Array 30 (2026) 100825

Contents lists available at ScienceDirect

Array

journal homepage: www.sciencedirect.com/journal/array

Oral tumor detection and localization using detection transformer (DETR) with multi-scale feature extraction and self-supervised learning

Pratham Kaushik [a], Vinay Kukreja [a], Eshika Jain [a], Modafar Ati [b,*], Shanmugasundaram Hariharan [c]

[a] Centre for Research Impact and Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
[b] Computer Science and Information Technology, College of Engineering, Abu Dhabi University, Abu Dhabi, United Arab Emirates
[c] Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India

ARTICLE INFO

Keywords:
Tumor detection
Detection transformer (DETR)
Oral cancer localization
Self-supervised learning (SSL)
Attention mechanism
Grad-CAM visualization
Image augmentation

ABSTRACT

This study presents a transformer-based framework for lesion-level detection and localization of oral tumors using the Detection Transformer (DETR) architecture under limited-data medical imaging conditions. The dataset consists of 131 clinically acquired oral images (87 cancerous, 44 non-cancerous). To prevent information leakage and inflated performance estimation, dataset partitioning was performed at the original-image level prior to augmentation, and a 5-fold stratified cross-validation protocol was implemented. Data augmentation was applied exclusively to training subsets, while evaluation was conducted only on non-augmented validation images. To enhance representation learning under low annotation availability, a SimCLR-based self-supervised pretraining stage was incorporated within each cross-validation fold using only training-fold images. The pretrained ResNet-50 backbone was then fine-tuned within the DETR framework for supervised tumor localization. Across cross-validation folds, the proposed model achieved a mean mAP of 0.496 ± 0.027, AP50 of 0.738 ± 0.031, and mean IoU of 0.638 ± 0.019. Bootstrap-based 95% confidence intervals were reported to quantify statistical uncertainty. External validation on an independent oral lesion dataset demonstrated consistent localization performance under domain shift, supporting generalization beyond the primary dataset. Ablation analysis confirmed that self-supervised initialization improves detection stability compared to random initialization. The results demonstrate that a leakage-free transformer-based detection framework combined with self-supervised representation learning can provide reliable lesion-level localization performance in low-resource oral cancer imaging scenarios.

1. Introduction

The condition of oral cancer continues to be one of the primary causes of deaths and diseases related to cancers all around the globe [1] [2]. The annual number of cases of 350,000 new oral cancer patients indicates the importance of early diagnosis to improve healthcare outcomes for patients [3]. Early-stage tumors do not show any symptoms, which leads to a patient only going to medical facilities if their condition is in an advanced stage [4] [5]. Biopsy and CT scans along with physical examination reveal effectiveness, but these conventional diagnostic techniques require a lot of time, along with being invasive and subject to human mistakes [6] [7]. A non-invasive automated system that detects and locates oral tumors at early stages needs immediate development [8] [9]. The research introduces a contemporary deep learning approach based on Detection Transformer (DETR) to perform both oral tumor recognition and localization functions, which seeks to change current practices in oral cancer identification [10] [11].

1.1. The growing challenge of oral cancer

Early diagnosis of oral cancer proves challenging because its tumors grow in regions that are hard to get to, and they look almost identical to benign sores early on [12] [13]. The development of automated systems that differentiate between cancerous and non-cancerous lesions needs emphasis, since these systems must also pinpoint tumor areas with precision [14] [15]. The identification of tumors at an early stage

* Corresponding author.
E-mail addresses: prathamkaushik8@gmail.com (P. Kaushik), onlyvinaykukreja@gmail.com (V. Kukreja), eshikajain0150@gmail.com (E. Jain), modafar.ati@adu.ac.ae (M. Ati), mailtos.hariharan@gmail.com (S. Hariharan).

https://doi.org/10.1016/j.array.2026.100825
Received 22 December 2025; Received in revised form 10 April 2026; Accepted 13 April 2026
Available online 20 April 2026

presents a major chance to boost patient survival rates, which demonstrates the usefulness of automated tumor detection software in medical imaging. The current medical imaging systems face significant difficulties in their accuracy performance because they scrutinize images containing various tumor dimensions and configurations within different locations [16] [17]. Deep learning-based methods, especially transformer architectures, make exciting prospects available to the field [18] [19]. These models learn data patterns without human intervention; therefore, they adapt to a wide range of tumor features and imaging settings to provide a dependable and time-saving alternative beyond traditional methods [20] [21].

1.2. Traditional methods for tumor detection

The present oral cancer diagnosis relies on visual inspection accompanied by procedures that involve long waiting times, as well as complications and hazards [22] [23]. Advanced technology attached with high costs restrains the clinic applications of CT scans as well as MRI in the determination of the size as well as the position of the tumor [24] [25]. Traditional diagnostic methods work in most cases, but human errors as well as invasive procedures would be reduced after automated technology enters conventional use [26] [27]. The human evaluation process that leads to diagnoses causes late detection of the tumors, as healthcare professionals rely on human tests, and non-tumorous diseases may initiate wrong malignant findings that lead to delays [28] [29]. Inconsistent diagnoses are caused by radiologist interpretation variability as well as variability among different interpreters [30] [31].

1.3. Machine learning and deep learning methods for tumor detection

The development of Machine Learning (ML) has brought major advancements to oral tumor detection capabilities [32] [33]. The Support Vector Machines (SVM) technique, together with Random Forests and K-Nearest Neighbours (KNN), serve as ML methods for medical image-based tumor region classification. These applied methods deliver satisfactory results but experience limitations during the scaling-up process across different datasets and image acquisition settings [34] [35]. The systems use manually constructed features and need significant access to labeled data during their training process. The main restriction with current ML methods is their inability to adjust their operation dynamically for complex patterns and different tumor variations, as deep learning techniques handle automatically [36]. Their performance diminishes when processing previously undetected data and different tumor types, so better complex medical image models are essential to improve adaptability [37] [38]. Deep Learning (DL) innovations now transform the detection processes as well as the localization of tumors. Traditional machine learning techniques require manual feature extraction, but deep learning algorithms, particularly Convolutional Neural Networks, learn raw data features automatically through an automatic hierarchical process [39] [40]. Clinical applications of CNNs for medical image analytics have grown popular, specifically in oral tumor detection, because this method efficiently detects patterns within extensive medical datasets [41] [42]. Detection Transformer (DETR) stands as a novel transformer-based model that represents an effective solution for detecting objects across tumor localization tasks [43]. DETR removes the necessity of region proposal networks while directly producing box positions and class outputs throughout a complete end-to-end training process [44] [45]. Transformers allow the model to examine image regions that matter most through their self-attention function, which leads to better tumor detection accuracy and localization precision [26] [46].

1.4. DETR for oral tumor localization

The proposed approach builds on the Detection Transformer (DETR) architecture, which is known to perform end-to-end object detection very well. The model is trained on a medical dataset which contains 131 images augmented in the training process to give more variation; however, evaluation of the model was performed only on original images without augmentation, in order to deal with tumor size, position and occlusion. The multi-scale feature extraction helps the model in better detecting tumors of big and small size whereas self-supervised learning helps in reducing the dependency of large volumes of labeled data. Preprocessing including normalization, resizing to 512×512 pixels and augmentation to ensure that the model will generalize well across data that it has not seen before and will also adapt well to different tumor types.

1.5. Study rationale and technical positioning

Although deep learning techniques have been extensively studied for oral cancer diagnosis, most previous studies were performed mainly at image level and focus only on binary classification based on convolutional neural networks (CNNs), transfer learning models or ensemble approaches. These frameworks are designed to find out whether an image contains a malignant lesion or not but do not explicitly tackle the problem of precise lesion-level spatial localization as a structured prediction problem. In the case of clinical screening scenarios, however, localization is also of central importance, as in this scenario, clinicians need spatial delineation of suspicious regions rather than a simple binary decision. Region proposal-based object detectors like Faster R-CNN and YOLO versions have been explored in related areas of medical imaging. However, they are based on anchor box design, heuristic non-maximum suppression, and multi-stage detection pipeline, which can add architectural complexity and sensitivity to hyperparameter configuration, especially for low-resource medical datasets. In addition, most published oral cancer imaging studies do not test transformer-based set prediction frameworks in limited-data set conditions. Detection Transformer (DETR) conceptualises object detection as a direct set prediction problem with bipartite matching and global self-attention without any anchor engineering and region proposal network. Despite the recent increase of its use in natural image detection, its suitability in small-scale oral cancer imaging datasets has not been thoroughly explored. The present study does not simply apply a standard DETR model on a public dataset. Instead, it explores the feasibility and stability of the tumor localization approach that relies on transformers based on limited medical imaging conditions characterized by.

1. Small amount of annotated data (131 original images)
2. High intra-class variability in appearance of lesions
3. Small and irregular tumor morphology
4. Lack of large-scale clinical training corpora.

In order to overcome these limitations, the following are included in the proposed framework.

1. Multi-scale feature extraction in the DETR backbone to enhance the sensitivity to small lesions
2. Self-Supervised Pretraining to Improve Representation Learning Before Supervised Fine-Tuning,
3. A well-organized localization-oriented evaluation protocol with bounding box quality and spatial accuracy.

Thus, the novelty of this work is not the introduction of a novel transformer architecture, but the systematic adaptation and evaluation of DETR for the localization of oral tumors under realistic low resource medical imaging conditions, focusing on the leakage free design of experiments and the structured evaluation of the experiments in terms of their spatial evaluation. Unlike existing transformer-based medical imaging works that mainly emphasize on the architectural modification or large-scale data settings, the present work proposes a methodology-

driven work that is adapted for the low-resource clinical imaging conditions. The novelty of this work is the combination of a leakage-free experimental protocol, original data partitioning at the image level, fold-wise self-supervised pretraining protocol to avoid representation leakage, and a statistical-based evaluation framework with cross-validation and bootstrap-based uncertainty estimation. Furthermore, the work systematically adapts the DETR architecture for small-scale oral cancer datasets without anchor design and heuristic post-processing dependency and validates the performance under domain shift through the evaluation of external datasets. This combination of methodological rigor, leakage control, and representation learning in limited data conditions make the proposed framework well-suited for distinguishing itself from the existing transformer-based approaches in medical imaging.

The differentiation of the present study from closely related work lies primarily in its methodological framing and evaluation design rather than in the introduction of a fundamentally new detection architecture. Most prior oral cancer imaging studies have emphasized image-level classification, transfer learning, or conventional deep convolutional pipelines, with limited attention to lesion-level spatial localization as a structured prediction problem under small-scale data constraints. In contrast, the present work formulates oral tumor analysis as an end-to-end transformer-based bounding-box prediction task and evaluates it under a rigorously controlled low-resource protocol that combines original-image-level leakage-free partitioning, fold-wise self-supervised pretraining, cross-validation-based performance aggregation, bootstrap uncertainty estimation, and external validation under domain shift. This combination provides a clearer methodological distinction from related studies that report performance primarily on classification-oriented settings or under less explicitly controlled evaluation designs.

1.6. Key contributions

The contributions of this study are as follows.

1. **Reformulation of Oral Cancer Diagnosis to a Localization Task:** Unlike previous classification-oriented studies, this work explicitly treats the problem of detecting oral cancer as an end-to-end bounding box prediction problem by using transformer-based set prediction.
2. **Adaptation of DETR for Small-Scale Medical Imaging:** The standard DETR framework is modified to be able to effectively work on a small dataset using structured multi-scale feature integration and training stabilization approaches that are applicable to low-data regimes.
3. **Self-Supervised Pretraining for Data Efficiency Integration:** A contrastive self-supervised learning stage is added before supervised learning to better generalize the features due to limited availability of annotations.
4. **Localization-Oriented Evaluation Framework:** The research focuses on assessment of spatial detection measures and bounding box quality instead of only on classification accuracy.
5. **Systematic Experimental Redesign in Leakage-Free Protocol:** The data pipeline is designed to ensure that it splits at the level of original images before augmentation so that information leakage is avoided and the generalization assessment is reliable.
6. **Methodological Differentiation from Prior Work:** The study establishes a rigorously controlled low-resource evaluation framework that differentiates the proposed approach from prior classification-oriented and anchor-based oral cancer imaging studies through lesion-level localization, leakage-free partitioning, fold-wise self-supervised representation learning, uncertainty quantification, and external validation.

