2025 IEEE Symposium on Computational Intelligence in Image, Signal Processing and Synthetic Media Companion (CISM Companion) | 979-8-3315-0851-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/CISMCompanion65074.2025.11032552

Improving oral cancer classification via segment-driven photographic deep learning imaging

Marco Parola
University of Pisa
marco.parola@ing.unipi.it
0000-0003-4871-4902

Edoardo Malaspina
University of Pisa
edoardo.malaspina@sma-rty.com
0009-0009-8628-6225

Mario G.C.A. Cimino
University of Pisa
mario.cimino@unipi.it
0000-0002-1031-1959

Gaetano La Mantia
University of Palermo
gaetano.lamantia@community.unipa.it
0000-0002-3135-7462

Giuseppina Campisi
University of Palermo
campisi@odonto.unipa.it
0000-0002-9443-0495

Olga Di Fede
University of Palermo
odifede@odonto.unipa.it
0000-0002-5562-7420

***Abstract*—Oral cancer is a major health problem requiring accurate healthcare support systems, and Deep learning (DL) based medical imaging has proven to be an effective solution. This work addresses the oral cancer classification task by employing different convolutional architectures. Our goal is to improve the classification tasks by incorporating segmentation information. We propose two segment-driven strategies to strengthen the traditional classification training. The first one involves training a dedicated neural network (NN) to predict masks, which are then used to classify masked images to hide unuseful information. Specifically, we introduce an approach relying on soft-masks to weigh the contribution of each pixel to the final classification against the already proposed hard-mask strategy. The second proposed approach involves training the NN via CrossEntropyIoU, a loss function consisting of the CrossEntropy for identifying the correct label, and the Intersection over Union measuring the mismatch between the activation map and the mask. Experiments show that implementing segment-driven strategies enhances accuracy and training speed using both convolutional and transformer architectures.**

***Index Terms*—Oral cancer, Segment-driven classification, Soft segmentation, Deep learning, CNNs, transformers.**

I. INTRODUCTION

With a significant overall burden of morbidity and mortality, oral squamous cell carcinoma (OSCC) poses a problem in oncology. It represents a significant risk to public health and highlights the urgent need to improve early detection methods [1]. Currently, radiotherapy and chemotherapy are the main adjuvant treatments for OSCC, with surgical removal as the main therapeutic intervention. Even though these therapies cause severe mutilation and a consequent decline in quality of life, the 5-year overall survival rate remains 60% [2]. To achieve a more accurate effective response, early diagnosis of OSCC is essential. Screening relying on healthcare support systems emerges as an essential tool for early diagnosis and treatment of oral cancer in this context [3]. Accurate diagnosis promotes early initiation of appropriate therapies and reduces the potential for unnecessary invasive procedures, thereby reducing patient discomfort and associated healthcare costs. This practice is also promoted by the wide availability of cameras, which enables large-scale screening relying on more widespread and frequent testing in the population to detect cancer in its early stages [4].

Deep Learning (DL) has emerged as a powerful approach to support medical personnel in various diagnostic tasks [5, 6]. In particular, adopting DL to solve classification problems in oral cancer can improve the efficiency of diagnostic procedures. Taking advantage of a DL approach, we propose to develop a classification model for OSCC recognition.

In image analysis, numerous solutions have been proposed to improve the performance of classifiers. Among these, one approach is to exploit segment information [7]. Integrating segmentation data into the classification process has been shown to improve the overall accuracy of classification models. This research is concerned with exploring the effectiveness of DL techniques in oral cancer screening using a dataset of photographic images [8]. The dataset used has been collected and annotated through the collaboration of qualified professionals, including dental consultants and trainees specializing in oral medicine. Focusing on the classification task, the work explores integrating segmentation information in DL training to enhance performance. Against the traditional classification training, we adopt a strategy involving segment prediction and explicit input mask which is exploited in different medical works [9, 10]. In addition to obscuring images by applying a hard mask, we introduced a soft mask approach in which different image areas are emphasized using soft masks as weights. The conceptual insight comes from a similar study [11], which demonstrated improved effectiveness of using uncertain segments. The weakness identified in expressed binary masks is directly related to the structural complexity of lesions, where uncertain or inconsistent areas conceal valuable anatomical information crucial for accurate diagnosis. Inspired by the results of the cited article, we introduce to the field of medical imaging a soft mask.

We propose training a NN using CrossEntropyIoU (CEn-

tIoU), a combined loss function that considers classification performance and the mismatch between grad cam activation maps and physician-provided masks. This guides the NN to focus on specific image areas without masking redundant parts. Experiments show this technique speeds up deep learning model training for both convolutional and transformer architectures, as the additional cost of generating saliency maps is offset by fewer training epochs.

II. Background and literature review

A. Image classification

Advances in DL architectures have made computer vision popular in recent years, specifically thanks to two well-known architectures: convolutional neural networks (CNNs) and transformers [12].

Such architectures are widely adopted for addressing image classification problems. Let $\mathscr{X}$ be the input space, and $\mathscr{Y} = \{1, 2, \ldots, K\}$ be the set of possible classes, a classifier training requires a labeled dataset $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathscr{X}$ is an input sample and $y_i \in \mathscr{Y}$ is the corresponding class label. The goal is to learn a mapping $f : \mathscr{X} \to \mathscr{Y}$ from the training data, such that for any new input x the classifier can predict its class label $\hat{y} = f(x)$. The learning process involves finding a model θ minimizing the *cross entropy* loss function (CEnt) quantifying the dissimilarity between the predicted output and the true labels. The optimization problem is formulated as:

$$\theta^* = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} CEnt(f_\theta(x_i), y_i) \tag{1}$$

where $f_\theta(x_i)$ is the predicted label for the input x_i based on the model parameters θ, and θ^* represents the optimal set of parameters.

