Computers in Biology and Medicine 189 (2025) 109913

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Automated classification of oral cancer lesions: Vision transformers vs radiomics

Eva Chilet-Martos [a], Joan Vila-Francés [a], José V. Bagan [b], Yolanda Vives-Gilabert [a,*]

[a] IDAL, Electronic Engineering Department, ETSE-UV, University of Valencia, Avgda. Universitat s/n, Burjassot, València, 46100, Spain
[b] Head Service Stomatology and Maxillofacial Surgery, University General Hospital, University of Valencia, Valencia, Spain

ARTICLE INFO

Keywords:
Oral cancer
Transformer-based object detection
Transformer-based segmentation
Vision transformers
Radiomics

ABSTRACT AND KEY WORDS

Background and objective: Early diagnosis is paramount in the effective management of oral cancer, offering numerous benefits including improved treatment outcomes, reduced morbidity and mortality, preservation of function and appearance, cost-effectiveness, and enhanced quality of life for patients. Transformer-based models, increasingly used in medical image analysis, are the focus of our study. We aim to compare a vision transformer (ViT) classification method with a fully automated radiomics approach. This involves using object detection and segmentation algorithms to effectively classify oral lesions in both cancer and control cases. A combined approach is also presented.

Methods: The analysis included 322 patients with oral lesions, comprising 120 cancer cases and 202 controls, with standard JPG images. Pretrained transformer-based algorithms like DEtection TRansformer (DETR), Segment Anything (SAM), and Vision Transformers (ViT) were used to explore different pipelines for lesion classification. For the ViT approach, images were inputted in three configurations: the entire image, a bounding box around the lesion, and a lesion delineation. The radiomics approach involved pipelines with bounding boxes and lesion segmentations. Additionally, a ViT-Radiomics combined approach was proposed, using ViT attention maps as radiomics masks. To validate the models, a five-fold cross validation was used.

Results: The combined ViT-radiomics model demonstrated superior performance for small training sets, achieving specificity = 0.97 ± 0.04, sensitivity = 0.96 ± 0.05, and accuracy = 0.97 ± 0.02 for the 100 % of the training set. When analyzed independently, the ViT approach using the entire image achieved the best results with specificity = 0.99 ± 0.02, sensitivity = 0.96 ± 0.05, and accuracy = 0.96 ± 0.02. Following closely was the pipeline using automatically obtained segmentations, while the one with bounding boxes had the least favourable outcomes. In the radiomics approach, the most effective classifier used the attention masks from the ViT classifier (derived from the entire image), achieving specificity = 0.97 ± 0.05, sensitivity = 0.95 ± 0.08, and accuracy = 0.94 ± 0.03. Manual segmentations yielded the best results for both approaches, indicating potential for performance enhancement through improved lesion segmentation.

Conclusions: The ViT classification surpassed the radiomics-based approach yet combining ViT with radiomics yielded similar results. However, the attention maps generated by ViT tend to associate oral lesions in cancer patients with regions distant from the lesions in control patients. For tasks requiring the examination and comparison of features within cancer and control oral lesions, utilizing the radiomics approach with an automatic lesion segmentation algorithm is recommended.

1. Introduction

Oral cancer is a significant health concern worldwide, with its incidence varying across different populations. According to the World Health Organization (WHO), there were approximately 354,864 new cases of oral cancer reported globally in 2020 [1]. The incidence rates vary geographically, with higher rates observed in regions with prevalent risk factors such as tobacco use, alcohol consumption, and betel quid chewing [2]. In the United States, oral cancer accounts for approximately 3 % of all cancer cases, with an estimated 54,010 new cases and 10,850 deaths expected in 2021 [3]. The incidence of oral cancer has been increasing in certain demographic groups, including

* Corresponding author. IDAL, Universitat de València, Avinguda de la Universitat s/n, Burjassot 46100, Spain.
E-mail address: yolanda.vives@uv.es (Y. Vives-Gilabert).

https://doi.org/10.1016/j.compbiomed.2025.109913
Received 16 April 2024; Received in revised form 9 January 2025; Accepted 21 February 2025
Available online 27 February 2025

young adults and non-smokers [4].

Early diagnosis of oral cancer is crucial for improving patient outcomes and reducing mortality rates. The prognosis of oral cancer depends on various factors, including the stage of the disease at diagnosis, tumour size, location, histological grade, and the presence of metastasis [5–8]. Early detection of oral cancer through routine screening and diagnostic examinations can lead to timely intervention and improved outcomes. Several screening methods, including visual examination, toluidine blue staining, and adjunctive techniques such as autofluorescence imaging and brush biopsy, have been employed for early detection of oral cancer [9]. Additionally, advancements in molecular biomarkers and imaging technologies present promising opportunities for early diagnosis and risk stratification of oral cancer patients. Although many cases of oral squamous cell carcinoma (OSCC) are still diagnosed at advanced stages—resulting in prognosis and survival rates that have remained largely unchanged for the past 25 years—early diagnosis is essential. Despite progress in surgical and oncological techniques, accurately distinguishing precancerous lesions from oral cancers is critical for improving survival outcomes.

Tissue biopsy remains the gold standard for diagnosing oral cancer, as with other cancers; however, prior to performing a biopsy, it is highly beneficial to identify suspicious areas using available screening methods. In addition to the basic inspection and palpation of oral lesions, several screening techniques have been developed. These include the application of toluidine blue to selectively stain areas of dysplasia or malignancy [10], autofluorescence, which uses specific wavelengths of light to detect tissue changes and highlight abnormal areas [11], chemiluminescence [12] and brush biopsy-based cytology [13]. Despite their utility, these methods are often limited by high false positive and false negative rates, necessitating the exploration of alternative approaches such as deep learning.

Artificial Intelligence (AI) and particularly deep learning offers transformative potential in overcoming the limitations of traditional screening methods by analyzing large datasets and complex patterns. This improves diagnostic accuracy by minimizing both false positives and false negatives. Furthermore, deep learning can integrate diverse patient data, such as genetic and lifestyle factors, enabling personalized screening that moves beyond the "one-size-fits-all" approach. It also fosters standardized diagnoses by minimizing operator dependency, ensuring consistent and accurate interpretations. Importantly, AI models can process non-invasive data with remarkable precision, reducing the need for painful or risky procedures. Their scalability and cost-efficiency further make these models viable in low-resource settings. While deep learning cannot entirely replace traditional methods, it serves as a valuable complement, enhancing screening accuracy and efficiency, and paving the way for faster, more accurate, and personalized diagnostics.

AI has witnessed significant growth in recent years. This progress has been fuelled by the availability of vast datasets and a substantial increase in computing power, particularly through graphical processing units (GPUs). In the field of oral oncology, two predominant classification pipelines are commonly employed: the radiomics-based approach [14] and the more recently embraced deep learning approach [15–18].

The radiomics approach revolves around extracting a multitude of quantitative features from medical images, which can include computed tomography, magnetic resonance images, or simple photographic JPG images. These extracted features serve as inputs for ML algorithms within decision support systems [19]. Statistical, shape and texture features are components of radiomics-based approaches in medical imaging and offer valuable insights into spatial patterns and variations in tissue or tumour structures. Texture features in particular play a crucial role in capturing subtle nuances not immediately discernible to the human eye, proving instrumental in diagnosis, prognosis, and treatment planning across various medical conditions, including oral oncology [17].

In the domain of the deep learning approach, early papers in the field harnessed the power of deep convolutional neural networks (CNNs). This architecture gained prominence following the victory of AlexNet [20] in the ImageNet [21] image classification competition in 2012 and in 2015 the U-net architecture emerged as one of the most efficient CNN architectures for image segmentation [22]. Numerous publications have leveraged variants of CNN algorithms for their studies [17,23–25].

Transformers emerged in 2017 as efficient models for natural language processing (NLP) with the groundbreaking claim "attention is all you need" [26]. Drawing inspiration from the success in NLP, Dosovitskiy et al. [27] introduced the vision transformer (ViT) by framing image classification as a sequence prediction task, converting images into a series of patches with positional encoding to retain spatial information. This approach enables the model to capture long-term dependencies within the input image. In the field of medical image analysis, the ViT architecture has become the state of the art for many disease diagnosis, finding applications in image detection [28], segmentation [29] and classification tasks [30].

In the specific domain of oral oncology, Islam et al. [31] compared two popular CNN models, VGG19 and MobileNet, with the transformer-based DeIT model to discriminate between benign and malignant oral lesions. In this study, CNN-based models slightly outperformed DeIT. On the contrary, Singha Deo et al. [32] asserted the supremacy of vision transformers in the classification of oral cancer using histopathological images. In the realm of medical image segmentation, MedSAM [33] has emerged as a recent development aimed at customizing the segment anything model [29]. on a large-scale medical image dataset with 1,570,263 image-mask pairs, covering 10 imaging modalities and over 30 cancer types.

In an effort to compare radiomic-based machine learning approaches with deep learning methods, Aubreville et al. introduced a deep learning-based classification approach for cancerous tissue using laser endomicroscopy images, demonstrating superior performance of DL compared to classical machine learning textural feature-based classifiers [34]. These results are in line with the review presented by Aydin Demircioğlu [35].