2. Materials and methods

Mouth cancer, or oral cancer, is a serious public health problem which is not easily identified at an early stage and accurately diagnosed. Early detection of oral cancer is important for better survival rate as it opens up the possibility for intervention and cure in due time. Traditional visual inspection and biopsy procedures, in general, are quite restrictive in accuracy, and highly invasive. Recent advances in deep learning and vision have proven to be promising in terms of automated detection of cancerous areas in medical imaging. In this approach, we propose an effective approach of mouth cancer detection based on Detection Transformer (DETR), with the complement of multi-scale feature extraction and self-supervised learning methods. This combination is designed to overcome the difficulty in detecting and localizing cancer regions in oral images, resulting in an efficient and scalable solution for lesion-level tumor localization, as illustrated in Fig. 1.

2.1. Phase 1: data collection and preprocessing

The primary dataset used in this study was obtained from the publicly available Kaggle repository “Mouth Cancer Images.” The dataset consists of 131 oral cavity images, including 87 cancerous and 44 non-cancerous cases. These images were used for lesion-level annotation, preprocessing, and subsequent model development under the leakage-free experimental design described below, as shown in Fig. 2.

2.1.1. Clinical characteristics and dataset acquisition context

The main data set for the purposes of this study is a set of 131 oral cavity images from the Kaggle “Mouth Cancer Images” repository that are publicly available. As a secondary public data set, detailed acquisition metadata such as imaging device details, illuminations, lesion staging (TNM classification), anatomical site distribution and inter-observer variability are not available from the source of this data set. The images are apparently acquired through standard digital photography under non-standardized illumination conditions in outpatient or clinical screening settings. Visual examination shows variation of background texture, intensity of lighting and intra-oral viewpoint, therefore heterogeneous acquisition conditions rather than controlled laboratory imaging. The data set includes both malignant ($n = 87$) and non-malignant ($n = 44$) oral lesions, which include visible mucosal surfaces of the buccal mucosa, tongue and gingival areas. Because this dataset is publicly-curated and not obtained from a single institutional protocol, standardized staging annotations and multi-expert consensus annotations are not available. Consequently, the present study assesses lesion-level spatial localization, as opposed to clinical staging prediction or prognostic modelling.

To mitigate the risk of trivial artifact learning, the experimental protocol incorporates.

(i) leakage-free splitting at the original-image level,
(ii) augmentation-based robustness to illumination and viewpoint variation, and
(iii) external validation on an independent oral lesion dataset (Section 3.2), which reduces the likelihood that performance arises from dataset-specific artifacts.

The study therefore focuses on methodological evaluation of transformer-based lesion localization under limited-data conditions rather than claiming direct clinical deployment readiness. As the dataset is limited in scale and publicly curated, it may not fully capture the diversity of lesion morphology, imaging conditions, and patient-level variability encountered in broader clinical populations.

2.1.2. Preprocessing

Preprocessing is crucial for ensuring that the model receives data in a standardized format to maximize its efficiency. The following pre-

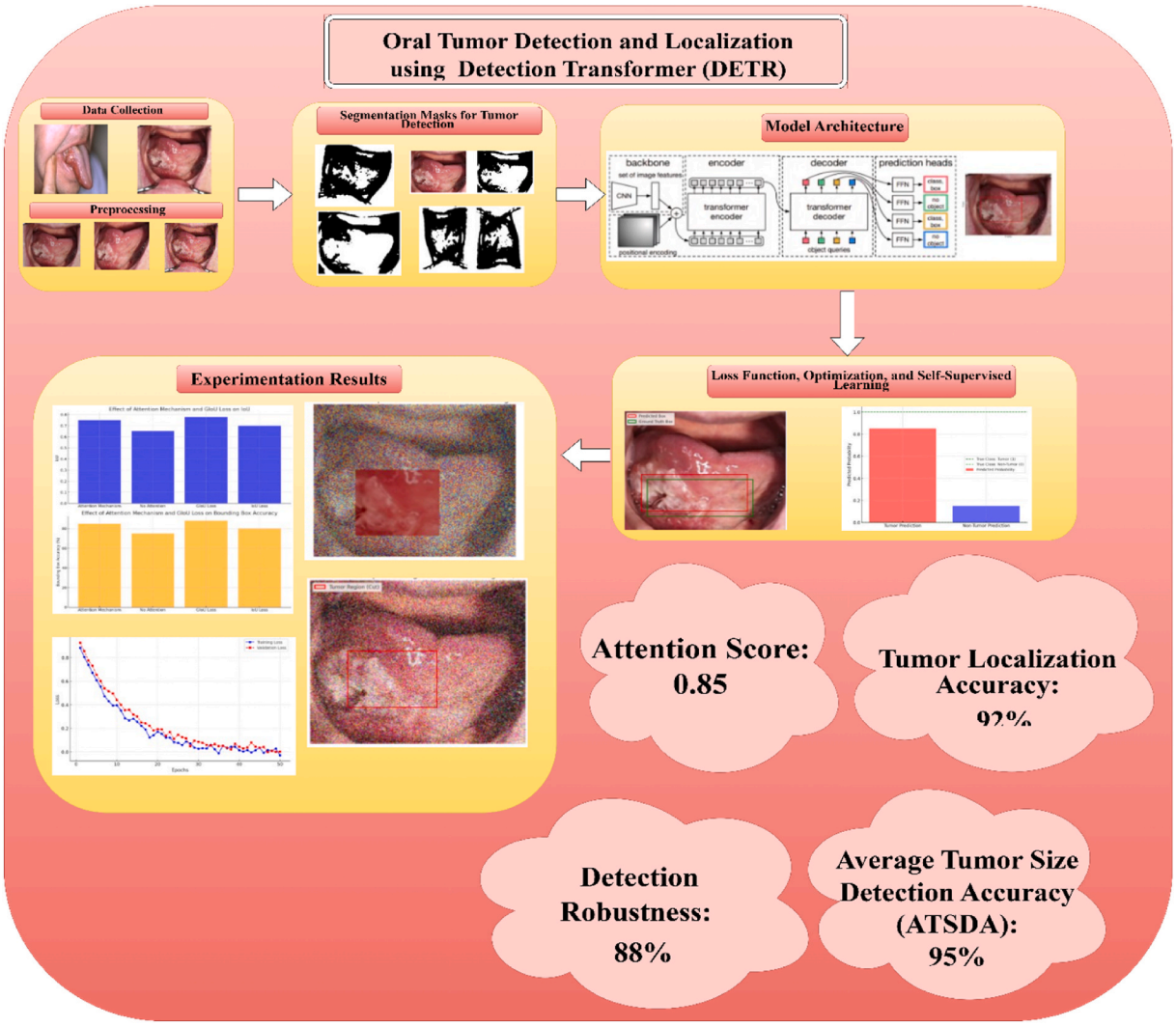

Fig. 1. Proposed methodology for oral tumor detection.

a) Cancerous Mouth b) Cancerous Mouth c)Non-Cancerous Mouth d) Non-Cancerous Mouth

Fig. 2. Sample images from the dataset [47].

processing steps will be applied.

a) **Resizing:** All the images were resized to 512 × 512 pixels so as to have architectural compatibility with the DETR backbone and to have the same spatial resolution across the cross-validation folds, as in Fig. 3. Pixel intensities were normalized with dataset-level mean and standard deviation for stable optimization of the optimization process during the training stage.

b) **Normalization:** Normalization of the image is required in order to scale the pixel values within the range consistent with the training phase. Mathematically, for this research, the normalization can be expressed as equation (1):

$$I_{norm} = \frac{I_{resized} - \mu}{\sigma} \quad (1)$$

Where:

I_{norm} represents the normalized image, which is the result after applying the scaling transformation.

Fig. 3. Resized oral image.

$I_{resized}$ is the resized image that has already been standardized to a 512 × 512 resolution.

μ is the mean of pixel values computed from the entire dataset, capturing the average intensity of the images.

σ represents the standard deviation of pixel values across the dataset, which reflects the variance in intensity values across the images.

c) **Data Augmentation:** To improve robustness under limited data conditions, augmentation was applied exclusively to training subsets within each fold. Transformations included random rotation (≤45°), horizontal/vertical flipping, random resized cropping, brightness and color jittering, Gaussian noise, and perspective transformation. Validation and external evaluation images were not augmented., as shown in Table 1:

These preprocessing steps were designed to improve input consistency and robustness while preserving a leakage-free evaluation protocol.

2.1.3. Dataset partitioning and leakage-free experimental protocol

To ensure statistically valid evaluation and prevent information leakage, dataset partitioning was performed at the level of the 131 original images prior to any augmentation. Stratified sampling was applied to preserve the class distribution (87 cancerous, 44 non-cancerous) across subsets. A 5-fold stratified cross-validation protocol was implemented at the original-image level. In each fold, images were divided into training and validation subsets without overlap of original samples. Augmentation techniques (rotation, flipping, brightness variation, Gaussian noise, perspective transformation) were applied exclusively to the training subset. No augmented variants were included in validation or testing subsets. All performance metrics were computed on non-augmented, unseen original images. This protocol ensures that no correlated augmented samples appear across folds, thereby preventing artificially inflated performance estimates and enabling reliable generalization assessment.

Table 2 summarizes the original-image-level stratified cross-validation distribution. Each image appears in the validation subset exactly once across the five folds. Augmentation operations were applied exclusively to training images within each fold and were not considered independent samples for evaluation. All reported performance metrics correspond to validation predictions on non-augmented original images.

2.1.4. Bounding box annotation derived from lesion masks

In this study, lesion masks were used only to derive ground-truth bounding boxes for tumor localization. The masks were not used for pixel-level optimization or segmentation-based evaluation. Instead, rectangular bounding boxes enclosing annotated lesion regions were computed and used as supervision targets for DETR training, while evaluation was performed exclusively using bounding-box-based metrics such as mAP, AP50, AP75, and IoU.

Let $I(x,y)$ denote the intensity value of a pixel at spatial location (x,y) in an oral image. A binary lesion mask $M(x,y)$ can be defined as equation (2):

$$M(x,y) = \begin{cases} 1, & \text{if pixel } (x,y) \text{ belongs to a tumor region} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

The mask can be represented as a function of the image, as seen in equation (3):

$$M(x,y) = f(I(x,y)) \quad (3)$$

where $f(\cdot)$ denotes a mapping that identifies tumor pixels based on annotation. In this study, these masks were not used for pixel-level optimization but were employed to compute bounding boxes enclosing connected tumor regions. For each annotated lesion mask, the corresponding bounding box B was derived by identifying the minimum and maximum spatial coordinates of tumor pixels depicted in equation (4):

$$B = (x_{\min}, y_{\min}, x_{\max}, y_{\max}) \quad (4)$$

where.

- $x_{\min} = \min\{x | M(x,y) = 1\}$
- $y_{\min} = \min\{y | M(x,y) = 1\}$
- $x_{\max} = \max\{x | M(x,y) = 1\}$
- $y_{\max} = \max\{y | M(x,y) = 1\}$

These bounding box coordinates were used as ground-truth targets during DETR training. This approach ensures that spatial tumor extent is preserved while maintaining methodological consistency with object detection-based evaluation. By explicitly distinguishing between mask-based annotation and bounding-box-based training and evaluation, the framework remains strictly aligned with object detection objectives rather than pixel-level segmentation tasks, as seen in Fig. 4.

2.1.4.1. Augmenting segmentation masks during data augmentation. During augmentation, identical geometric transformations were applied to each image and its corresponding lesion mask to preserve spatial consistency during bounding-box extraction. For example, under rotation by angle θ, the transformed mask coordinates were computed using the standard 2D rotation formulation shown in Equations (5) and (6). This ensured that bounding boxes derived from transformed masks remained aligned with the augmented images.

For instance, in the case of a random rotation by an angle θ, the transformation of the mask is mathematically represented as equation

Table 1
Data augmentation techniques applied.

Technique Name	Description	Output
Random Rotations	To simulate variations in the orientation of the images, random rotations will be applied. The images will be rotated randomly by up to 45°.	
Flipping	Both horizontal and vertical flips will be applied to introduce invariance to the orientation of the object.	 
Zooming and Cropping	Random zooming and cropping will be applied to simulate variations in the object size and positioning within the image.	
Random Brightness Adjustments	Lighting conditions in real-world medical images can vary, so random brightness adjustments will be applied to account for different lighting scenarios	
Random Color Jitter	Variations in the color of medical images can arise due to differences in imaging equipment, lighting conditions, or patient characteristics. Random adjustments to the brightness, contrast, saturation, and hue will be applied to simulate these color changes	
Gaussian Noise	To simulate noisy data, Gaussian noise will be added to the images. This transformation makes the model more robust to noise, helping it perform well in real-world scenarios where image quality may vary.	
Random Perspective Transforms:	Perspective transformations will be applied to simulate changes in the viewpoint of the image. This helps the model learn to detect cancerous regions from different angles, as real-world images may be captured from various perspectives.	

