 2023.
33. Sakamoto K, Ikeda T. Deep-learning application for identifying histological features of epithelial dysplasia of tongue. Journal of Oral and Maxillofacial Surgery, Medicine, and Pathology. 2022;34(4):514–22.
34. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022. pp. 11976–11986.
35. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J. An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020.
36. Peng J, Xu Z, Dan H, Li J, Wang J, Luo X, Xu H, Zeng X, Chen Q. Oral epithelial dysplasia detection and grading in oral leukoplakia using deep learning. BMC Oral Health. 2024;24(1):434.
37. Liu Y, Bilodeau E, Pollack B, Batmanghelich K. Automated detection of premalignant oral lesions on whole slide images using convolutional neural networks. Oral Oncol. 2022;134:106109.
38. Baik J, Ye Q, Zhang L, Poh C, Rosin M, MacAulay C, Guillaud M. Automated classification of oral premalignant lesions using image cytometry and Random Forests-based algorithms. Cell Oncol (Dordr). 2014;37(3):193–202.
39. Gupta RK, Kaur M, Manhas J. Cellular level based deep learning framework for early detection of dysplasia in oral squamous epithelium. InProceedings of ICRIC 2019: recent innovations in computing. Cham: Springer International Publishing. 2019. pp. 137–149.
40. Suzuki K. Overview of deep learning in medical imaging. Radiol Phys Technol. 2017;10(3):257–73.
41. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. J Pathol Inform. 2016;7:29.
42. Pontalba JT, Gwynne-Timothy T, David E, Jakate K, Androutsos D, Khademi A. Assessing the Impact of Color Normalization in Convolutional Neural Network-Based Nuclei Segmentation Frameworks. Front Bioeng Biotechnol. 2019;7:300.
43. Ghofrani A, Toroghi RM, Tabatabaie SM: Catiloc: Camera Image Transformer for Indoor Localization. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): 2021; 2021: 1450–1454.
44. Parvaiz A, Khalid MA, Zafar R, Ameer H, Ali M, Fraz MM. Vision Transformers in medical computer vision—A contemplative retrospection. Eng Appl Artif Intell. 2023;122:106126.
45. Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, Fu H. Transformers in medical imaging: A survey. Med Image Anal. 2023;88:102802.
46. Matsoukas C, Haslum JF, Söderberg M, Smith K. Is it time to replace cnns with transformers for medical images?. arXiv preprint arXiv:2108.09038. 2021.
47. S Band S, Yarahmadi A, Hsu C-C, Biyari M, Sookhak M, Ameri R, Dehzangi I, Chronopoulos AT, Liang HW. Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. Informatics in Medicine Unlocked. 2023;40:101286.

Publisher's Note