While deep-learning algorithms bring considerable advantages in terms of process automation and task performance, radiomics offers a higher level of interpretability, an asset in the biomedical field. This study aims to harness machine learning-based classification, integrating state-of-the-art vision transformer classifiers, to enhance the automation and performance of radiomics. The objective is to compare the transformer-based classification approach with a fully automated radiomics process to establish an optimal pipeline for maximizing the accuracy of classifying oral lesions as cancerous or non-cancerous. To achieve this, advanced transformer-based algorithms for detection [28], segmentation [29], and classification [27] were employed for the precise identification, delineation, and classification of oral lesions with the aim of facilitating the early diagnosis of OSCCs and thereby improving prognosis and survival rates.

2. Materials and methods

2.1. Dataset

During the period 1997–2020, we collected a sample comprising a total of 322 patients, all recruited in the Stomatology and Maxillofacial Surgery Service of the General University Hospital of Valencia. Of the 322 patients, 120 presented oral squamous cell carcinoma diagnosed, according to Warnakulasuriya et al. [36] criteria. The rest of the patients (202) presented non-cancerous oral lesions. The oral cancer group was composed of 75 men (62.5 %) and 45 women (37.5 %). The tongue had the highest frequency in 43 cases (35.9 %), followed by the gingiva in 40 patients (33.3 %). The ulcerated form predominated with 76 cases (63.3 %). We had 65 (54.2 %) in the initial stages of cancer, and the remaining 55 (45.8 %) were in advanced stages. The mean age of this group with cancer was 70.12, with a SD of 10.63.

The control group of patients without cancer consisted of 202

patients with oral leukoplakia. Of them, 92 (45.5 %) were men and 110 (54.5 %) were women. The mean age of this group was 57.88, with a SD of 12.82.

The dataset was comprised of JPG images taken with an ordinary digital camera (Nikon Digital Camera D5600, objective: Nikon 60 mm, flash Viltrox Macro Ring Life JY670n). The images had the following characteristics: width: 25.33 cm, height: 16.93 cm, resolution: 300 pixels/inch, photo dimensions 2992 px x 2000 px. The images were manually segmented as regions of interest (ROIs) by a clinical expert with 30 years of experience (Dr. Bagan, co-author) using ImageJ software [37]. These segmentations will serve as the gold standard throughout this manuscript. The gold standard bounding boxes (BB) were derived directly from the manual segmentations.

A five-fold cross-validation strategy was applied to ensure the robustness of the results while considering the constraints of computational resources.

To address the challenges posed by limited and imbalanced data, a robust technique known as data augmentation was employed. Data augmentation involves applying diverse transformations to the original data, artificially expanding the training dataset. This technique proves particularly beneficial in enhancing the performance and generalization capability of machine learning models. Common data augmentation methods encompass image rotation, flipping, cropping, resizing, addition of noise, adjustments to brightness and contrast, and geometric transformations.

Given the constraints of a reduced and imbalanced dataset, a comprehensive $\times 10$ (in controls) and $\times 16$ (in cancer) data augmentation strategy was implemented for the training dataset. This augmentation process included horizontal and vertical flips ($p = 0.5$), brightness and contrast adjustments ($p = 0.3$), and the addition of Gaussian noise ($p = 0.3$), all implemented using the Albumentations package. Additionally, Floyd-Steinberg dithering was applied using the Pillow library. Since the number of cancer images was divisible by 5, all five folds contained 24 cases, becoming for each iteration 96 cases in the training set (augmented to 1536) and 24 in the test set. However, as the number of control images was not divisible by 5, the number of control images varied among folds (40 or 41 cases per fold), becoming for each iteration 161 or 162 cases in the training set (augmented to 1610 or 1620) and 40 or 41 in the test set.

2.2. Methodology

As outlined in the introduction, this study aims to compare the effectiveness of a ViT classifier with a fully automated radiomics approach in classifying oral lesions. While ViT classification offers a direct analysis from the original images, we hypothesize that its performance could be enhanced by pre-selecting the oral lesions before classification. Conversely, the radiomics process needs the prior detection and/or segmentation of the lesion, requiring the input of a specific mask. As a result, separate pipelines were employed for each approach, sometimes requiring the detection of oral lesions through bounding boxes (referred to as detection throughout the manuscript) or their delineation using segmentation algorithms (referred to as segmentation throughout the manuscript). Additionally, a combined ViT-Radiomics approach was also introduced to leverage the strengths of both methods for enhanced classification.

Transfer learning is a machine learning technique where a model developed for one task serves as the starting point for a model on a second related task. It leverages knowledge gained from solving one problem and applies it to another, potentially speeding up model development and improving performance, especially when data for the target task is limited. Fine-tuning, one approach to transfer learning, involves retaining most of the layers of a pretrained model and adapting them using target task data. This allows the model to adapt to task-specific patterns while retaining knowledge from the source task [38]. In this study, all transformer-based algorithms were fine-tuned with gold standard masks, including those for detection, segmentation, and ViT classification.

The algorithms were executed on an Intel Core i9 machine with 64 GB of RAM and an NVIDIA RTX 3060 GPU. Implementation was done using Python (version 3.10.9), PyTorch (version 2.0.0), and torchvision (version 0.15.1).

Detailed descriptions of the algorithms and pipelines used in this study are provided in subsequent sections. We first introduce the algorithms for oral lesion detection and segmentation, followed by descriptions of the different pipelines for the ViT and radiomics approaches and finally, we present the ViT-Radiomics combined approach.

The code corresponding to this study is available in the following GitHub repository: https://github.com/evachi27/Automated_Oral_Cancer_Classification.

2.2.1. Lesion detection and segmentation with transformer-based models

In this study, oral lesion detection was performed with the Detection Transformer (DETR) algorithm [28] accessed through the Hugging Face library. DETR was pioneering in applying Transformers to computer vision tasks. It combines a convolutional neural network (CNN) backbone with a transformer architecture. Initially, the model utilizes a standard CNN backbone to generate a 2D representation of the input image. This representation is then flattened and enriched with positional encoding. Subsequently, the data is fed into a transformer encoder. In the following step, a transformer decoder is introduced, which takes a fixed number of positional embeddings called “object queries” as input. The decoder also attends to the output of the encoder. Each output embedding from the decoder passes through a shared feedforward network (FFN), responsible for making predictions, including either a detection with class and bounding box information or a “no object” class, as in this study.

The hyperparameters for DETR were configured with a batch size of 8, trained for 100 epochs and a learning rate of $1e^{-4}$. The metric employed to assess the performance of this algorithm was the IoU (Intersection over Union (Eq. (1)).

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \quad \text{Eq. 1}$$

The segmentation of oral lesions was performed using the Segment Anything Model (SAM), a pioneering model architecture developed by Meta research group [29]. SAM is designed for generating segmentation masks for various objects in an image. Trained on 1 billion masks from 11 million images, SAM allows automatic segmentation of everything in an image or segmentation based on manual prompts. The model architecture includes an image encoder using a Mean Absolute Error (MAE) pre-trained Vision Transformer (ViT) [39], a prompt encoder capable of handling sparse (points, bounding boxes, text) or dense (masks) inputs, and an efficient mask decoder that maps image and prompt embeddings to a mask. For this study, a sparse prompt encoder using bounding boxes was employed.

SAM, despite its impressive performance in various general tasks, may not be as accurate as specialized algorithms designed explicitly for medical image segmentation when applied directly to medical images without re-training or fine-tuning [40,41]. To address this issue, the model was fine-tuned using the gold standard segmentations of the oral lesions.

The hyperparameters were configured with a batch size of 1, trained for 100 epochs using the dice coefficient as loss function and a learning rate of $1e^{-5}$. The performance of SAM was evaluated with the Dice coefficient (Eq. (2)).

$$Dice = \frac{2 \times Area\ of\ Overlap}{Total\ Area} \quad \text{Eq. 2}$$

2.2.2. Vision transformer-based classification

The Vision Transformer (ViT) classification architecture revolutionized image classification by applying the transformer architecture, originally designed for natural language processing, to visual data [27]. ViT starts by dividing the input image into non-overlapping patches, treating each patch as a token. These patch tokens are then linearly embedded and combined with positional encodings to preserve spatial information. The heart of the architecture lies in the transformer encoder, a stack of layers incorporating self-attention mechanisms and feedforward sub-layers. This enables ViT to capture intricate patterns and dependencies across the entire image, both locally and globally. The final classification is performed using a classification head on top of the transformer encoder, predicting the class labels.

The self-attention mechanism in ViTs is a key component that enables the model to capture contextual information from different regions of the input image. Self-attention masks are matrices used to control which regions can attend to each other in the self-attention mechanism of transformer-based models. These masks determine the patterns of attention across the input regions and help the model focus on relevant information while ignoring irrelevant or distracting regions. These masks are implemented as binary matrices of zeros and ones, where a value of 1 indicates allowed attention and a value of 0 indicates forbidden attention.