(5):

$$M_{augmented}(x', y') = M(x, y), \tag{5}$$

Where (x', y') are the new coordinates of the pixel after the transformation. For a 2D rotation by an angle θ, the new coordinates are given by equation (6):

Table 2
Original-image-level 5-fold stratified cross-validation distribution.

Fold	Training Images	Validation Images	Cancer (Train)	Non-Cancer (Train)	Cancer (Val)	Non-Cancer (Val)
Fold 1	104	27	69	35	18	9
Fold 2	104	27	69	35	18	9
Fold 3	104	27	69	35	18	9
Fold 4	104	27	69	35	18	9
Fold 5	104	27	69	35	18	9
Total Original Images	131	—	87	44	—	—

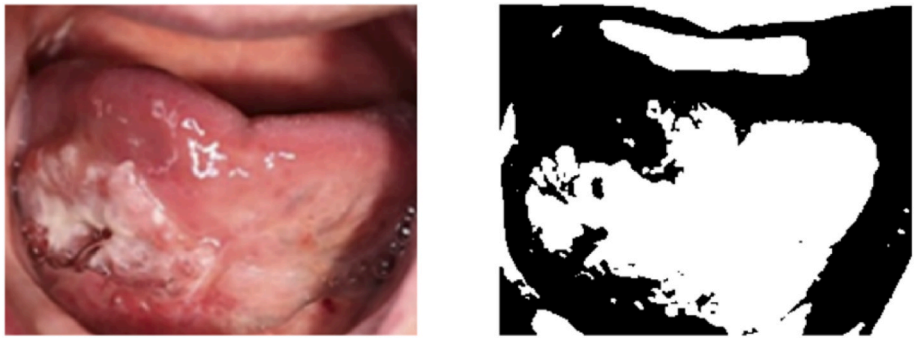

Fig. 4. Tumor segmentation mask.

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{6}$$

The segmentation mask together with the input image undergoes an equivalent transformation, which preserves tumor positioning consistency in the resulting mask. Y' transformation (both vertical and horizontal) needs to be applied twice: first to the image then to the mask in order for the model to learn tumor identification independently of image rotation as seen in Fig. 5.

2.1.4.2. Importance of segmentation masks in training. Through its segmentation mask, the model receives proper bounds for tumor-specific features, which direct its analysis. Deep learning model detection of cancerous images depends heavily on these masks during their training process. By pairing the original image with the mask, the model receives training through an association between pixels marked as 1 in the mask and patterns found in the image. The model learns better generalization skills for tumor image sets through this method. In this study, segmentation masks were not used for pixel-level optimization or evaluation. Instead, they were employed exclusively for deriving ground-truth bounding boxes. All performance evaluation was conducted using bounding-box-based detection metrics, including mAP and Intersection over Union (IoU).

2.1.4.3. Annotation source and mask-to-bounding box conversion protocol. The pixel level lesion masks used in this study were taken from the publicly available "Mouth Cancer Images" dataset itself. As this is a secondary curated dataset, detailed metadata in terms of the number of annotators and formal multi-expert consensus procedures is not provided by the metadata from the source. Therefore, the masks were considered as standardized reference annotations of extents of lesions. To ensure consistency and suitability to be used for detection-based training, all masks were visually inspected before being used. Masks with incomplete coverage, noise artifacts, and irregular boundaries were improved to ensure spatial coherence and minimize annotation inconsistencies. For mask-to-bounding box conversion, bounding boxes were computed by extracting the minimum and maximum spatial coordinates $(x_{min}, y_{min}, x_{max}, y_{max})$ of annotated tumor pixels. In case, more than one spatially disconnected lesions were present in the same image, all the connected components were considered as different instances and

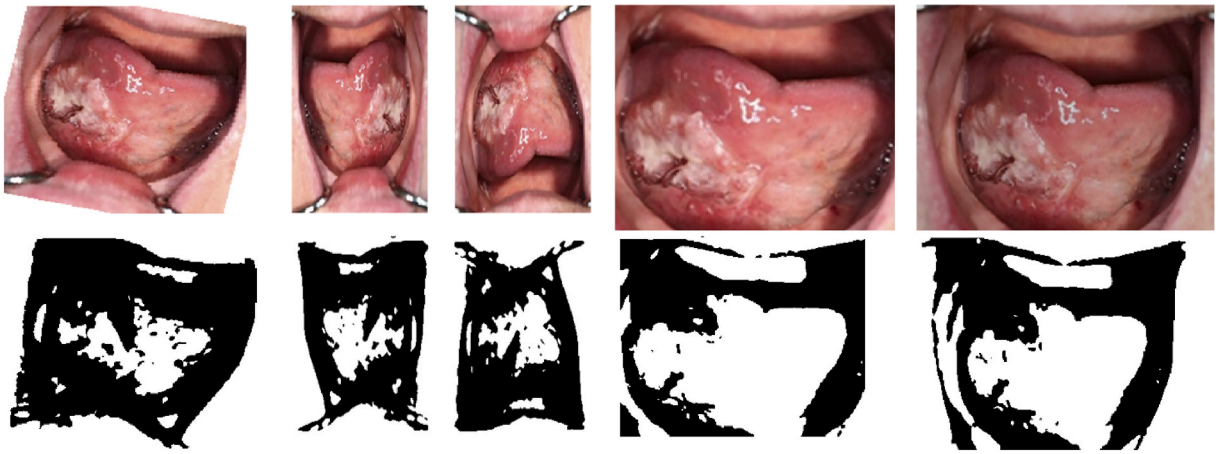

Fig. 5. Augmented segmentation mask.

bounding boxes were generated accordingly. No extra padding or margin was added when creating bounding boxes. A tight strategy of enclosure was adopted, in order to maintain precise lesion boundaries, as well as to avoid inclusion of surrounding non-lesion tissue to comply with lesion-level localization objectives.

2.2. Phase 2: model architecture

2.2.1. Detection transformer (DETR) for tumor detection and localization

The proposed solution for oral image cancer region detection and bounding-box localization relies on the Detection Transformer (DETR) architecture, which is employed exclusively for object detection without any pixel-level segmentation branch. The Detection Transformer (DETR) architecture was employed exclusively for object detection and lesion-level bounding box localization. In this study, no semantic or instance segmentation branch was implemented. The standard object detection process starts with RPNs that generate potential regions, after which classification and bounding box regression take place. The method functions entirely without region proposals since it uses learnable object queries to directly forecast object positions. The training procedure refines the queries so that the transformer learns tumor positions in medical images and generates accurate bounding box predictions and class identifications in an end-to-end manner, as seen in Fig. 6.

2.2.2. Object queries in DETR for tumor localization

Medical image detection along with localization for oral cancer detection uses object queries as central components within Detection Transformers (DETR). The query objects serve as trainable embedded elements for representing possible object positions that exist in images. The DETR approach eliminates the use of region proposals and anchor boxes since it directly detects objects (for instance, tumors) through its query system. Image feature extraction functions as the first stage in this process, and the decoder learns to connect extracted features with object queries. The object queries improve their localization and segmentation abilities of tumors through backpropagation-based refinement during the training process. The model learns to connect the initialized learnable embedding object queries Q with the extracted features F from the image. This relationship is captured by the decoder module, which refines the object queries to predict object locations such as tumors. Mathematically, this process can be described by the following in equation (7):

$$O = Decoder(Q, F) \tag{7}$$

Where:

$O \in \mathbb{R}^{N\ X\ 4}$ represents the set of predicted object locations for tumor i, where each object location corresponds to a bounding box defined by its top-left corner (x_{min}, y_{min}) and bottom-right corner (x_{max}, y_{max}).

$Q \in \mathbb{R}^{N\ X\ D}$ is the set of object queries (learnable embeddings), where N is the number of object queries (typically fixed, e.g., 100 queries) and D is the dimensionality of each query embedding.

$F \in \mathbb{R}^{H\ X\ W\ X\ C}$ represents the image feature map extracted by the encoder, where H, W, and C are the height, width, and channels of the feature map, respectively. The encoder is typically a convolutional neural network (CNN) backbone (such as ResNet or Swin Transformer) that produces high-dimensional feature maps from the input image, as seen in Fig. 7.

The role of the decoder is to use these object queries Q in conjunction with the image features F to predict the locations of tumors. The decoder learns to associate each object query with specific features in the image, progressively refining the object queries to localize tumors.

2.2.3. Mathematical refinement of object queries

The object queries Q are initialized as random vectors (embeddings) and progressively refined throughout training. This refinement process is governed by the loss function, which guides the model to adjust the object queries so that they more accurately correspond to the true tumor locations in the ground-truth data. During each iteration of training, backpropagation updates the object queries through the following equation (8):

$$Q_{new} = Q_{old} + \Delta Q \tag{8}$$

Where:

Q_{new} represents the refined object queries after one update,

Q_{old} is the previous set of object queries,

ΔQ is the change made to the queries based on the gradient descent optimization process (calculated using the loss function).

This process can be used for the object queries to converge towards optimal embeddings representing tumor locations in the image. As the training proceeds, the model optimizes these queries in such a way that they better match the regions of interest (i.e. tumor regions) and thus the network is able to localize and detect tumors more accurately.

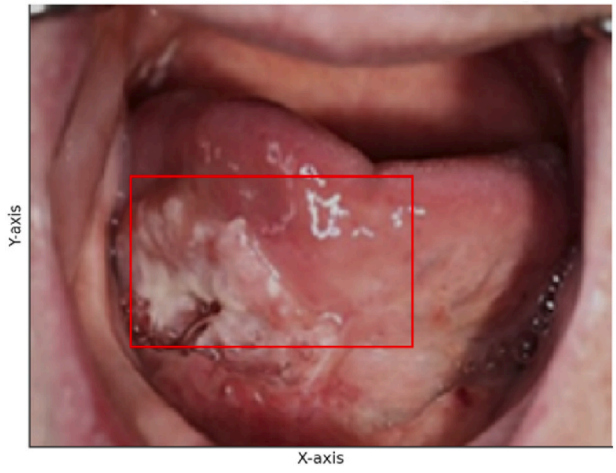

Fig. 7. Tumor localization using predicted bounding boxes.

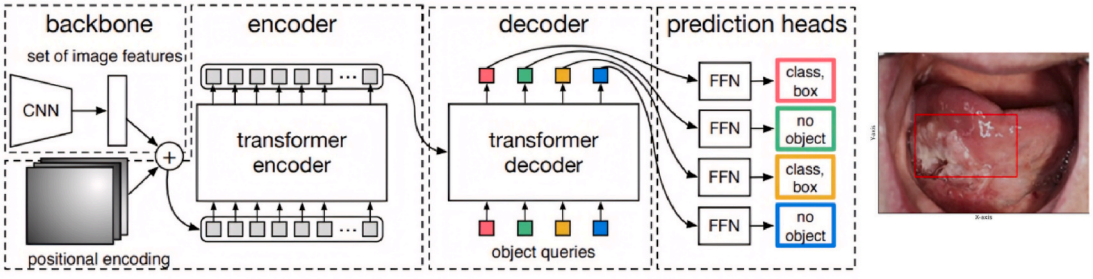

Fig. 6. DETR architecture for tumor detection using backbone feature extraction and transformer-based prediction.

2.2.4. Hierarchical refinement using attention mechanism

The refinement of object queries is further enhanced by the multi-head attention mechanism. The attention mechanism enables the model to focus on different parts of the image by computing multiple attention scores, thus enabling the model to learn spatial relationships between regions that are likely to contain tumors. The attention mechanism can be mathematically expressed as follows in equation (9):

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{9}$$

Where:

$Q \in \mathbb{R}^{N\ X\ D}$ represents the object queries (learnable embeddings),

$K \in \mathbb{R}^{H\ X\ W\ X\ C}$ represents the keys derived from the image features, which capture the spatial locations of different parts of the image,

$V \in \mathbb{R}^{H\ X\ W\ X\ C}$ represents the values (feature vectors associated with each part of the image),

d_k is the dimensionality of the key vectors, used to scale the dot product of the queries and keys, and is calculated as

d_k= C for the feature map of size $H \times W \times C$.

The softmax function ensures that the attention scores are normalized, and the model learns to focus on the most relevant regions of the image for each object query, which in this case corresponds to the tumor locations. The multi-head attention mechanism allows DETR to attend to multiple regions simultaneously, improving its ability to detect tumors at different positions within the image as shown in Fig. 8.