In this study, the vit-base-patch16-224-in21k model, accessible via the Hugging Face library was employed. This model aligns with the original architecture introduced by Dosovitskiy et al. [27]. Initially pre-trained on the extensive ImageNet-21k dataset (comprising 14 million images with 21,843 classes) at a resolution of 224x224, it underwent further fine-tuning on the ImageNet 2012 dataset (1 million images, 1000 classes) at the same resolution. To tailor it to our specific application, the model underwent additional fine-tuning using our training dataset. Prior to input into the ViT model, images were resized to a lower resolution (224x224), rescaled, and normalized across the RGB channels with mean (0.5, 0.5, 0.5) and standard deviation (0.5, 0.5, 0.5). ViT was configured with a batch size of 2, 10 epochs and a learning rate of 5×10^{-5}.

The vision transformer-based classification in this study involved three distinct pipelines (Fig. 1):

1. Direct Classification from original Image, which entailed classifying oral lesions directly from the original image.
2. Detection and Subsequent Classification, where oral lesions were detected through bounding boxes, and the identified small areas were subsequently classified.
3. Detection and Segmentation before Classification, which involved the prior detection and the subsequently delineation of the oral lesions, followed by classification using the segmented areas.

2.2.3. Radiomics-based classification

Radiomics, a specialized field in medical imaging, involves extracting and analysing quantitative features from images to provide detailed insights into tissue or pathology. The workflow begins with image acquisition, followed by lesion segmentation and feature extraction, encompassing first-order, shape-related, texture-related, and higher-order filtered features. Feature selection algorithms are then used to identify significant features, with subsequent application of machine learning classification algorithms (Fig. 2).

In this classification approach, two distinct models were constructed based on the ROIs used to train the model: 1) using gold standard bounding boxes of the lesions and 2) leveraging gold standard segmentation masks of the lesions.

To assess the performance of these models, efforts were made to automate the radiomics process as much as possible, facilitating a comparison between gold standard masks and automatically obtained masks. Two pipelines were employed to segment the lesion mask for radiomics feature calculation. The first pipeline used DETR to generate a bounding box around the oral lesion (pipeline A2). The second pipeline combined detection (using DETR) and segmentation (via SAM) of the lesions (pipeline B2) (Fig. 3).

For comparison purposes, a pre-trained Convolutional Neural Network (CNN) was applied, specifically the ResNet50 architecture [42]. ResNet50 is widely recognized for its deep architecture and residual learning capabilities, making it a strong candidate for transfer learning tasks.

2.2.3.1. Radiomic features. As already introduced before, the radiomics process implies the calculation of radiomic features obtained from the

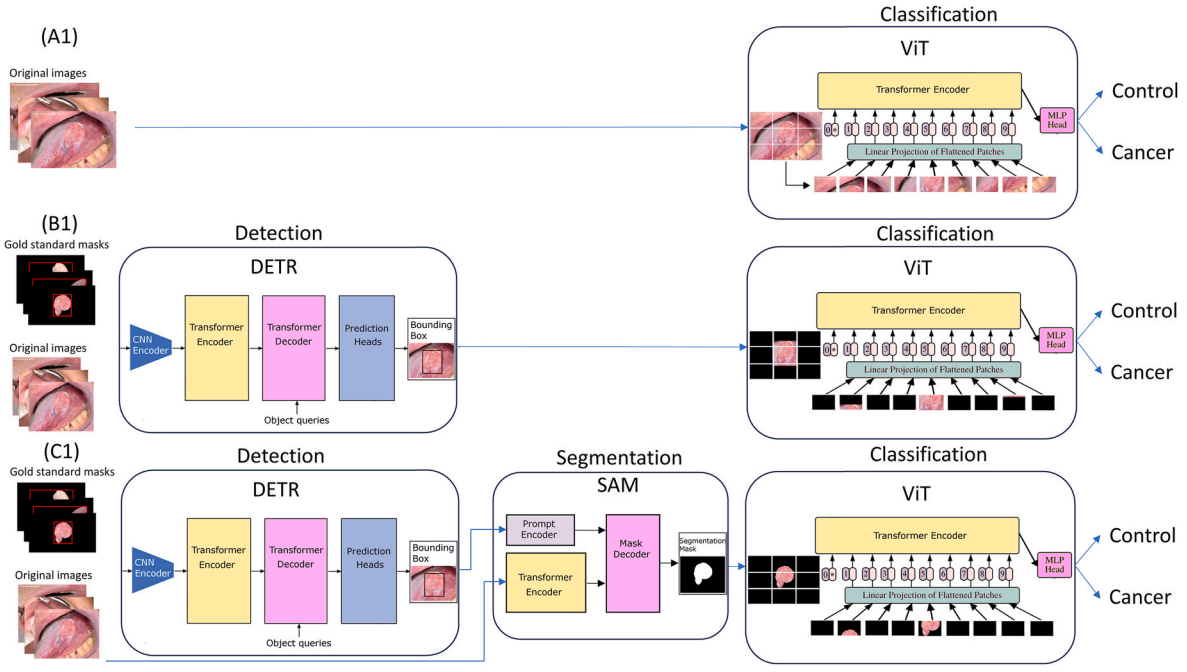

Fig. 1. Training pipelines for the vision-based classification procedure using (A1) directly the original images; (B1) a detection of the lesion before the classification and (C1) a detection and a segmentation algorithm before the classification.

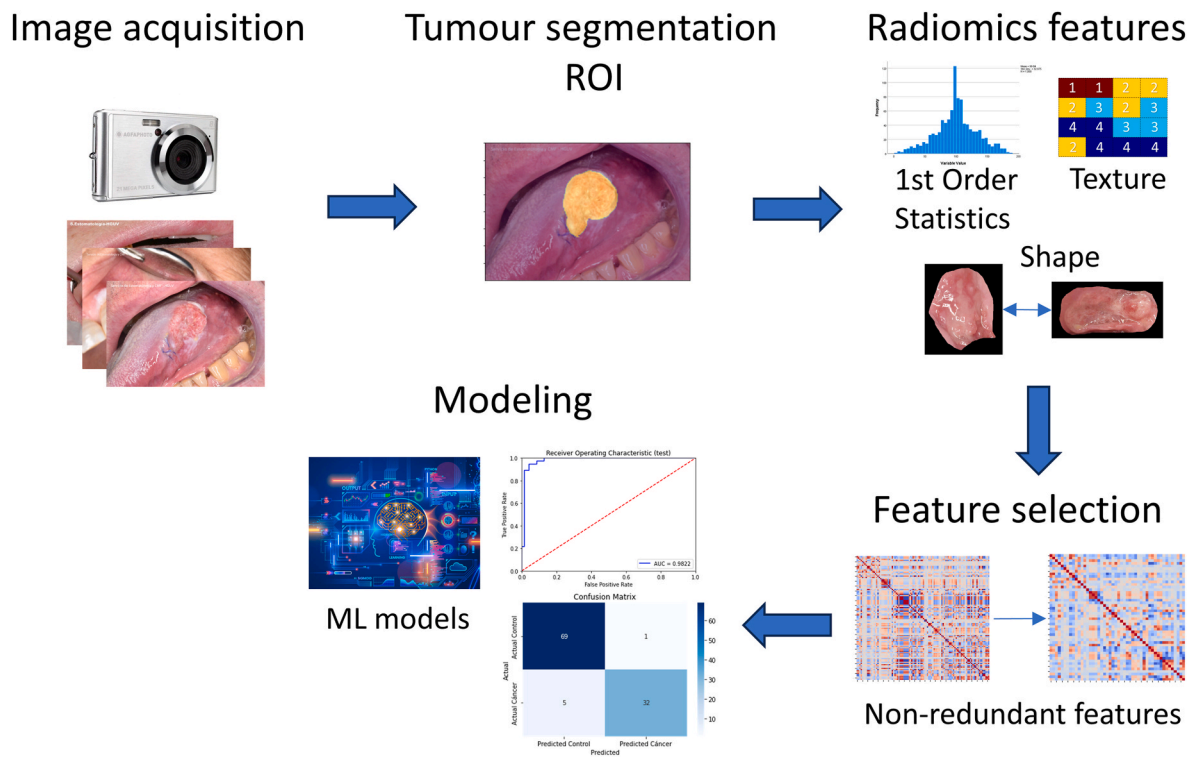

Fig. 2. Radiomics process.