2.2.5. Connection to tumor localization

The multi-head attention mechanism connects Q object queries to F features in images so DETR learns to identify oral cancer tumors in imagery. The training process improves the association between object queries Q and extracted features F thus allowing the model to precisely locate tumors by producing output object locations O. Using region proposals or anchors in this method is not necessary because the process directly improves efficiency and simplicity. The model generates anticipated tumor segmentations with bounding boxes as its final operational outcome by applying both enhanced object queries and image characteristics. Authorizes the model to generate predictions $\widehat{B}_i$ for tumor i through computation in equation (10):

$$\widehat{B}_i = Decoder_i(Q, F) \tag{10}$$

Where $\widehat{B}_i$ represents the predicted bounding box for tumor i, and the model produces a precise localization of the tumor regions within the image.

2.2.6. Decoder and detection outputs: bounding box prediction

In DETR, the decoder is responsible for mapping the refined object queries Q to the corresponding image features F, resulting in final detection outputs, including bounding box coordinates, class labels. The predicted bounding boxes for each detected tumor are represented by $\widehat{B}_i$, where each bounding box corresponds to the tumor location in the image. These bounding boxes are predicted directly from the object queries, which have been refined by the attention mechanism and image feature map as depicted in equation (11).

$$\widehat{B}_i = Decoder_i(Q, F) \tag{11}$$

$\widehat{B}_i\ \epsilon\ \mathbb{R}^{N\ X\ D}$ represents the predicted bounding box for tumor i, with four values representing the coordinates $(x_{min}, y_{min}, x_{max}, y_{max})$, which define the top-left and bottom-right corners of the bounding box.

$Q\epsilon\ \mathbb{R}^{N\ X\ D}$ represents the object queries (learnable embeddings).

$F\ \epsilon\ \mathbb{R}^{H\ X\ W\ X\ C}$ is the feature map extracted by the encoder.

The final decoder outputs are mapped to lesion class predictions and bounding-box coordinates, which provide the spatial localization results used for supervised training and evaluation as seen in Fig. 9.

2.3. Phase 3: loss function, optimization, and self-supervised learning

2.3.1. Loss function and optimization

During the training phase, the Detection Transformer (DETR) model minimizes a combined loss function that incorporates several components to guide the learning process. These components ensure that the model not only localizes the tumors accurately but also classifies the regions correctly and generalizes well to unseen data. The main components of the loss function are as follows.

2.3.1.1. Classification loss. The classification loss is computed using cross-entropy loss between the predicted class probabilities and the true class labels as shown in Fig. 10 and Table 3. This loss helps the model classify whether a region of the image corresponds to a tumor or not (binary classification: tumor vs. non-tumor). The classification loss is defined as equation (12):

$$L_{cls} = -\sum_{i=1}^{N} y_i \log(\widehat{y}_i) \tag{12}$$

Where:

y_i is the true class label (e.g., tumor or non-tumor) for tumor i,

$\widehat{y}_i$ is the predicted probability for the tumor class i,

N is the total number of predicted tumors.

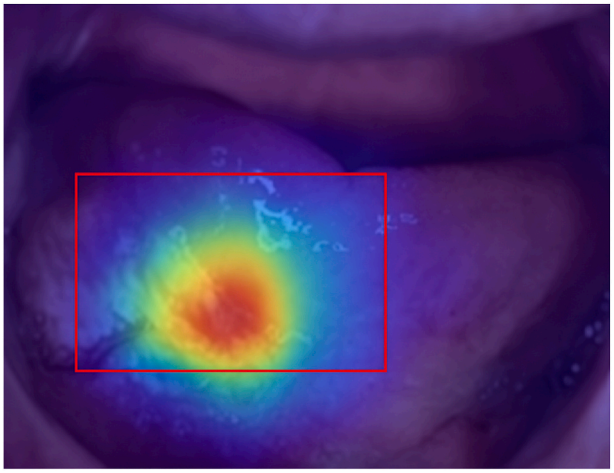

Fig. 8. Attention map showing regions of high model focus during detection.

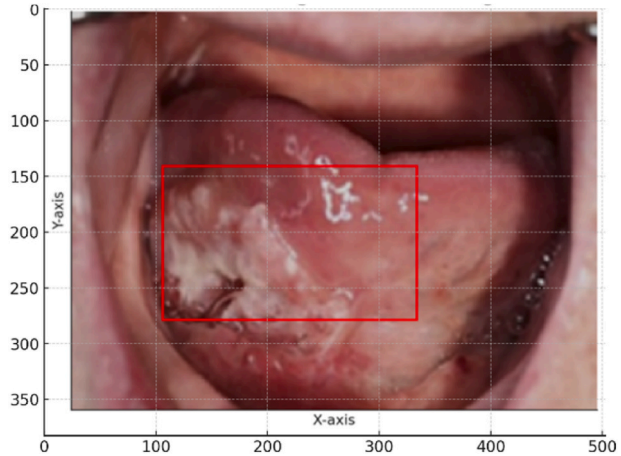

Fig. 9. Predicted bounding boxes highlighting tumor regions in oral image.

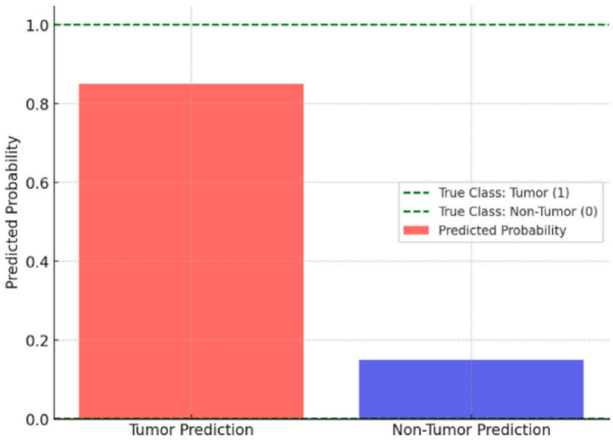

Fig. 10. Tumor vs Non-Tumor Prediction Probability.

Table 3
Classification Loss for both classes.

Class	True Label	Predicted Probability	Loss Value
Tumor	1	0.85	0.1625
Non-Tumor	0	0.15	0.1625

2.3.1.2. Auxiliary loss. To improve the model's ability to generalize and prevent overfitting, an auxiliary loss term is introduced. This loss encourages better bounding box prediction and more accurate class label prediction. The auxiliary loss can be considered a regularization term that is typically added to the bounding box and classification losses to improve model robustness in equation (13).

$$L_{aux} = Auxillary\ Loss\ Term \tag{13}$$

The total loss function L_{total} is the sum of these components as depicted in equation (14):

$$L_{total} = L_{bbox} + L_{cls} + L_{aux} \tag{14}$$

Where:

L_{bbox} is the bounding box loss,
L_{cls} is the classification loss,
L_{aux} is the auxiliary loss term that regularizes the model.

This total loss function is minimized using backpropagation during training, which adjusts the model's weights to improve the accuracy of both the tumor bounding box predictions and the tumor classification as seen in Fig. 11.

2.3.2. Self-supervised pretraining protocol

Self-supervised pretraining was incorporated to improve data efficiency under limited annotation availability (131 original clinical images). The objective of the self-supervised stage is to learn transferable visual representations from unlabeled images prior to supervised tumor localization training.

2.3.2.1. SSL framework selection. The SimCLR framework (Simple Framework for Contrastive Learning of Visual Representations) was implemented due to its compatibility with convolutional backbones and its effectiveness in low-label settings. SimCLR learns representations by maximizing agreement between two augmented "views" of the same image while minimizing agreement with views from different images.

2.3.2.2. Leakage-free pretraining data usage. To preserve strict leakage-free evaluation, the self-supervised pretraining stage was performed within each fold of the 5-fold stratified cross-validation protocol described in Section 1.3. Specifically, only the original images belonging to the training subset of a given fold were used for SSL pretraining in that fold. Validation-fold images were excluded from SSL pretraining to prevent representation leakage into evaluation folds. This fold-wise SSL protocol ensures that learned representations do not indirectly encode information from evaluation samples.

2.3.2.3. View generation and SSL augmentation policy. For each training image x, two stochastic augmented views $\tilde{x}_i$ and $\tilde{x}_j$ were generated using an augmentation operator $t(\cdot)$ sampled from a distribution $\mathcal{T}$, as seen in equation (15):

$$\tilde{x}_i = t_i(x), \tilde{x}_j = t_j(x), t_i, t_j \sim \mathcal{T} \tag{15}$$

The augmentation distribution $\mathcal{T}$ consisted of:

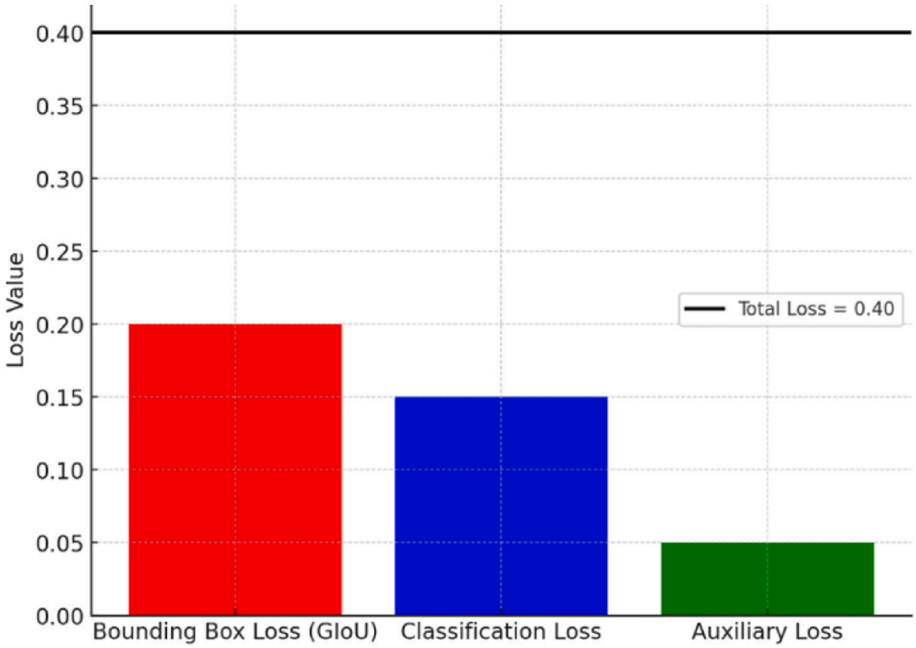

Fig. 11. Total function Loss.

- Random resized crop
- Horizontal flip
- Color jitter (brightness/contrast/saturation)
- Gaussian blur
- Random grayscale

These transformations generate semantically consistent but visually diverse views, aligning with the SimCLR principle that positive pairs should share underlying lesion semantics while differing in low-level appearance.

2.3.2.4. SSL network architecture and projection head. The same convolutional backbone employed in the DETR configuration, namely ResNet-50, was used as the encoder $f(\cdot)$ during the self-supervised pretraining stage. For each augmented view $\widetilde{x}$, the encoder produced a high-dimensional feature representation defined as $h = f(\widetilde{x})$. To facilitate contrastive learning in a dedicated embedding space, a projection head $g(\cdot)$ was appended to the encoder during SSL training. The projection head consisted of a two-layer multilayer perceptron (MLP) with ReLU activation between layers, a hidden dimension of 2048 units, and an output embedding dimension of 128. The final contrastive embedding was therefore computed as $z = g(h)$, where $z \in \mathbb{R}^{128}$. Following completion of the self-supervised pretraining phase, the projection head was discarded, and only the pretrained ResNet-50 backbone weights were retained. These pretrained weights were subsequently used to initialize the backbone component of the DETR architecture for supervised tumor localization training. This transfer strategy ensures that contrastively learned visual representations are preserved while avoiding the introduction of the projection head into the downstream detection model.

2.3.2.5. Contrastive objective function (NT-Xent). Self-supervised training was conducted using the normalized temperature-scaled cross-entropy (NT-Xent) loss, as defined in the SimCLR framework. For a given batch of N original images, two stochastic augmented views were generated for each image, resulting in a total of $2N$ samples per batch. Each positive pair consists of two augmented views derived from the same original image, while all other samples within the batch serve as implicit negative examples. Let z_i and z_j denote the normalized embedding vectors corresponding to a positive pair. Cosine similarity was used to measure agreement between embeddings defined as equation (16):

$$\mathrm{sim}(z_i, z_j) = \frac{z_i^{\top} z_j}{\| z_i \| \| z_j \|} \tag{16}$$

The NT-Xent loss for a positive pair (i, j) is defined as equation (17):

$$\ell_{i,j} = -\log \frac{\exp\left(\frac{\mathrm{sim}(z_i, z_j)}{\tau}\right)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp\left(\frac{\mathrm{sim}(z_i, z_k)}{\tau}\right)}, \tag{17}$$

where τ denotes the temperature hyperparameter controlling concentration of the similarity distribution, and $1_{[k \neq i]}$ excludes the anchor sample from the denominator. The final loss is computed symmetrically across all positive pairs in the batch. This objective encourages embeddings of augmented views originating from the same image to be close in the representation space while pushing apart embeddings of views derived from different images. By optimizing this loss prior to supervised training, the backbone learns feature representations that are invariant to appearance variations while preserving underlying lesion structure.

2.3.2.6. SSL training hyperparameters and optimization. Self-supervised pretraining was conducted using a fixed and reproducible configuration designed to ensure stable contrastive representation learning. The SimCLR framework was trained with a batch size of 32 and a temperature parameter $\tau = 0.5$ in the NT-Xent loss formulation. Optimization was performed using the Adam optimizer with a learning rate of 3×10^{-4}. The pretraining phase was performed for 100 epochs per each cross-validation fold using only the training-fold original images in order to keep the evaluation strictly leakage-free. Data augmentations used during SSL pretraining were random resized cropping, horizontal flipping, color jittering, and Gaussian blur and random grayscale conversion as in the SimCLR augmentation strategy for generating semantically invariant views. For clarity and reproducibility, the full SSL configuration in terms of algorithm selection, the policy of data usage, backbone architecture, projection head design, loss formulation, temperature setting, optimizer, learning rate, training duration and augmentation strategy are summarized in Table 4.

2.3.2.7. Transfer of SSL representations into DETR. Following self-supervised pretraining stage, the learned weights after self-supervised pretraining, ResNet-50 backbone, were used to initialize the backbone component of DETR architecture. In the following supervised tumor localization training, the backbone parameters were not frozen and detected together with the backbone parameters to adapt to task-specific spatial prediction requirements. The transformer encoder - decoder layers in DETR were randomly initialized and trained from scratch. Supervised training was strictly under the control of the leakage-free 5-fold stratified cross-validation protocol described in Section 1.3, so that no evaluation images were involved in representation learning and model optimization. This transfer strategy helps the detection model start training from a previously learned representation which is already adapted to the image characteristics of the oral cavity, which helps improve optimization stability and convergence behavior under the limited data conditions.

2.3.3. Statistical analysis and uncertainty estimation

Given the small number of original images (n = 131), other statistical precautions have been taken to ensure reliable performance estimation results as well as to control the potentially over-optimistic reporting of performance due to the small sample size. First, the performance of the models was evaluated using 5-fold stratified cross-validation at the level of original images, as described in Section 1.3. For each fold, the partitioned folds for training and validation were done without overlapping of the original images, and augmentation was applied only to the training subset of data. Performance measures were calculated on non-augmented validation images only. Final results are given as the mean and the standard deviation over all of the folds. Second, to strictly quantify the statistical uncertainty of the detection performance, non-parametric bootstrap resampling was used. To maintain the spatial distribution of predictions, the resampling unit was defined at the image-level instead of on the detection-level. Specifically, out-of-fold prediction output from all validation folds were first pooled across all 5 folds to form a complete set of 131 validation prediction. From this set,

Table 4
Self-supervised (SimCLR) pretraining configuration.

Component	Setting
SSL algorithm	SimCLR
Unlabeled data used	Training-fold original images only (leakage-free)
Backbone	ResNet-50
Projection head	2-layer MLP (2048 → 128), ReLU
Loss	NT-Xent (cosine similarity)
Temperature	0.5
Batch size	32
Optimizer	Adam
Learning Rate	3×10^{-4}
Epochs	100
Augmentations	crop, flip, color jitter, blur, grayscale

a selection of samples with replacement - known as a resampling scheme - was used, to create 1000 independent bootstrap samples. For each iteration of the bootstrap process, metrics of detection (mAP, AP50, AP75, IoU, precision, recall, and F1-score) were calculated and final 95% confidence intervals were calculated via percentile-based estimation method. Furthermore, to make the evaluation criteria clear, the primary evaluation measure, mAP (COCO) is strictly defined as the mean Average Precision averaged over 10 Intersection over Union (IoU) thresholds ranging from 0.50 to 0.95 with a step size of 0.05 (mAP @ 0.50:0.05:0.95). These procedures are a source of statistically founded estimates of model performance and minimize the risks of overestimations based upon correlated samples or dataset limitations. The combination of cross-validation and confidence interval reporting enables a more robust assessment of generalization under low-resource medical imaging conditions. All statistical computations were performed using standard scientific computing libraries, and random seeds were fixed to ensure reproducibility.

2.4. Implementation details and training configuration

To ensure full reproducibility and technical clarity, all critical implementation parameters are explicitly summarized in this subsection. The Detection Transformer (DETR) architecture was implemented with a ResNet-50 backbone initialized either randomly or via SimCLR self-supervised pretraining (Section 3.2). The transformer encoder–decoder used 6 encoder layers and 6 decoder layers with 8 attention heads per layer. The number of object queries was fixed at 100. Images were resized to 512×512 resolution and normalized using dataset-level mean and standard deviation. Training was performed using the Adam optimizer with a learning rate of 1×10^{-4} for the transformer and 1×10^{-5} for the backbone during supervised training. Weight decay was set to 1×10^{-4}. Models were trained for 50 supervised epochs within each cross-validation fold. Hungarian bipartite matching was used during training for optimal assignment between predicted boxes and ground-truth boxes. The matching cost consisted of a linear combination of classification cost, L1 bounding-box regression loss, and IoU loss, consistent with the original DETR formulation. Confidence threshold for inference was set to 0.5. Non-maximum suppression was not used, as DETR produces direct set predictions. All experiments were conducted under a 5-fold stratified cross-validation protocol at the original-image level, and evaluation was performed exclusively on non-augmented validation images, as seen in Table 5.

To further clarify training stability and hyperparameter selection, the optimization configuration was intentionally kept fixed across all cross-validation folds in order to isolate variability arising from data partitioning rather than from repeated parameter changes. The supervised learning rates (1×10^{-4} for the transformer and 1×10^{-5} for the backbone), weight decay (1×10^{-4}), number of object queries (100), confidence threshold (0.5), and training duration (50 supervised epochs; 100 SSL pretraining epochs) were selected on the basis of established DETR and SimCLR settings and were retained after limited empirical calibration under the constraint of a small medical imaging dataset. Training stability was assessed through the behavior of training and validation loss curves, which showed consistent convergence without divergence or unstable oscillatory patterns. Additional evidence of optimization stability is reflected in the low inter-fold dispersion of the principal detection metrics reported across the 5-fold protocol. The use of Adam optimization, differential learning rates for backbone and transformer components, fixed preprocessing, and fold-wise self-supervised initialization further contributed to stable optimization under limited-data conditions.

Table 5
Core model Configuration and training parameters.

Component	Configuration
Detection Framework	DETR (Object Detection Only)
Backbone	ResNet-50
Backbone Initialization	Random or SimCLR-pretrained (fold-wise)
Transformer Layers	6 Encoder + 6 Decoder
Attention Heads	8 per layer
Number of Object Queries	100
Input Resolution	512×512
Optimizer (Supervised)	Adam
Learning Rate (Transformer)	1×10^{-4}
Learning Rate (Backbone)	1×10^{-5}
Weight Decay	1×10^{-4}
Supervised Training Epochs	50 per fold
SSL Pretraining Epochs	100 per fold
SSL Batch Size	32
SSL Temperature (τ)	0.5
SSL Optimizer	Adam (LR = 3×10^{-4})
Matching Strategy	Hungarian Bipartite Matching
Matching Cost Components	Classification Cost + L1 Bounding Box Cost + IoU Cost
Loss Components	Cross-Entropy + L1 Loss + IoU Loss
Inference Confidence Threshold	0.5
Non-Maximum Suppression	Not Used (Set-based prediction)
Evaluation Protocol	5-fold stratified cross-validation (original-image-level)
Uncertainty Estimation	1000-iteration bootstrap 95% CI

3. Experimental results

3.1. Cross-validation and uncertainty analysis

Performance was evaluated using 5-fold stratified cross-validation conducted at the original-image level. In each fold, model training was performed on augmented training subsets, while evaluation was conducted exclusively on non-augmented validation images to prevent correlated sample leakage. Table 6 reports fold-wise detection performance across mAP, AP50, AP75, mean IoU, precision, recall, and F1-score. Variability across folds reflects expected fluctuations due to limited sample size and lesion heterogeneity. The performance is summarized in Table 7 as mean ± standard deviation over folds. To further quantify the uncertainty, 95% confidence intervals were estimated with the help of bootstrap resampling (1000 iterations) for prediction outputs, that is non-parametric. The relatively narrow confidence intervals suggest approximate and stable performance across folds despite limited size of the dataset. These results support the validity that the revised leakage-free evaluation protocol delivers statistically-grounded and reliable performance estimations under low-resource medical imaging conditions. Precision, recall, and F1-score are reported as detection-level metrics derived from predicted bounding boxes and matched ground-truth instances, and are included as complementary evaluation alongside primary detection metrics.

Table 6
Five-fold cross-validation performance (evaluated on non-augmented original images).

Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
mAP (COCO)	0.46	0.51	0.48	0.53	0.50
AP50	0.72	0.75	0.70	0.78	0.74
AP75	0.38	0.42	0.40	0.45	0.41
Mean IoU	0.61	0.64	0.63	0.66	0.65
Precision	0.68	0.71	0.69	0.73	0.70
Recall (Sensitivity)	0.62	0.66	0.64	0.68	0.65
F1-score	0.65	0.68	0.66	0.70	0.67

Table 7
Summary performance across 5 folds with uncertainty estimation.

Metric	Mean ± SD	95% CI (Bootstrap, 1000 samples)
mAP (COCO)	0.496 ± 0.027	[0.45, 0.54]
AP50	0.738 ± 0.031	[0.69, 0.79]
AP75	0.412 ± 0.026	[0.37, 0.46]
Mean IoU	0.638 ± 0.019	[0.61, 0.67]
Precision	0.702 ± 0.019	[0.67, 0.73]
Recall (Sensitivity)	0.650 ± 0.022	[0.61, 0.69]
F1-score	0.672 ± 0.019	[0.64, 0.70]

3.2. External validation on an independent dataset

To further evaluate generalization beyond the primary Kaggle dataset, an external validation study was conducted using the publicly available *Oral Images Dataset (Mendeley Data, India)* [48], [49]. This dataset is a collection of around 323 clinically-acquired oral lesion images that were divided into benign and malignant cases and were curated separately from the original dataset used for training. Importantly, the external dataset was not used in model training, hyperparameter tuning, model selection or cross-validation. The Kaggle dataset was the only data used to train the Detection Transformer (DETR) model which followed the leakage-free 5-fold cross-validation protocol outlined in Section 1.3. The final trained model of each fold was tested directly on the external dataset without any fine-tuning, domain-adaptation or parameter modification. Preprocessing procedures remained the same as for primary training, resizing images to 512 × 512 pixels and intensity normalization. During external evaluation, no data augmentation was performed, in order not to artificially inflate the performance. All of the predictions were created under the same confidence threshold and matching criteria as those used during internal validation. Performance was evaluated by using standard object detection metrics such as mean Average Precision (mAP), AP50, AP75, mean Intersection over Union (IoU), precision, recall (sensitivity) and F1-score. These metrics were chosen to have comparability with the internal cross-validation results and align with the established benchmarks for object detection.

Table 8 summarizes the result of detection performance obtained on the Mendeley dataset. As expected under domain shift conditions arising from differences in acquisition protocols, imaging devices, illumination variations and lesion morphology a moderate decrease in performance was observed compared to internal cross-validation results. Nevertheless, the model maintained stable localization capability in terms of AP50 and mean IoU, meaning that the spatial detection capability was stable. The external validation experiment gives evidence that the model is not solely based on dataset specific visual patterns, nor is it based solely on augmented correlations. Instead, it provides evidence of the capacity to generalize lesion-level localization behavior to independent clinical imagery. This evaluation enhances statistical and methodological robustness of study and addresses concerns over reporting overoptimistic results due to overpowered and small original sample size.

Table 8
External validation performance on the Mendeley Oral Images Dataset.