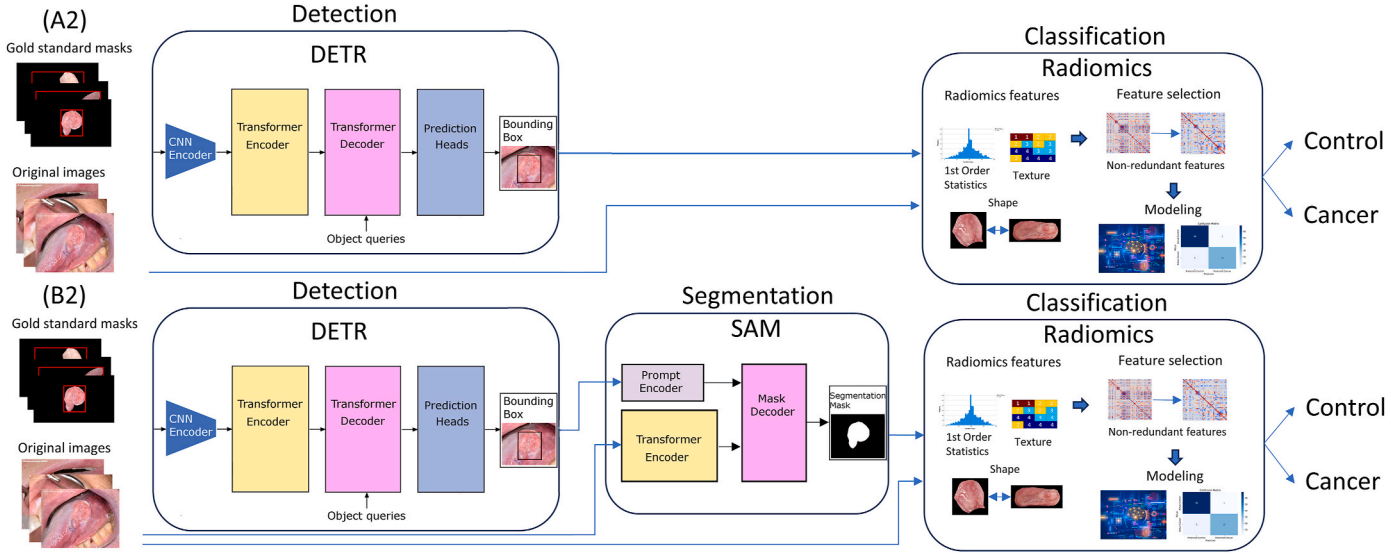

Fig. 3. Training pipelines for obtaining the masks automatically for radiomics classification procedure using (A2) DETR as detection algorithm to obtain a bounding box of the oral lesion and (B2) a detection (DETR) and a segmentation (SAM) algorithm before the classification.

masks. In this study, we extracted a total of 102 radiomic features obtained with PyRadiomics library [43], divided in different categories (all the features are listed in the Supplementary material):

- **First Order Statistics**: they describe the distribution of voxel intensities within the image region defined by the mask.
- **Shape-based features**: features obtained directly from the mask, independent from the gray level intensity distribution in the ROI.
- **Gray Level Co-occurrence Matrix (GLCM)**: this matrix describes the second-order joint probability function of an image region constrained by the mask.
- **Gray Level Size Zone Matrix (GLSZM)**: this matrix quantifies the gray level zones in an image and is defined as the number of connected voxels that share the same gray level intensity.
- **Gray Level Run Length Matrix (GLRLM)**: this matrix quantifies gray level runs and is defined as the length in number of consecutive pixels that have the same gray level value.
- **Neighbouring Gray Tone Difference Matrix (NGTDM)**: this matrix quantifies the difference between a gray value and the average gray value of its neighbours within distance δ.
- **Gray Level Dependence Matrix (GLDM)**: this matrix quantifies gray level dependences in an image and is defined as the number of

connected voxels within distance δ *that are dependent on the center voxel.*

A detailed description of how these features are calculated can be found on the documentation website of PyRadiomics (https://pyradiomics.readthedocs.io). The default values were used for all parameters. All the features were standardized using Standard Scaler function of Scikit-learn package [44].

Out of the entire set of features, the high correlated ones ($r \geq 90$ %) were excluded (54 features), resulting in a total of 48 features. Most of the excluded features were derived from the GLCM matrix (19 out of 24) and from the GLDM matrix (12 out of 14). The features that were included and excluded from the analysis are listed in the Supplementary material.

2.2.3.2. Machine learning classifiers. Three different popular ML algorithms were tested with the remaining 47 selected features: Support Vector Machine [45], Random Forest [46] and a Gradient Boosting [47] using the implementation of Scikit-learn package [44]. A Grid Search strategy was employed in all the cases to optimize the most relevant hyperparameters of the classifiers, while default values were retained for the remaining ones.

2.2.3.3. Most relevant features. Permutation feature importance serves as a model inspection method applicable to any trained estimator and tries to gauge the impact of individual features on the model's performance [48]. It quantifies feature importance by measuring the reduction in the model's score when the values of a single feature are randomly shuffled. This shuffling disrupts the connection between the feature and the target, allowing the observed drop in the model score to reveal the extent to which the model relies on that specific feature. Notably, this approach is model agnostic, offering flexibility, and can be iteratively computed with various permutations of the feature. Its validation performance, measured via the r^2 score, is significantly larger than the chance level. In this study, feature importance was applied with Scikit-learn library to the three tested models (training, validation, and test datasets) to study the importance of the different features (n_repeats = 30).

2.2.4. ViT-radiomics combined approach

As previously noted, one of the primary advantages of ViT is its ability to function without the need for prior detection or segmentation of oral lesions, a requirement in the radiomics approach. In response, a ViT-Radiomics combined approach was proposed (Fig. 4). This model initiates with ViT classification on the original images (pipeline A1), using the resulting attention maps as masks to extract radiomics features for subsequent machine learning classification. Each method (ViT and radiomics) generates probabilities for classifying samples as either belonging to the control or cancer group, from which weighted mean values were calculated. Given that the ViT pipeline produced better results than radiomics, its probabilities were assigned a weight of 60 %, while those from radiomics were given a weight of 40 %. These combined probabilities were then binarized using a threshold of 0.5.

To evaluate the robustness of the three methods (ViT, radiomics using ViT attention maps, and the combined ViT-Radiomics approach) across different training data sizes, the training dataset was progressively reduced, with random samples taken from 20 % to 100 % of the dataset. The test dataset remained unchanged in all cases.

3. Results

3.1. Lesions detection and segmentation with transformer-based models

Table 1 presents the performance results obtained from the different algorithms, including DETR, SAM (using gold standard bounding boxes) and the combination of DETR + SAM. It's noteworthy that the DETR algorithm demonstrated a higher ease of detection for the cancer group, with the IoU of the cancer group averaging around 20 % higher than that of the control group across all datasets.

In evaluating the SAM segmentation performance in test, notably

Table 1
Performance of the detection and segmentation algorithms.

	DETR (IOU)	SAM (DICE)	DETR + SAM (DICE)
TRAIN	0.67 ± 0.02	0.57 ± 0.01	0.52 ± 0.02
CONTROL	0.58 ± 0.02	0.49 ± 0.01	0.43 ± 0.02
CANCER	0.75 ± 0.02	0.66 ± 0.01	0.61 ± 0.02
TEST	0.53 ± 0.02	0.55 ± 0.02	0.45 ± 0.03
CONTROL	0.46 ± 0.03	0.47 ± 0.02	0.37 ± 0.030
CANCER	0.64 ± 0.04	0.67 ± 0.03	0.59 ± 0.04

Results in mean ± std of the mean IoU and Dice of the five iterations.

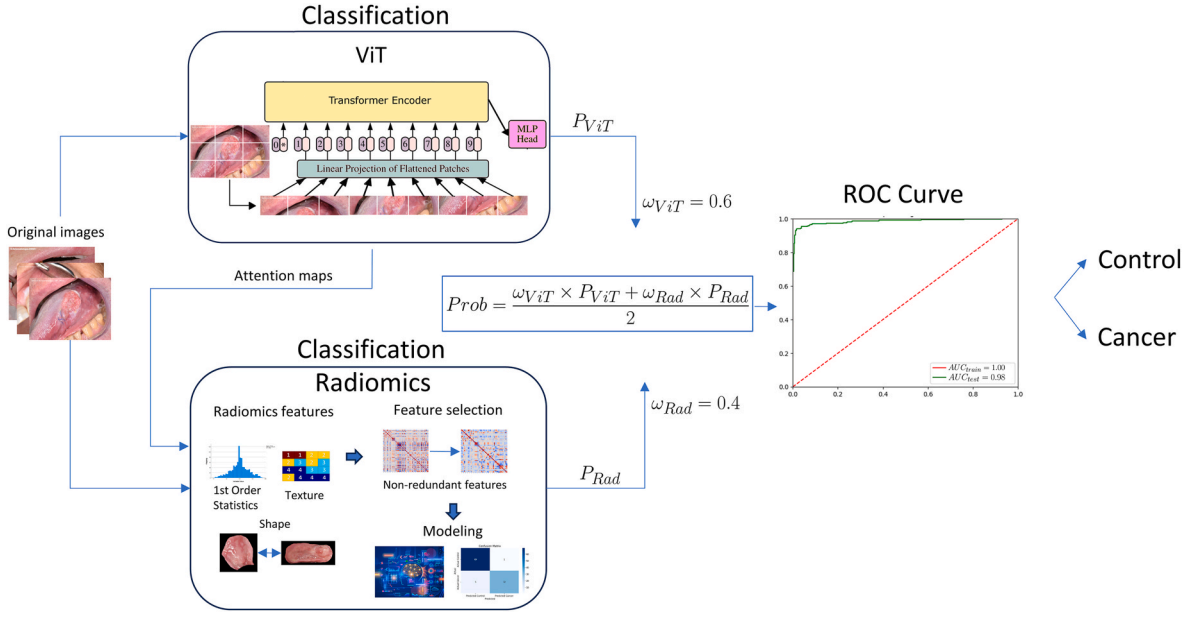

Fig. 4. Integrating ViT-Radiomics for enhanced classification of oral cancer lesions.