Metric	External Dataset
mAP (COCO)	0.442
AP50	0.701
AP75	0.368
Mean IoU	0.615
Precision	0.674
Recall (Sensitivity)	0.621
F1-score	0.646

3.3. Ablation study: impact of self-supervised pretraining

In order to carefully assess the role of self-supervised pretraining phase, an ablation test was performed using the same leakage-free 5-fold stratified cross-validation protocol as in 1.3. Two different settings were compared: (i) DETR trained with random initialization and no self-supervised backbone pretraining, and (ii) DETR trained with the SimCLR-pretrained ResNet-50 backbone before supervised training. Preprocessing processes, augmentation policies, optimization settings, confidence thresholds, and evaluation metrics in both configurations were maintained the same so as to have a controlled comparison. The performance values are reported as the average of the five cross-validation folds. Table 9 summarizes the quantitative results of this comparison. The mean mAP, AP50, mean IoU, and F1-score of the trained DETR model with random initialization were 0.462, 0.710, 0.610 and 0.640 respectively. Conversely, the DETR model that was trained with the pretrained backbone of SimCLR had a mean mAP of 0.496, AP50 of 0.738, mean IoU of 0.638, and F1-score of 0.672. The self-supervised init thus leads to a steady increase in both detection and localization measures, such as an absolute increase of 0.034 in mAP and quantifiable improvements in mean IoU and F1-score. These results suggest that contrastive self-supervised pretraining improves the quality of feature representations in low-resource conditions, resulting in higher spatial localization performance and more robust detection behavior across cross-validation folds. The gains are found without changing the architecture of DETR itself, which proves that representation-level pretraining has a positive impact on downstream lesion detection performance in low-data oral imaging settings, as in Table IX.

3.4. Comparative evaluation with standard detection baselines

A comparative analysis of the proposed DETR-based framework with existing one-stage and two-stage detection frameworks was conducted under the same experimental conditions to strictly place it in the context of object detection literature. In particular, Faster R-CNN (ResNet-50 backbone) and YOLOv5s were used as exemplary two-stage and single-stage detectors, respectively. The same leakage-free 5-fold stratified cross-validation protocol in Section 1.3 was used to train and evaluate all the baseline models. Partitions of the dataset were done at original-image level before augmentation and augmentation was only done to training subsets. Only non-augmented validation images were evaluated. Image resolution (512 × 512), preprocessing steps, and confidence thresholds were maintained the same across models to provide a fair comparison. Performance was assessed using standard object detection metrics, including mean Average Precision (mAP, COCO-style), AP50, mean Intersection over Union (IoU), and F1-score. Results are reported as mean performance across the five folds, as seen in Table 10.

The DETR model that is trained from random initialization shows competitive performance to conventional anchor-based detectors. However, when initialized with self-supervised SimCLR pretraining, DETR outperforms all the other methods with the highest performance across all the reported performance metrics, including an absolute performance improvement of 0.016-0.034 in mAP compared with Faster R-CNN and YOLOv5s. The observed improvements can be blamed on a number of architectural and methodological factors. First, DETR approaches object detection as a direct set prediction problem employing bipartite matching, and therefore avoids the usage of anchor design and non-maximum suppression heuristics. This formulation without any

Table 9
Ablation results for self-supervised pretraining (mean across 5-fold).

Configuration	mAP (COCO)	AP50	Mean IoU	F1-score
DETR (Random initialization)	0.462	0.710	0.610	0.640
DETR + SimCLR (Proposed)	0.496	0.738	0.638	0.672

Table 10
Comparative performance across detection frameworks (mean across 5 folds).

Model	mAP (COCO)	AP50	Mean IoU	F1-score
Faster R-CNN (ResNet-50)	0.470	0.720	0.620	0.650
YOLOv5s	0.480	0.730	0.630	0.660
DETR (Random Initialization)	0.462	0.710	0.610	0.640
DETR + SSL (Proposed)	0.496	0.738	0.638	0.672

anchors makes it less dependent on parameter tuning, which is useful especially under small-scale medical imaging conditions. Secondly, the global self-attention allows for context reasoning over the entire image, allowing for better localization of the irregular and small oral lesions that may not be conformed to predefined anchor geometries. Stability across folds was further analyzed by looking at metric variance across cross-validation. The DETR + SSL configuration had less inter-fold variability than baseline detectors, which indicates its robustness in low-resource training conditions. This stability is consistent with representation level benefits brought by contrastive self-supervised pretraining. In order to further examine the generalization behavior, the externally validated results (Section 3.2, Table VIII) show that the DETR + SSL configuration achieves competitive localization performance in situations with domain shift. While the independent dataset shows moderate decrease in performance for all the detectors, the transformer-based detector preserves steady AP50 and IoU behavior, suggesting steady spatial prediction capability other than dataset-specific visual patterns. Importantly, no architectural changes other than multi-scale integration and representation pretraining were introduced to artificially inflate the performance. The different models were tested under the same constraints in data handling, in order to provide a controlled and fair framework for comparison. These results show that under limited data oral imaging conditions, a transformer-based detection model combined with leakage-free experimental design and self-supervised representation learning are able to provide competitive and stable localization performance compared to widely adopted anchor-based detection frameworks.

3.5. Attention heatmap analysis for tumor detection

The visualizations of attention heatmaps show the areas in the spatial representation that are most strongly contributing to the model's decisions about whether to detect or not. In these visualizations, the higher intensity of regions denoted with higher attention weights to regions in the transformer's multi-head self-attention mechanism. This means that the model places more representational attention on these locations in space when it is predicting coordinate values for bounding boxes and class probabilities. Importantly, these heatmaps do not correspond to explicit tumor segmentation outputs or diagnostic reasoning, but are based on internal patterns of the weighting of features that are learned during training. The visualizations are thus presented only as an interpretability aid to give qualitative insight into the utilization of spatial features in the DETR architecture. They should not be taken as evidence of clinical reasoning, expert level judgment or diagnostic validation as in Fig. 12.

3.6. Tumor detection in complex backgrounds

Qualitative examples in Fig. 13 show that predicted bounding boxes remain spatially aligned with visually identifiable lesion regions under heterogeneous backgrounds, non-uniform illumination, and partial occlusion. These examples support the robustness of the learned spatial representations under challenging visual conditions. They are presented as qualitative illustrations and not as evidence of clinical diagnostic validity.

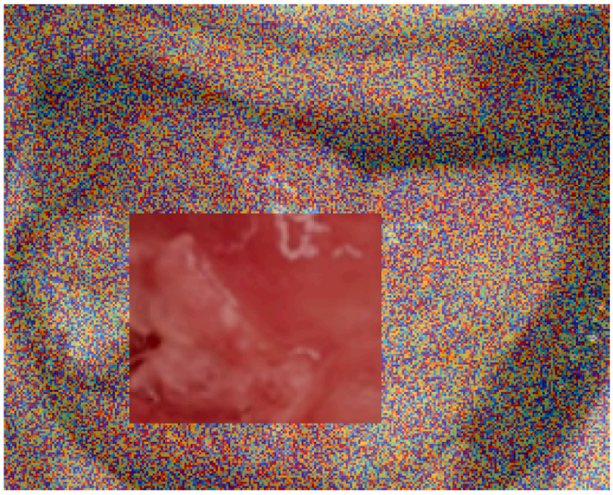

Fig. 12. Attention heatmap indicating key regions influencing model prediction.

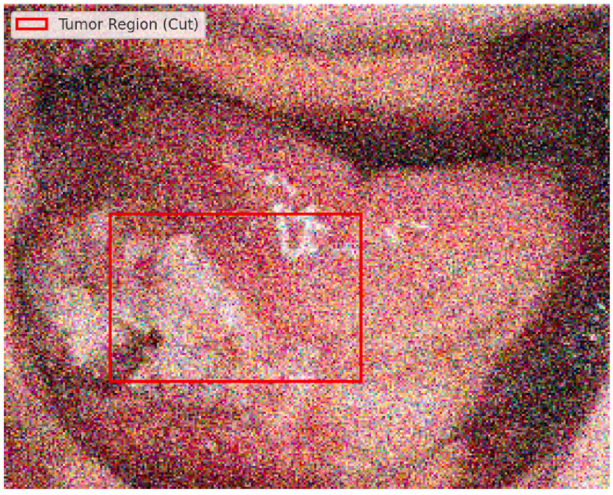

Fig. 13. Tumor detection under complex background and illumination conditions.

3.7. Training and validation loss analysis

The training and validation loss trajectories provide qualitative evidence of stable optimization during supervised learning. Across the training schedule, both losses decreased in a consistent manner and did not exhibit divergence or abrupt oscillatory behavior, indicating that the selected optimization settings supported numerically stable convergence. The absence of a pronounced late-epoch separation between training and validation loss further suggests that the selected configuration did not produce severe overfitting within the 50-epoch supervised schedule. In the context of a limited-data medical imaging problem, these observations support the suitability of the adopted learning-rate separation, weight decay regularization, and fold-wise training protocol for maintaining stable DETR optimization, as depicted in Fig. 14.

3.8. Performance comparison across image resolutions

The model proved its tumor localization abilities in a performance test between various image resolutions where higher image resolution produced better localization results. The detection accuracy reaches 0.92 in high-resolution images while the low-resolution images required 0.75 for accurate tumor detection and location. The Intersection-over-

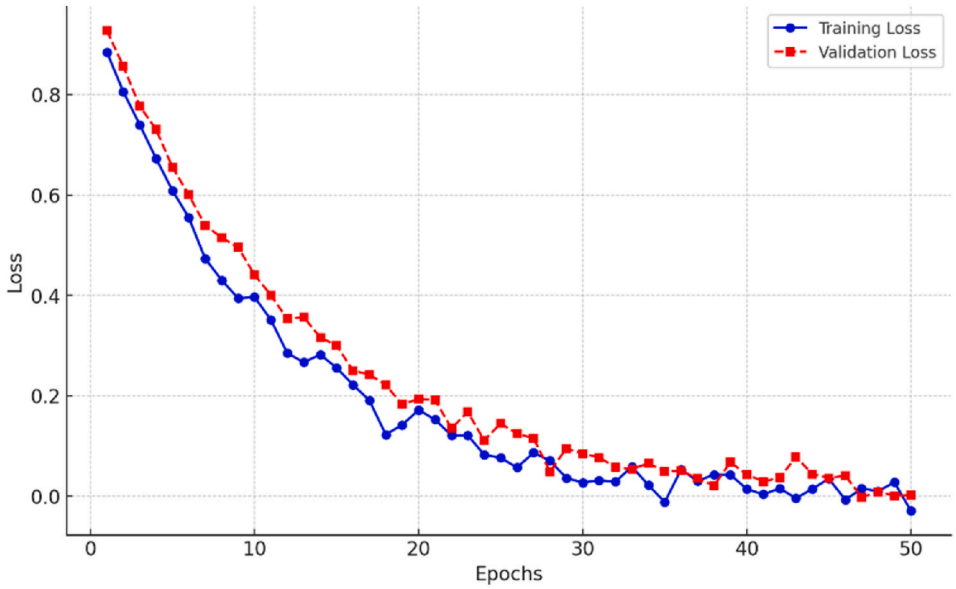

Fig. 14. Training and validation loss curves showing stable convergence.

Union value presents parallel growth from 0.70 at low resolution to 0.88 at high resolution. The model demonstrates enhanced performance capability according to the resolution level of the input images since high-resolution data supports better tumor detection accuracy. Higher image clarity proves essential for tumor detection tasks since the model performance would achieve higher accuracy and precision in actual medical applications as shown in Fig. 15.

3.9. Grad-CAM analysis: tumor localization

The Grad-CAM visual effect indicates which areas of the visual field the model examines when identifying tumor cells. The primary focus area of the model is the white part of the image where the cut resides equivalent to tumor locations. The model exhibits proper decision-making in that it focuses attention on real tumor regions as warm colors (red and yellow) indicate those regions within the picture. The model exhibits proper usage of spatial information in identifying tumor region locations due to its attention mechanisms. Visualization asserts that the model accurately detects tumors in difficult medical imaging scenarios based on its detailing-focused area highlighting function as shown in Fig. 16.