high Dice coefficients were observed, attributed to the assistance provided by gold standard bounding boxes. While SAM segmentation exhibited better results for cancer cases compared to controls, the disparity was less pronounced than in the case of DETR. However, the application of SAM segmentation following DETR detection led to a significant decrease in Dice coefficients. This decline was attributed to the accumulation of errors from both DETR and SAM algorithms. Fig. 5 illustrates this dependence, especially when DETR fails to detect a lesion (e.g. Patient 3). The interdependence of DETR and SAM errors is further evidenced in Fig. 6, where the IoU of DETR correlates with the Dice coefficient of SAM in the DETR + SAM segmentation ($r^2 = 0.34 \pm 0.04$ for train and $r^2 = 0.41 \pm 0.07$ for test considering the five iterations). Additionally, the cancer group consistently exhibited higher IoU and Dice coefficients than the control group in these evaluations.

3.2. Vision transformer classification

After lesion detection and segmentation, the ViT classifier was applied across three distinct pipelines, as illustrated in Fig. 1. To gauge the performance of the ViT model in isolation, free from the influence of accumulated errors from various pipelined algorithms, an evaluation was conducted using the gold standard ROIs. The ViT model with bounding boxes underwent testing with bounding boxes derived from gold standard segmentations. Similarly, the ViT model with segmentations was assessed using SAM algorithm segmentations derived from gold standard BB and gold standard manually segmented masks. For comparison purposes, a fine-tuned ResNet50 model was also applied to the data. The detailed results are shown in Table 2.

The ViT classification demonstrated exceptional accuracy when using original images without prior lesion detection or segmentation. This accuracy further increased to a mean value 0.97 when original images were masked with gold standard segmentations. However, in the pursuit of process automation, model performance decreased when applying automatic detection (pipeline B1) to 0.88. With the inclusion of detection and segmentation algorithms (pipeline C1), the accuracy showed a decrease to 0.86, with notable reduction in sensitivity to 0.71. The specificity, however, remained high. The results from the ResNet50 were generally significantly lower than those achieved with ViT.

Upon analyzing attention masks for all models and calculating Dice coefficients, it was observed that the ViT models predominantly focused on cancer lesions while showing less emphasis on control lesions, evidenced by the Dice coefficients of the control group obtained from the different pipelines (Table 3).

Specifically, attention maps derived from the cancer group in pipeline A1 exhibited higher Dice coefficients in test than those from any other pipeline, also coinciding with the highest classification accuracy, as indicated in Table 2. It's important to note that the precision of these results is significantly influenced by the transformer's patch size (16x16 in this case).

3.3. Radiomics classification

Within the radiomics procedure, the use of a mask is imperative to extract the different radiomic features. In this context, two distinct models were examined based on different types of masks: 1) trained with gold standard bounding boxes and 2) trained with gold standard segmentation masks. These two models were tested with different input masks, as shown in Table 4.

Regarding the radiomics classification, the three ML classifiers demonstrated similar metrics on the test dataset; however, SVM exhibited slightly superior performance. Table 4 presents the results obtained by the SVM classifier.

In the concluding phase, a comprehensive analysis of the most crucial features for each of the two models was undertaken.

For the bounding boxes model, some of the most significant features included Run Entropy, which had higher values in the cancer group compared to the control group, and Long Run High Gray Level Emphasis and the 10th percentile, both of which showed lower values in the cancer group than in the control group. In the mask-dependent models, key features were the Perimeter Surface Ratio and High Gray Level Run Emphasis, both of which were smaller in the cancer group, while the Major Axis Length exhibited the opposite trend, being larger in the cancer group.

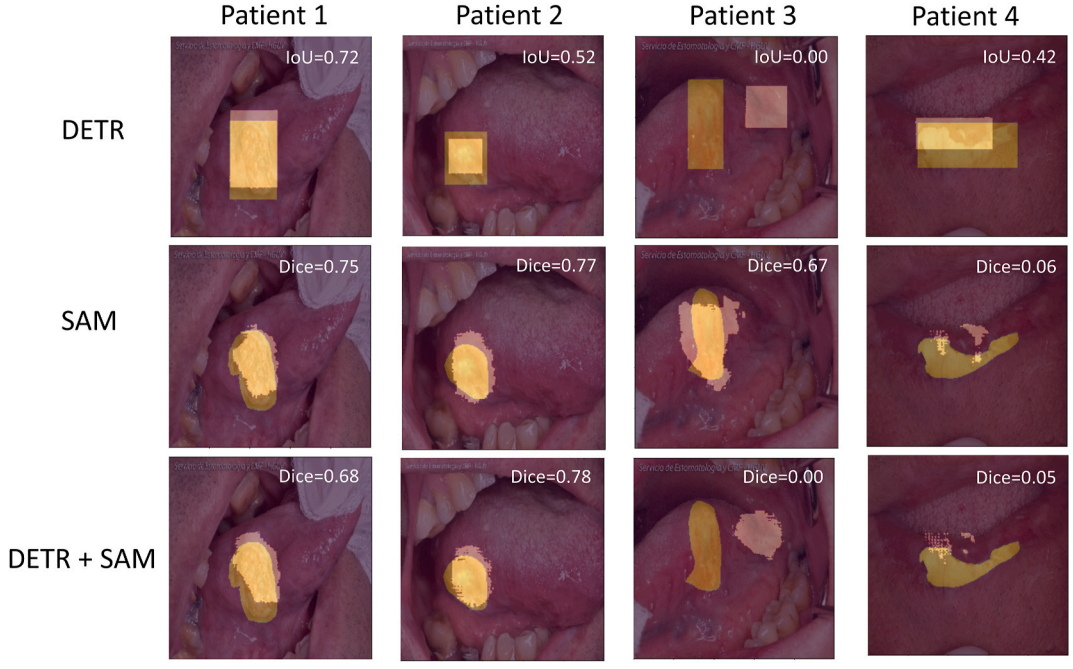

Fig. 5. Segmentation output (light pink) vs gold standard (in dark yellow) of DETR, SAM and DETR + SAM algorithms of four different patients (patients 1, 2, 3 are cancer, patient 4 is control). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

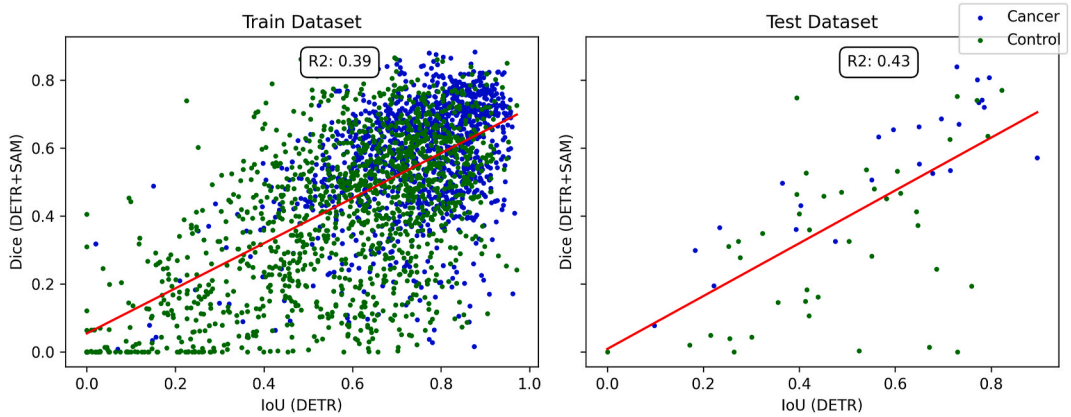

Fig. 6. Example of the relationship between IoU obtained by DETR and the Dice coefficients of the masks obtained by DETR + SAM for the two datasets and their corresponding r^2 value (fold = 2).

Table 2
Classification metrics for the different ViT pipelines.

Training Modality	Testing images	Specificity	Sensitivity	Accuracy	F1-score	AUC	AUC (ResNet50)
Whole image	**Original images (pipeline A1)**	0.99 ± 0.02	0.96 ± 0.05	0.96 ± 0.02	0.95 ± 0.03	0.96 ± 0.03	0.82 ± 0.05
Gold Standard Bounding Boxes	**Gold standard BB**	0.96 ± 0.04	0.99 ± 0.02	0.97 ± 0.04	0.96 ± 0.03	0.97 ± 0.03	0.77 ± 0.06
	DETR (pipeline B1)	0.82 ± 0.05	0.98 ± 0.02	0.88 ± 0.03	0.86 ± 0.03	0.90 ± 0.03	0.75 ± 0.04
Gold Standard Segmentation Masks	**Gold standard masks**	0.97 ± 0.02	**0.98 ± 0.02**	**0.98 ± 0.02**	**0.97 ± 0.02**	**0.98 ± 0.02**	**0.85 ± 0.03**
	SAM using gold standard BB	**0.99 ± 0.01**	0.72 ± 0.12	0.90 ± 0.05	0.83 ± 0.08	0.86 ± 0.05	0.75 ± 0.03
	DETR + SAM (pipeline C1)	0.95 ± 0.04	0.71 ± 0.11	0.86 ± 0.04	0.79 ± 0.07	0.83 ± 0.05	0.78 ± 0.04

Results in mean ± std of the five iterations.

Table 3
Dice/IoU of the mean attention maps for the ViT classifiers.

	PIPELINE A1 (DICE)	PIPELINE B1 (IOU)	PIPELINE C1 (DICE)
TRAIN	0.29 ± 0.03	0.31 ± 0.01	0.23 ± 0.01
CONTROL	0.04 ± 0.05	0.02 ± 0.02	0.05 ± 0.03
CANCER	0.55 ± 0.05	0.61 ± 0.02	0.43 ± 0.05
TEST	0.23 ± 0.02	0.23 ± 0.02	0.21 ± 0.02
CONTROL	0.04 ± 0.03	0.04 ± 0.02	0.07 ± 0.03
CANCER	0.57 ± 0.02	0.54 ± 0.05	0.44 ± 0.05

Results in mean ± std of the mean IoU and Dice of the five iterations.

3.4. ViT-radiomics combined approach

The combined approach involves the use of the attention maps derived from the ViT classification with original images as masks for the radiomic feature extraction. This approach provides an automatic alternative to obtain the ROIs, eliminating the need for manual delineation.

By applying the radiomics pipeline with the ViT attention maps, we achieved an accuracy of 0.94 ± 0.03, which is slightly lower than that of the ViT classification alone (0.96 ± 0.02). By combining both approaches as depicted in Fig. 4, we managed to enhance slightly the accuracy, F1-score and AUC of each method individually. The results are summarized in Table 5.

To examine the impact of training size on the results, we applied the combined approach using varying percentages of images from the training dataset. The corresponding results are shown in Fig. 7. The combined approach showed higher accuracy when the training set was small (20 % and 30 % of the total training images). However, from 40 % onwards, the accuracy of the ViT-Radiomics method was nearly equivalent to that of ViT alone. The radiomics procedure using attention

Table 4
Performance of the radiomics pipelines.

Training Modality	Input masks	Specificity	Sensitivity	Accuracy	F1-score	AUC
Gold Standard Bounding Boxes	**Gold standard BB**	0.89 ± 0.04	0.77 ± 0.06	0.85 ± 0.04	0.85 ± 0.04	0.91 ± 0.04
	DETR (pipeline A2)	0.83 ± 0.07	0.73 ± 0.07	0.80 ± 0.06	0.80 ± 0.06	0.86 ± 0.06
Gold Standard Segmentation Masks	**Gold standard masks**	**0.93 ± 0.05**	**0.81 ± 0.06**	**0.89 ± 0.05**	**0.89 ± 0.05**	**0.96 ± 0.02**
	SAM using gold standard BB	0.92 ± 0.04	0.74 ± 0.10	0.86 ± 0.05	0.87 ± 0.03	0.9 ± 0.04
	DETR + SAM (pipeline B2)	0.87 ± 0.02	0.72 ± 0.10	0.81 ± 0.05	0.82 ± 0.05	0.87 ± 0.04

Results in mean ± std of the five iterations.

Table 5
Performance of the best classifiers.

Classification Pipeline	Specificity	Sensitivity	Accuracy	F1-score	AUC
ViT with original images	**0.99 ± 0.02**	**0.96 ± 0.05**	0.96 ± 0.02	0.95 ± 0.03	0.96 ± 0.03
Radiomics with ViT attention maps	0.97 ± 0.05	0.95 ± 0.08	0.94 ± 0.03	0.94 ± 0.03	0.98 ± 0.01
ViT-Radiomics combined approach	0.97 ± 0.04	**0.96 ± 0.05**	**0.97 ± 0.02**	**0.97 ± 0.02**	**0.98 ± 0.02**

Results in mean ± std of the five iterations.

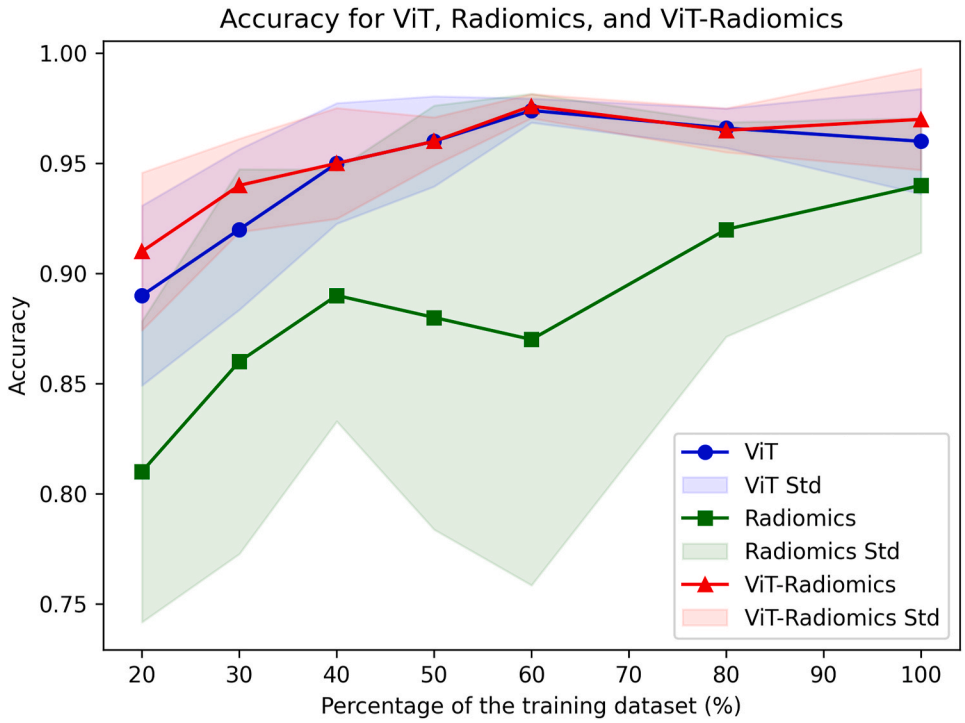

Fig. 7. Accuracy of the ViT, Radiomics and ViT-Radiomics approaches with different sizes of the original training dataset (mean ± std of the five iterations).

masks showed a lower performance and a high variability across folds.

4. Discussion

Oral lesion detection, segmentation and classification is critical to human life, as the incidence of oral cancer has increased considerably and is seriously threatening human health. These tasks are particularly challenging due to many factors such as indistinct boundaries, variable contrast, and colour differences. In the recent years, transformer-based models have dominated the field of natural language processing and have more recently been applied to the area of computer vision, outperforming tasks like object detection, segmentation, and classification, and in particular to the field of medical image analysis [30]. As far as we know, this is the first study comparing automatized transformer-based radiomics procedure with vision transformer classifiers in oral cancer. For each of these two approaches (transformer-based classification and radiomics-based classification), different pipelines were tested. These pipelines included the lesion detection and the segmentation using different transformer-based algorithms. A combined approach was also proposed.

4.1. Lesions detections and segmentations

For oral lesion object detection, we achieved promising results for the cancer group; however, the average IoU for the control group was approximately 20 % smaller than that of the cancer group, highlighting its dependency on the type of lesion. In a study by Ickler et al. [49], various transformer-based medical object detection algorithms, including DETR, Conditional DETR, and DINO DETR, were compared with a strong anchor-based one-stage detector, Retina U-Net. Interestingly, DINO DETR emerged as the best-performing model in three out of four datasets, offering the added advantage of being computationally less demanding than Retina U-net [49].

Concerning the segmentation of oral lesions, we opted for SAM instead of the U-Net architecture, which was used in a previous work of our group [17]. While the cancer group exhibited similar segmentation Dice coefficients to our previous work, the results of the control group notably decreased. It's worth noting that the datasets in both studies were not identical. Despite SAM demonstrated superior performance in natural image segmentation, it appears to deliver less favourable results in applications like healthcare, as also noted by Ji et al. [50]. Additionally, the study by Huang et al. [41] revealed that SAM exhibited remarkable performance in specific objects but proved to be unstable, imperfect, or even failed in certain situations, as we also noted and illustrated in Fig. 5 (patient 4). The results presented in Table 1, specifically SAM with gold standard bounding boxes versus DETR + SAM, highlight that SAM requires substantial prior human knowledge to achieve relatively accurate results in biomedical tasks, consistent with observations by Huang et al. [41]. The decrease in the segmentation performance inevitably impacts the classification performance as well.