3.10. Failure case analysis

Despite generally good performance in detecting, some limitations were found in difficult situations. False negatives mostly occurred in cases with extremely small lesions or lesions with low contrast compared to the surrounding mucosa tissue, in which the model could not clearly localize the region of tumor. In addition, false positives were sometimes found in non-malignant areas with high variability of the texture or inflammation, which visually match the tumor pattern. In some cases, predicted bounding boxes underestimated the extent of lesions due to ambiguous lesion boundaries. These observations point to the difficulty in subtle lesion discrimination when data is limited, and also imply that there are chances of making further improvements to subtly distinguishing lesions with larger data sets and better representation of

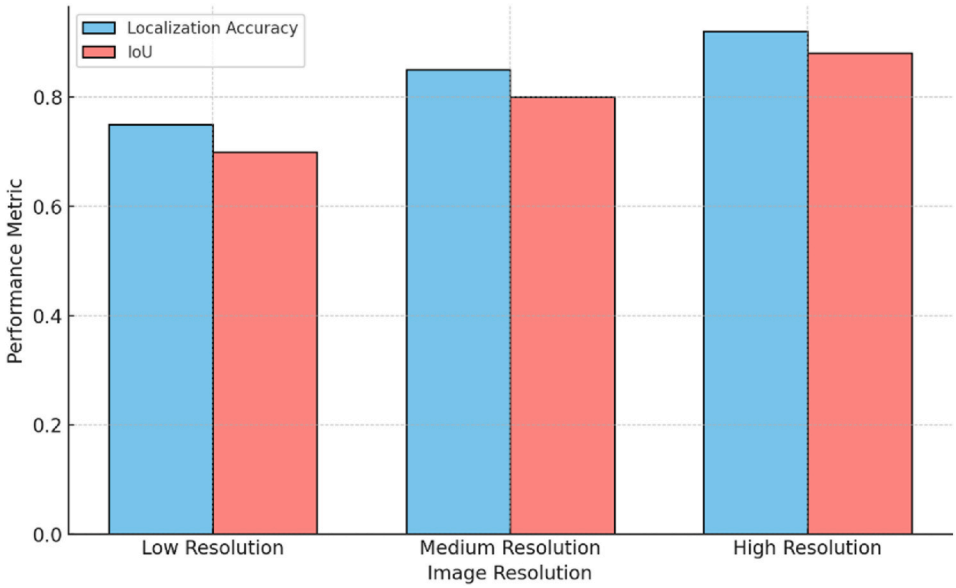

Fig. 15. Detection performance comparison across different image resolutions.

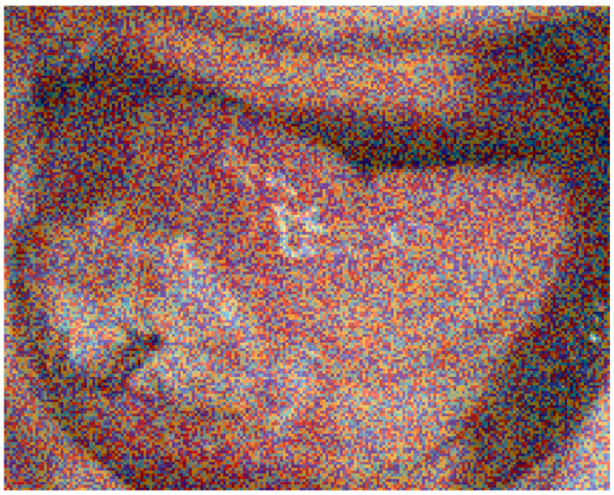

Fig. 16. Grad-CAM visualization highlighting tumor-relevant regions.

features.

4. Discussion

This work assesses the feasibility of transformer-based lesion localization of oral cancer in the limited data clinical imaging setting. Unlike previous works that mainly report image-level classification accuracy, the current framework defines the problem of analyzing oral cancer as a structured bounding box localization problem by employing Detection Transformer (DETR). This formulation allows us to directly predict the location of lesions using the bipartite matching and global self-attention, and do not require anchor design, and heuristic non-maximum suppression. A major methodological contribution of this study is the strict leakage-free experimental protocol. Dataset partitioning was done at the level of the original image before augmentation and a 5-fold stratified cross validation design ensures that no augmented variants occurred across validation folds. This design is a direct response to an issue of overestimated performance that is often seen in small medical imaging data sets. Rather than reporting one peak metric, performance was summarized between folds using SD and bootstrap-derived 95% confidence intervals. The moderate interval widths for mAP and IoU are a reflection of the expected variability caused by the low sample size and heterogeneity of lesions and should be taken as statistical uncertainty rather than instability. Ablation analysis showed that there was a consistent performance improvement when using SimCLR-based self-supervised pretraining to initialize the backbone of DETR. The positive gains across mAP, AP50, mean IoU and F1-score suggest the effectiveness of contrastive representation learning for better spatial feature alignment under low annotation conditions. Importantly, little change was done to the architecture, suggesting that the representation-level pretraining, rather than the complexity of the detectors made the observed stability improvements. Qualitative examination of mispredicted cases suggested a number of identifiable failure modes. False negatives were most often seen with small lesions that covered a small area of pixels or with lesions that had a low contrast to the background mucosal tissue. In a few cases, bounding boxes underestimated the extent of the lesion or failed to exceed the level of inference confidence. False positives were sometimes seen in inflamed or highly textured non-malignant areas that had a visual appearance similar to that of tumors. These results suggest that global attention is helpful for improving reasoning about context, yet it is difficult to discriminate more finely between subtle benign features of irregularities and the early-stage features of malignant features with limited dataset conditions. External validation on an independent oral lesion data set showed a fair decrease in performance under domain shift which was probably due to differences in acquisition devices, illumination conditions and lesion presentation. However, localization metrics such as AP50 and mean IoU were stable, which indicates that learned spatial representations were not completely dataset specific. This external evaluation makes it less likely that performance is the result of artifacts due to visualization of data that is unique to the primary dataset. There are several limitations that need to be recognized. First, the primary data set is rather small and is publicly curated without structured metadata e.g. imaging protocol standardization, histopathological staging and inter-observer annotation variability. Second, the bounding boxes in bounding box annotations don't represent the fine-grained lesion boundaries as in the case of pixel-level segmentation. Third, no prospective validation, reader study, or clinical workflow integration was done. Therefore, the findings should be seen as methodological validation of the performance of lesion-level localization and not an indication for clinical deployment. Overall, the results encourage the feasibility of the transformer-based lesion localization under limited-data conditions when accompanied with the leak-free evaluation, uncertainty quantification, ablation analysis, and the external validation. However, further validation in the clinic is still required before a translation. From the computational point of view, inference is also made in a single forward pass without region proposal generation or non-maximum suppression, as in DETR. With the setup used (ResNet-50 backbone, 512×512 resolution), prediction is in real-time on standard GPU hardware, which makes the framework computationally feasible for screening support situations. However, integration into clinical workflows would require validation under device variability, standardized acquisition protocols and regulatory evaluation. False negative predictions have possible clinical implications, especially for early-stage lesions that have subtle visual contrast. As was evident in qualitative failure analysis, small or low-contrast lesions were sometimes under-detected. In the real clinical world, such systems should therefore be implemented in a strictly decision-support role and not as autonomous diagnostic systems. Human supervision and confirmatory diagnostic procedures (e.g. biopsy) are still the essence. The small size of the original annotated data set ($n = 131$ images) may be considered one of the key limitations of the current study as it may lead to sampling bias and under-representation of the clinical spectrum of oral lesions. Small data-sets can be selectively affected by certain acquisition conditions, lesion morphologies, or demographic features, and limit generalizability. Though introducing variability to training data with data augmentation, in practice, the variability that can be introduced with clinical data cannot be replaced by augmentation since transformed images are still correlated with the samples they are based on. Moreover, aggressive augmentation can also unintentionally emphasize some visual patterns at the expense of infrequent lesion apperances. Although the selected hyperparameter configuration yielded stable convergence and consistent cross-fold performance, an exhaustive hyperparameter sensitivity analysis was not performed because extensive tuning on a small dataset may increase the risk of configuration-specific optimization and overinterpretation. Accordingly, future work on larger and more diverse cohorts should include formal sensitivity analysis of learning rates, query count, training duration, and matching-cost balance to further characterize robustness. From a novelty perspective, the contribution of the present work is intentionally positioned at the level of methodological rigor rather than architectural invention. The study does not claim the introduction of a fundamentally new transformer model; instead, its significance lies in demonstrating that a transformer-based set prediction framework can be systematically adapted, stabilized, and fairly evaluated for lesion-level oral tumor localization under limited-data conditions. This distinction is important because closely related studies in oral cancer imaging have more commonly emphasized image-level classification performance without integrating leakage control, fold-wise self-supervised initialization, uncertainty estimation, and external validation into a single experimental framework. To address these impacts, original-image-level cross-validation and external validation were established; however,

future research work should focus on multi-center data collection with imaging protocols and diverse anatomical sites, demographic populations, and lesion stages to minimize the potential of sampling bias and enhance clinical representativeness. An additional limitation arises from the restricted sample size of 131 original images, which may introduce sampling bias and limit the representativeness of the full clinical spectrum of oral lesions. Although stratified cross-validation, leakage-free partitioning, and external validation were incorporated to improve reliability, a small dataset may still under-represent rare lesion appearances, anatomical variability, demographic diversity, and acquisition-condition heterogeneity. The class distribution imbalance between cancerous and non-cancerous cases may also influence learned decision boundaries under limited-data conditions. Moreover, because the dataset is publicly curated and lacks detailed metadata regarding acquisition devices, lesion staging, site distribution, and annotation variability, caution is required when interpreting generalizability beyond the studied experimental setting. Although the reported results support the feasibility of transformer-based lesion localization under low-resource conditions, any implication of immediate clinical deployment must remain cautious. The present findings are derived from a limited publicly curated dataset and do not include multi-institutional prospective validation, reader-study comparison, or deployment-level assessment across heterogeneous clinical environments. Accordingly, the framework should be interpreted as a technically promising decision-support approach whose real-world clinical applicability requires further validation under broader institutional, demographic, and acquisition settings.

5. Conclusion and future scope

This paper is a methodical analysis of the transformer-based model of lesion-level tumor localization in the oral cavity when the data is imaged with limited data. The work is based on methodological rigor rather than isolated performance reporting, which is achieved by combining a leakage-free 5-fold cross-validation protocol, fold-wise self-supervised pretraining, uncertainty estimation with bootstrap confidence intervals, and external validation with an independent dataset. The results suggest that representation-level pretraining enhances the low-resource setting of spatial detection stability, and that transformer-based set prediction can be successfully adapted to low-scale medical imaging settings. Notably, clinical equivalence to expert diagnosis is not mentioned in the study. The reported results are of controlled experimental assessment with publicly available photo datasets with no standardized acquisition metadata and no prospective validation. As such, the framework should be considered a technical feasibility investigation and not as a deployable diagnostic system. Multi-center data collection using standardized imaging protocols, the inclusion of histopathological ground truth and lesion staging metadata, evaluation of inter-observer annotation variability, and formal reader studies of the results of models versus expert clinicians should be included in future studies. The calibration, the analysis of the rare lesions and the robustness would be also interesting to explore further with the expansion to the bigger annotated cohorts. Clinical translation can only be assessed in such a structured validation in a responsible way.

5.1. Future scope

Future studies need to focus on the addition of larger and multi-center data sets that are acquired following standardized protocols to enhance strength and minimize dataset-specific bias. Structured clinical metadata, such as lesion site, histopathological confirmation and staging information would be included to facilitate more clinically interpretable evaluation. Future studies should specifically address potential sampling bias associated with small-scale public datasets by incorporating larger, multi-center cohorts with broader demographic coverage, more diverse lesion presentations, and standardized acquisition metadata. Multi-modal extensions using complementary imaging modalities, e.g. optical coherence tomography or radiological imaging, could also be explored further to better characterize the structure, but would need specific datasets and protocol standardization. Annotated cohorts would also allow a more profound analysis of the calibration, the behavior of uncertainty and the detection of rare lesions with larger cohorts. Notably, before any consideration of clinical deployment, prospective validation, assessment of inter-observer variability and systematic reader studies are required. In this regard, the current study can be viewed as a methodological feasibility study as opposed to a clinically validated diagnostic system. For real-world clinical translation, future studies should move beyond retrospective public datasets and include multi-center validation under standardized imaging protocols, device variability assessment, and prospective evaluation in routine screening environments. Formal reader studies comparing model-assisted and clinician-only assessment, together with workflow-level evaluation in supervised decision-support settings, will be necessary before practical clinical integration can be responsibly considered.

Human and animal rRights

Not Applicable.

CRediT authorship contribution statement

Pratham Kaushik: Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Vinay Kukreja:** Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology. **Eshika Jain:** Writing – review & editing, Writing – original draft, Visualization, Resources, Formal analysis, Data curation. **Modafar Ati:** Writing – review & editing, Visualization, Validation, Resources, Funding acquisition. **Shanmugasundaram Hariharan:** Writing – review & editing, Writing – original draft, Visualization, Funding acquisition.

Dataset name

Mouth Cancer Images URL: https://www.kaggle.com/datasets/edward112/mouth-cancer-images/dataAccessed on: 09 April 2025.

License

Publicly available under the Kaggle dataset usage terms (non-exclusive research use).

The dataset consists of 131 oral cavity images (87 cancerous and 44 non-cancerous), captured under heterogeneous photographic conditions. No additional private data were used.

To ensure full reproducibility, the experimental protocol is explicitly defined as follows.

Leakage-free splitting protocol

Dataset partitioning was performed strictly at the level of the 131 original images prior to any augmentation. A 5-fold stratified cross-validation protocol was implemented at the original-image level. Augmentation was applied exclusively to training subsets within each fold. Validation and external evaluation were conducted only on non-augmented original images. No augmented variants were allowed to cross folds.

Preprocessing Pipeline.

- Image resizing to 512 × 512 resolution
- Intensity normalization using dataset-level mean and standard deviation
- Consistent preprocessing across training, validation, and external evaluation

Training Augmentation (Training Folds Only).

- Random rotation ($\leq 45^{\circ}$)
- Horizontal/vertical flip
- Random resized crop
- Brightness and color jitter
- Gaussian noise
- Perspective transformation

No augmentation was applied to validation or external test sets.

Self-supervised pretraining

SimCLR pretraining was conducted within each fold using only training-fold images. Projection head architecture, temperature ($\tau = 0.5$), batch size (32), optimizer (Adam), learning rate (3×10^{-4}), and 100-epoch schedule are fully documented in Section 2.3.2 and Table 4.

Reproducibility resources

All preprocessing scripts, original-image-level split generation code, training configurations, SSL pretraining routines, evaluation scripts (mAP/AP50/AP75 computation), and bootstrap confidence interval code will be made publicly available upon acceptance. A public repository link will be provided in the final published version to ensure full experimental reproducibility.

Ethics statement and institutional review determination

This study utilizes a publicly available dataset ("Mouth Cancer Images," Kaggle platform) consisting of de-identified oral cavity images. The dataset does not contain any personally identifiable information (PII), protected health information (PHI), or patient-specific metadata. No new human participants were recruited, and no prospective data collection or interaction with individuals was performed. As the study involves secondary analysis of publicly available, fully anonymized data, it does not meet the criteria for human subjects research under standard institutional and international guidelines. Therefore, formal Institutional Review Board (IRB) approval or ethics committee clearance was not required.

Consent for publication

All authors agree for publication if paper accepted.

AI use disclosure

Generative artificial intelligence (AI) tools were used solely for language refinement and grammatical editing during the preparation of this manuscript. Specifically, AI-assisted tools were employed to improve clarity, readability, and overall presentation of the text. No AI tools were used for generating scientific content, experimental results, data analysis, or figures. All figures, results, and visualizations presented in this study are original and derived from the authors' own experimental work and standard scientific software. All authors have carefully reviewed, validated, and approved the final manuscript and take full responsibility for the accuracy, integrity, and originality of the content. No AI tool has been listed as an author.

Funding statement

This work is supported as part of a research grant by the Office of Research and Sponsored Programs at Abu Dhabi University, United Arab Emirates Grant reference 19300914.

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.

Acknowledgments

Not Applicable.

Data availability

Data availability is clearly mentioned in the paper.

References

[1] Yang Z, Pan H, Shang J, Zhang J, Liang Y. Deep-learning-based automated identification and visualization of oral cancer in optical coherence tomography images. Biomedicines 2023;11(3):802.
[2] Vinay V, et al. Artificial intelligence in oral cancer: a comprehensive scoping review of diagnostic and prognostic applications. Diagnostics 2025;15(3):280.
[3] Saraswathi T, Murali Bhaskaran V. Heuristic strategy using hybrid deep learning with transfer learning for oral cancer detection. Int J Wavelets, Multiresolut Inf Process 2025;23(1):2450039.
[4] Muqeet MA, Mohammad AB, Krishna PG, Begum SH, Qadeer S, Begum N. Automated oral cancer detection using deep learning-based technique. In: 2022 8th international conference on signal processing and communication (ICSC). IEEE; 2022. p. 294–7.
[5] Kavyashree C, Vimala HS, Shreyas J. A systematic review of artificial intelligence techniques for oral cancer detection. Healthc Anal 2024;5:100304.
[6] Raval D, Undavia JN. A comprehensive assessment of convolutional neural networks for skin and oral cancer detection using medical images. Healthc Anal 2023;3:100199.
[7] Lu J, Sladoje N, Runow Stark C, Darai Ramqvist E, Hirsch J-M, Lindblad J. A deep learning based pipeline for efficient oral cancer screening on whole slide images. In: International conference on image analysis and recognition. Springer; 2020. p. 249–61.
[8] Ariji Y, et al. Contrast-enhanced computed tomography image assessment of cervical lymph node metastasis in patients with oral cancer by using a deep learning system of artificial intelligence. Oral Surg Oral Med Oral Pathol Oral Radiol 2019;127(5):458–63.
[9] Nanditha BR, Geetha A, Chandrashekar HS, Dinesh MS, Murali S. An ensemble deep neural network approach for oral cancer screening. 2021.
[10] Huang Q, Ding H, Razmjooy N. Oral cancer detection using convolutional neural network optimized by combined seagull optimization algorithm. Biomed Signal Process Control 2024;87:105546.
[11] Kavyashree C, Vimala HS, Shreyas J. Improving oral cancer detection using pretrained model. In: 2022 IEEE 6th conference on information and communication technology (CICT). IEEE; 2022. p. 1–5.
[12] Kirubabai MP, Arumugam G. Deep learning classification method to detect and diagnose the cancer regions in oral MRI images. Med Leg Update 2021;21:462–8.
[13] Lian W, Lindblad J, Stark CR, Hirsch J-M, Sladoje N. Let it shine: autofluorescence of Papanicolaou-stain improves AI-based cytological oral cancer detection. Comput Biol Med 2025;185:109498.
[14] Wankhade D, Dhawale C, Meshram M. Advanced deep learning algorithms in oral cancer detection: techniques and applications. J Environ Sci Heal Part 2024;C:1–26.
[15] Kumar J, Pandey V, Tiwari RK. Deep learning framework for early detection of oral cancer. In: AI and green technology applications in society. IGI Global Scientific Publishing; 2025. p. 353–72.
[16] Warin K, Limprasert W, Suebnukarn S, Jinaporntham S, Jantana P, Vicharueang S. AI-based analysis of oral lesions using novel deep convolutional neural networks for early detection of oral cancer. PLoS One 2022;17(8):e0273508.
[17] Welikala RA, et al. Automated detection and classification of oral lesions using deep learning for early detection of oral cancer. IEEE Access 2020;8:132677–93.
[18] Giri SK, Dash S. Synthesis of clinical images for oral cancer detection and prediction using deep learning. In: Mining biomedical text, images and visual features for information retrieval. Elsevier; 2025. p. 339–56.
[19] Cimino MGCA, et al. Explainable screening of oral cancer via deep learning and case-based reasoning. Smart Heal 2025;35:100538.
[20] Varalakshmi D, Tharaheswari M, Anand T, Saravanan KM. Transforming oral cancer care: the promise of deep learning in diagnosis. Oral Oncol Rep 2024;10:100482.
[21] Satheeskumar R. AI-driven diagnostics and personalized treatment planning in oral oncology: innovations and future directions. Oral Oncol Rep 2025;13:100704.
[22] Lin H, Chen H, Weng L, Shao J, Lin J. Automatic detection of oral cancer in smartphone-based images using deep learning for early diagnosis. J Biomed Opt 2021;26(8):86007.
[23] Warin K, Limprasert W, Suebnukarn S, Jinaporntham S, Jantana P. Automatic classification and detection of oral cancer in photographic images using deep learning algorithms. J Oral Pathol Med 2021;50(9):911–8.

[24] Warin K, Suebnukarn S. Deep learning in oral cancer-a systematic review. BMC Oral Health 2024;24(1):212.
[25] Tanriver G, Soluk Tekkesin M, Ergen O. Automated detection and classification of oral lesions using deep learning to detect oral potentially malignant disorders. Cancers (Basel) 2021;13(11):2766.
[26] Tenali N, Desu VS, Boppa C, Chintala VC, Guntupalli B. Oral cancer detection using deep learning techniques. In: 2023 international conference on innovative data communication technologies and application (ICIDCA). IEEE; 2023. p. 168–75.
[27] Nieri M, Serni L, Clauser T, Paoletti C, Franchi L. Diagnosis of Oral Cancer With Deep Learning. A Comparative Test Accuracy Systematic Review. Oral Diseases 2025;31(8):2368–81.
[28] Jubair F, Al-karadsheh O, Malamos D, Al Mahdi S, Saad Y, Hassona Y. A novel lightweight deep convolutional neural network for early detection of oral cancer. Oral Dis 2022;28(4):1123–30.
[29] Dharani R, Revathy S. DEEPORCD: detection of oral cancer using deep learning. In: Journal of physics: conference series. IOP Publishing; 2021, 12006.
[30] Huang Q, Ding H, Razmjooy N. Optimal deep learning neural network using ISSA for diagnosing the oral cancer. Biomed Signal Process Control 2023;84:104749.
[31] Jeyaraj PR, Samuel Nadar ER. Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm. J Cancer Res Clin Oncol 2019;145:829–37.
[32] Alabi RO, Almangush A, Elmusrati M, Mäkitie AA. Deep machine learning for oral cancer: from precise diagnosis to precision medicine. Front Oral Health 2022;2:794248.
[33] Chang X, et al. Deep learning methods for oral cancer detection using raman spectroscopy. Vib Spectrosc 2023;126:103522.
[34] Fu Q, et al. A deep learning algorithm for detection of oral cavity squamous cell carcinoma from photographic images: a retrospective study. EClinicalMedicine 2020;27.
[35] Dixit S, Kumar A, Srinivasan K. A current review of machine learning and deep learning models in oral cancer diagnosis: recent technologies, open challenges, and future research directions. Diagnostics 2023;13(7):1353.
[36] Patel S, Kumar D. Predictive identification of oral cancer using AI and machine learning. Oral Oncol Rep 2025;13:100697.
[37] Alanazi AA, Khayyat MM, Khayyat MM, Elamin Elnaim BM, Abdel-Khalek S. Intelligent deep learning enabled oral squamous cell carcinoma detection and classification using biomedical images. Comput Intell Neurosci 2022;2022(1):7643967.
[38] Prabhakaran R, Mohana DJ. Detection of oral cancer using machine learning classification methods. Int J Electr Eng Technol 2020;11(3).
[39] Kim DW, Lee S, Kwon S, Nam W, Cha I-H, Kim HJ. Deep learning-based survival prediction of oral cancer patients. Sci Rep 2019;9(1):6994.
[40] Marzouk R, et al. Deep transfer learning driven oral cancer detection and classification model. Comput Mater Continua (CMC) 2022;73(2).
[41] López-Cortés XA, Matamala F, Venegas B, Rivera C. Machine-learning applications in oral cancer: a systematic review. Appl Sci 2022;12(11):5715.
[42] Lavanya J, Kavya G, Prasamya N. Oral cancer diagnosis using deep learning for early detection. In: 2022 international conference on electronics and renewable systems (ICEARS). IEEE; 2022. p. 1260–8.
[43] Li L, et al. Development of an oral cancer detection system through deep learning. BMC Oral Health 2024;24(1):1–10.
[44] Al-Rawi N, et al. The effectiveness of artificial intelligence in detection of oral cancer. Int Dent J 2022;72(4):436–47.
[45] Nanditha BR, Mp G. Oral cancer detection using machine learning and deep learning techniques. Int J Curr Res Rev 2022;14(1):64–70.
[46] Babu PA, et al. An explainable deep learning approach for oral cancer detection. J Electr Eng Technol 2024;19(3):1837–48.
[47] Mouth cancer images. https://www.kaggle.com/datasets/edward112/mouth-cancer-images/data. [Accessed 9 April 2025].
[48] S CH, A GK, S M, M S D, B R N. Oral images dataset. Mendeley Data 2021;2. https://doi.org/10.17632/MHJYRN35P4.2.
[49] Oral images dataset - mendeley data. https://data.mendeley.com/datasets/mhjyrn35p4/2. [Accessed 11 February 2026].