4.2. Vision transformer and radiomics classification

In our study, the vision transformer approach demonstrated superior classification performance compared to the radiomics procedure, aligning with a recent systematic review conducted by Aydin Demircioğlu [35]. The review, which analyzed 1229 records, revealed that deep modelling outperformed generic radiomic modelling in 74 % of cases during internal validation. Additionally, fused models exhibited superior performance compared to both approaches individually in 72 % of cases, as we have also demonstrated with our ViT-Radiomics combined approach. The ViT classifier demonstrated exceptional performance with a 97 % mean accuracy when employing gold standard masks, which highlights the potential for further enhancing overall accuracy by improving lesion segmentation.

In a recent study conducted by Islam et al. on the classification of oral lesion images, various models, including convolutional neural networks (VGG19, MobileNet), and the transformer-based model Data-Efficient Image Transformer (DeIT) were employed [31]. The findings indicated that convolutional-based models achieved a remarkable accuracy rate of 100 %, while DeIT exhibited a slightly lower but still impressive accuracy of 98.73 %. The superior performance of convolutional-based models was attributed to their deeper architectures and utilization of pre-trained models. Despite a marginally lower accuracy, the

transformer-based model, DeIT, showcased notable generalization capabilities to unseen data. The study by Islam et al. emphasized that while convolutional networks may be better suited for medical problems, they demand higher computational resources, and their performance is more contingent on the quality and quantity of training data [31]. However, the fine-tuned ResNet50 model in our study demonstrated significantly lower performance compared to the ViT model (Table 2). Regarding the ViT results, we observed slightly lower performance metrics when using the original images compared to the findings of Islam et al. Several factors might contribute to this difference, including ViT's potential challenge in performing well with limited data. Moreover, DeIT, despite sharing the same architecture as ViT, incorporates an additional distillation token in the input, interacting with class and patch tokens through self-attention layers, potentially enhancing results [51]. Another possible factor could be variations in the homogeneity/heterogeneity of the two datasets.

In the context of the radiomics approach, using the average of the attention maps from the ViT classifier as ROIs yielded a classification closely aligned with that obtained by the ViT classifier alone (0.94 ± 0.03 and 0.96 ± 0.02, respectively). This approach offers the advantage of circumventing the need for manual segmentation, a laborious and resource-intensive task. However, a detailed analysis of the Dice coefficients in Table 3 revealed that the attention masks accurately cover oral lesions in the cancer group but not in the control group (Dice coefficients nearly 0). Comparatively, the model using segmentation masks outperformed the one employing bounding boxes.

Finally, our study demonstrates that integrating ViT and radiomics approaches enhances accuracy, particularly for smaller training sets (20 % and 30 % of the original size), while achieving comparable performance to ViT alone for larger datasets. However, careful handling of radiomic feature information remains essential in this context. This is because the comparison primarily involves cancer lesions with other parts of the oral cavity rather than with non-cancerous lesions, as evidenced by the Dice coefficients in Table 3. Additionally, shape-based features may not be highly accurate due to the resolution limitations of ViT attention maps, which depend on the patch size used in the transformer (in this case, 16x16). Moreover, in assigning a 60 % weight to ViT, we aimed to prioritize its contribution without overemphasizing it, though further experiments will be needed to refine this value depending on the dataset. Despite these challenges, combining multiple models remains an excellent option.

In summary, AI emphasizes non-invasiveness by utilizing images derived from standard modalities and eliminating the need for invasive procedures or chemical applications that may cause discomfort or tissue irritation. Additionally, AI excels in extracting features related to texture, intensity, and shape that traditional methods may miss, allowing for earlier and more precise detection of dysplastic or malignant changes. Combined with its cost-effectiveness, scalability, and minimal reliance on consumables or specialized tools, deep learning emerges as an advantageous alternative. These benefits not only enhance diagnostic performance but also make screening more accessible, particularly in low-resource settings, positioning it as a game-changer in oral cancer diagnostics.

4.3. Limitations of the study

We acknowledge that the main limitation of this study is the relatively small number of cases, particularly in the cancer group and the lack of external validation dataset. Additionally, future investigations should encompass testing additional models to further enhance the comprehensiveness and robustness of the findings.

5. Conclusions

The ViT classification outperformed the radiomics-based approach; however, integrating ViT with radiomics produced even better results, particularly for small training sets. Nonetheless, the attention maps produced by ViT often link oral lesions in cancer patients with areas far from the lesions observed in control patients. For tasks necessitating the analysis and comparison of characteristics within both cancerous and non-cancerous oral lesions, it is advisable to employ the radiomics approach alongside an automated lesion segmentation algorithm.

Our research facilitates the early diagnosis of OSCCs, thereby improving prognosis and survival rates. By focusing on the early detection of potentially malignant lesions and differentiating them from oral lesions, we hope to contribute significantly to the advancement of oral cancer diagnostics and patient care.

CRediT authorship contribution statement

Eva Chilet-Martos: Writing – review & editing, Software, Methodology, Investigation. **Joan Vila-Francés:** Writing – review & editing, Validation, Methodology, Formal analysis. **José V. Bagan:** Writing – review & editing, Project administration, Funding acquisition, Data curation, Conceptualization. **Yolanda Vives-Gilabert:** Writing – original draft, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Ethics statement

This study adheres to the Declaration of Helsinki and was approved by the "Comité Ético de Investigación con medicamentos" (CEIm) of the General University Hospital of Valencia (register number 10–2023, February 24, 2023). All participants provided their informed consent to participate.

Declaration of competing 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.

Acknowledgements

This work was supported by "Ministerio de Ciencia, Innovación y Universidades" and "Fondo Europeo de Desarrollo Regional" [PID2022-138398OB-I00/AEI/10.13039/501100011033/FEDER, UE] and "Generalitat Valenciana, Conselleria d'Innovació, Universitats, Ciència i Societat Digital" [CIAICO/2021/184].

This study adheres to the Declaration of Helsinki and was approved by the "Comité Ético de Investigación con medicamentos" (CEIm) of the General University Hospital of Valencia (register number 10–2023, February 24, 2023). All participants provided their informed consent to participate.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2025.109913.

References

[1] "World health organization, Cancer (2021). Retrieved from, https://www.who.int/health-topics/cancer#tab=tab_1.
[2] S. Warnakulasuriya, Global epidemiology of oral and oropharyngeal cancer, Oral Oncol. 45 (4–5) (Apr. 2009) 309–316, https://doi.org/10.1016/j.oraloncology.2008.06.002.
[3] American Cancer Society, Key Statistics for oral cavity and oropharyngeal cancers, Retrieved from, https://www.cancer.org/cancer/oral-cavity-and-oropharyngeal-cancer/about/key-statistics.html, 2021.
[4] P. Brennan, Pooled analysis of alcohol dehydrogenase genotypes and head and neck cancer: a HuGE review, Am. J. Epidemiol. 159 (1) (Jan. 2004) 1–16, https://doi.org/10.1093/aje/kwh003.

[5] M.P. Rethman, et al., Evidence-based clinical recommendations regarding screening for oral squamous cell carcinomas, J. Am. Dent. Assoc. 141 (5) (May 2010) 509–520, https://doi.org/10.14219/jada.archive.2010.0223.
[6] P.W. Yuen, K.Y. Lam, A.C.L. Chan, W.I. Wei, L.K. Lam, Clinicopathological Analysis of Local Spread of Carcinoma of the Tongue 11The study was supported by a research grant from the University of Hong Kong, grant number 337/048/0014 and 335/048/0081, Am. J. Surg. 175 (3) (Mar. 1998) 242–244, https://doi.org/10.1016/S0002-9610(97)00282-1.
[7] J.A. Woolgar, A. Triantafyllou, A histopathological appraisal of surgical margins in oral and oropharyngeal cancer resection specimens, Oral Oncol. 41 (10) (Nov. 2005) 1034–1043, https://doi.org/10.1016/j.oraloncology.2005.06.008.
[8] C.-T. Liao, et al., Risk stratification of patients with oral cavity squamous cell carcinoma and contralateral neck recurrence following radical Surgery, Ann. Surg Oncol. 16 (1) (Jan. 2009) 159–170, https://doi.org/10.1245/s10434-008-0196-4.
[9] F.W. Mello, et al., Prevalence of oral potentially malignant disorders: a systematic review and meta-analysis, J. Oral Pathol. Med. 47 (7) (Aug. 2018) 633–640, https://doi.org/10.1111/jop.12726.
[10] K.H. Awan, P.R. Morgan, S. Warnakulasuriya, Assessing the accuracy of autofluorescence, chemiluminescence and toluidine blue as diagnostic tools for oral potentially malignant disorders—a clinicopathological evaluation, Clin. Oral Invest. 19 (9) (Dec. 2015) 2267–2272, https://doi.org/10.1007/s00784-015-1457-9.
[11] J.R. Fernandes, L.C.F. Dos Santos, M.L. Lamers, Applicability of autofluorescence and fluorescent probes in the trans-surgical of oral carcinomas: a systematic review, Photodiagnosis Photodyn. Ther. 41 (Mar. 2023) 103238, https://doi.org/10.1016/j.pdpdt.2022.103238.
[12] C.S. Farah, M.J. McCullough, A pilot case control study on the efficacy of acetic acid wash and chemiluminescent illumination (ViziLite™) in the visualisation of oral mucosal white lesions, Oral Oncol. 43 (8) (Sep. 2007) 820–824, https://doi.org/10.1016/j.oraloncology.2006.10.005.
[13] E. Velleuer, et al., Diagnostic accuracy of brush biopsy–based cytology for the early detection of oral cancer and precursors in Fanconi anemia, Cancer Cytopathol 128 (6) (Jun. 2020) 403–413, https://doi.org/10.1002/cncy.22249.
[14] H.J.W.L. Aerts, et al., Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nat. Commun. 5 (1) (Jun. 2014) 4006, https://doi.org/10.1038/ncomms5006.
[15] M.Z.M. Shamim, et al., Automated detection of oral pre-cancerous tongue lesions using deep learning for early diagnosis of oral cavity cancer, Comput. J. 65 (1) (Jan. 2022) 91–104, https://doi.org/10.1093/comjnl/bxaa136.
[16] R.A. Welikala, et al., Automated detection and classification of oral lesions using deep learning for early detection of oral cancer, IEEE Access 8 (2020) 132677–132693, https://doi.org/10.1109/ACCESS.2020.3010180.
[17] A. Ferrer-Sánchez, J. Bagan, J. Vila-Francés, R. Magdalena-Benedito, L. Bagan-Debon, Prediction of the risk of cancer and the grade of dysplasia in leukoplakia lesions using deep learning, Oral Oncol. 132 (Sep. 2022) 105967, https://doi.org/10.1016/j.oraloncology.2022.105967.
[18] M.A.S. Tobias, B.P. Nogueira, M.C.S. Santana, R.G. Pires, J.P. Papa, P.S.S. Santos, Artificial intelligence for oral cancer diagnosis: what are the possibilities? Oral Oncol. 134 (Nov. 2022) 106117 https://doi.org/10.1016/j.oraloncology.2022.106117.
[19] A.J. Wong, A. Kanwar, A.S. Mohamed, C.D. Fuller, Radiomics in head and neck cancer: from exploration to application, Transl. Cancer Res. 5 (4) (Aug. 2016) 371–382, https://doi.org/10.21037/tcr.2016.07.18.
[20] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Commun. ACM 60 (6) (May 2017) 84–90, https://doi.org/10.1145/3065386.
[21] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, Li Fei-Fei, ImageNet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Miami, FL, Jun. 2009, pp. 248–255, https://doi.org/10.1109/CVPR.2009.5206848.
[22] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, ArXiv150504597 Cs [Online]. Available, http://arxiv.org/abs/1505.04597, May 2015. (Accessed 10 September 2018).
[23] P.R. Jeyaraj, E.R. Samuel Nadar, Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm, J. Cancer Res. Clin. Oncol. 145 (4) (Apr. 2019) 829–837, https://doi.org/10.1007/s00432-018-02834-7.
[24] S. Xu, et al., An early diagnosis of oral cancer based on three-dimensional convolutional neural networks, IEEE Access 7 (2019) 158603–158611, https://doi.org/10.1109/ACCESS.2019.2950286.
[25] K. Warin, W. Limprasert, S. Suebnukarn, S. Jinaporntham, P. Jantana, Automatic classification and detection of oral cancer in photographic images using deep learning algorithms, J. Oral Pathol. Med. 50 (9) (Oct. 2021) 911–918, https://doi.org/10.1111/jop.13227.
[26] A. Vaswani, et al., Attention is all you need, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017 [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
[27] A. Dosovitskiy, et al., An image is worth 16x16 words: transformers for image recognition at scale, in: International Conference on Learning Representations, 2021 [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy.
[28] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-End object detection with transformers, in: Eur. Conf. Comput. Vis., 2020, pp. 213–229, https://doi.org/10.48550/ARXIV.2005.12872.
[29] A. Kirillov, et al., Segment Anything, 2023, https://doi.org/10.48550/ARXIV.2304.02643.
[30] K. He, et al., Transformers in medical image analysis, Intell. Med. 3 (1) (Feb. 2023) 59–78, https://doi.org/10.1016/j.imed.2022.07.002.
[31] Md M. Islam, K.M.R. Alam, J. Uddin, I. Ashraf, M.A. Samad, Benign and malignant oral lesion image classification using fine-tuned transfer learning techniques, Diagnostics 13 (21) (Nov. 2023) 3360, https://doi.org/10.3390/diagnostics13213360.
[32] B. Singha Deo, M. Pal, P.K. Panigrahi, A. Pradhan, Supremacy of attention based convolution neural network in classification of oral cancer using histopathological images, Oncology (Nov. 2022), https://doi.org/10.1101/2022.11.13.22282265 preprint.
[33] J. Ma, Y. He, F. Li, L. Han, C. You, B. Wang, Segment anything in medical images, Nat. Commun. 15 (1) (Jan. 2024) 654, https://doi.org/10.1038/s41467-024-44824-z.
[34] M. Aubreville, et al., Automatic classification of cancerous tissue in laserendomicroscopy images of the oral cavity using deep learning, Sci. Rep. 7 (1) (Sep. 2017) 11979, https://doi.org/10.1038/s41598-017-12320-8.
[35] A. Demircioğlu, Are deep models in radiomics performing better than generic models? A systematic review, Eur. Radiol. Exp. 7 (1) (Mar. 2023) 11, https://doi.org/10.1186/s41747-023-00325-0.
[36] S. Warnakulasuriya, Newell W. Johnson, I. Van Der Waal, Nomenclature and classification of potentially malignant disorders of the oral mucosa: potentially malignant disorders, J. Oral Pathol. Med. 36 (10) (Jul. 2007) 575–580, https://doi.org/10.1111/j.1600-0714.2007.00582.x.
[37] C.A. Schneider, W.S. Rasband, K.W. Eliceiri, NIH Image to ImageJ: 25 years of image analysis, Nat. Methods 9 (7) (Jul. 2012) 671–675, https://doi.org/10.1038/nmeth.2089.
[38] G. Merlin, N. Vedant, R. Ruchit, and T. Mariya, "What Happens During Finetuning of Vision Transformers: An Invariance Based Investigation," in Proceedings of the 2nd Conference of Lifelong Learning Agents, pp. 601–619. [Online]. Available: https://proceedings.mlr.press/v232/merlin23a/merlin23a.pdf.
[39] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, R. Girshick, Masked autoencoders are scalable vision learners, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, New Orleans, LA, USA, Jun. 2022, pp. 15979–15988, https://doi.org/10.1109/CVPR52688.2022.01553.
[40] S. He, et al., Computer-Vision Benchmark Segment-Anything Model (SAM) in Medical Images: Accuracy in 12 Datasets, 2023, https://doi.org/10.48550/ARXIV.2304.09324.
[41] Y. Huang, et al., Segment Anything Model for Medical Images?, 2023, https://doi.org/10.48550/ARXIV.2304.14660.
[42] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *2016 IEEE Conference On Computer Vision And Pattern Recognition (CVPR)*, Las Vegas, NV, USA, IEEE, Jun. 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.
[43] J.J.M. van Griethuysen, et al., Computational radiomics system to decode the radiographic phenotype, Cancer Res. 77 (21) (Nov. 2017) e104–e107, https://doi.org/10.1158/0008-5472.CAN-17-0339.
[44] F. Pedregosa, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (85) (2011) 2825–2830 [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html.
[45] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (Sep. 1995) 273–297, https://doi.org/10.1007/BF00994018.
[46] T.K. Ho, Random decision forests, in: Proceedings of 3rd International Conference on Document Analysis and Recognition, IEEE, 1995, pp. 278–282.
[47] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. (2001) 1189–1232.
[48] A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: a corrected feature importance measure, Bioinformatics 26 (10) (May 2010) 1340–1347, https://doi.org/10.1093/bioinformatics/btq134.
[49] M.K. Ickler, M. Baumgartner, S. Roy, T. Wald, K.H. Maier-Hein, Taming detection transformers for medical object detection, in: T.M. Deserno, H. Handels, A. Maier, K. Maier-Hein, C. Palm, T. Tolxdorff (Eds.), Bildverarbeitung für die Medizin 2023, Springer Fachmedien Wiesbaden, Wiesbaden, 2023, pp. 183–188, https://doi.org/10.1007/978-3-658-41657-7_39. Informatik aktuell.
[50] W. Ji, J. Li, Q. Bi, T. Liu, W. Li, L. Cheng, Segment Anything Is Not Always Perfect: an Investigation of SAM on Different Real-World Applications, 2023, https://doi.org/10.48550/ARXIV.2304.05750.
[51] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jegou, Training data-efficient image transformers & distillation through attention, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 139, PMLR, Jul. 2021, pp. 10347–10357 [Online]. Available: https://proceedings.mlr.press/v139/touvron21a.html.