B. Image segmentation

As well for classification, DL models have been employed in the segmentation problems. The segmentation problem consists of training a model to predict segmentation masks given input images. Let x be an image of the input space $\mathscr{X}$ and y the corresponding mask in the output space $\mathscr{Y}$, the training dataset is comprised of N examples $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is an input image and y_i is its ground truth segmentation mask. The goal is to find a model f that maps input images to predicted segmentation masks. Binary cross entropy *BCEntropy* loss function is usually adopted to measure the dissimilarity between the predicted mask $f(x_i)$ and the ground truth mask y_i. The optimization problem is formulated as:

$$\theta^* = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} BCEntropy(f_\theta(x_i), y_i) \tag{2}$$

where $f_\theta(x_i)$ is the predicted mask for the input x_i based on the model parameters θ, and θ^* represents the optimal set of parameters.

Hard or soft segmentation can be performed, generating hard or soft masks respectively. Soft masks introduce a probabilistic element, assigning each pixel a probability value that reflects the likelihood of whether or not it belongs to the target object. Hard masks, on the other hand, refer to binary segmentation maps in which each pixel is assigned a discrete label indicating whether or not it belongs to the target object. These masks are characterized by a distinct delimitation of the object boundaries. They can be obtained from soft masks by applying a threshold. Figure 1 shows an example of hard and soft mask prediction.

(a) (b) (c)

Fig. 1: Prediction example of soft and hard masks: (a) an input image and segment provided by physicians, (b) the predicted hard mask, and (c) the predicted soft mask.

Since its introduction in 2015, the Fully Convolutional Network (FCN) [13] has significantly advanced semantic segmentation by enabling end-to-end, pixel-by-pixel training with convolutional networks. Chen et al. further improved this with Deeplab [14], which uses dilated convolutions to maintain resolution and atrous spatial pyramid pooling to capture multi-scale context. The U-Net architecture [15], with its U-shaped structure of encoder and decoder blocks, is also popular for segmentation. Its skip connections enhance the decoder's ability to produce accurate semantic features.

C. Saliency map

Saliency maps, which highlight the regions in an input image that contribute the most to the network's decision, serve as valuable tools for understanding model predictions.

Grad-CAM extends the concept of saliency maps to provide class-specific localization [16] by computing a weighted combination of the feature maps in the final convolutional layer of the CNN on the gradient of the predicted class score y_c with respect to the feature maps. The Grad-CAM map L_c for class c is computed as follows:

$$L_c = \text{ReLU}\left(\sum_k \alpha_k^c F_k\right) \tag{3}$$

where F_k represents the k-th feature map from the last convolutional layer, and α_k^c is the weight assigned to each feature map, computed as the global average pooling of the gradients:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y_c}{\partial A_{ij}^k} \tag{4}$$

where A_{ij}^k is the activation of the k-th feature map at position (i, j), and Z is the spatial dimension of the feature maps.

III. Methodology

In addressing the classification problem, we conducted some benchmark experiments on the use of pre-trained models as a baseline experiment indicated as CEnt (Cross Entropy). Next, we introduce two alternative strategies aimed at improving classification performance by exploiting segment information.

The first proposed approach, called "soft-mask", aims at improving [9]. This uses a two-stage architecture, employing two distinct consecutive networks, as shown in Figure 2. The first NN focuses on segmentation, generating a non-binary soft mask that outlines the relevant regions; then, the second network classifies the image based on the masked regions, allowing for a more refined classification process. Experiments following the workflow of [9] have also been reproduced for state-of-the-art comparison and are indicated as "hard-mask".

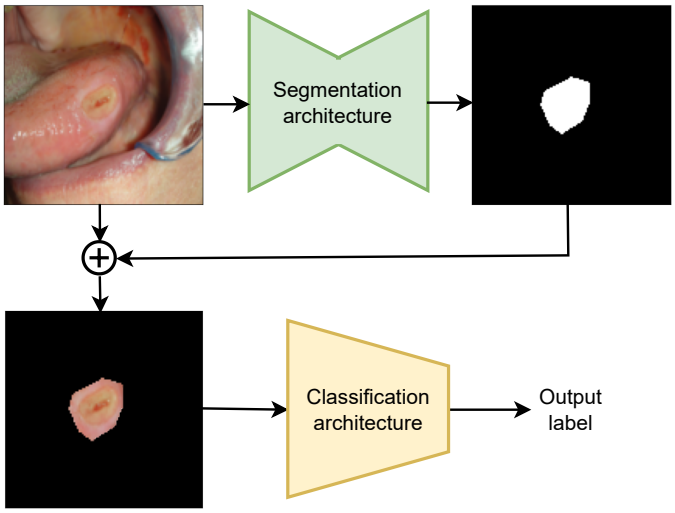

Fig. 2: Two-stage training DL pipeline.

The second approach trains a classifier using a new combined loss function, CEntIoU, as shown in Figure 3. During training, the classifier generates a predicted label and the corresponding saliency map highlighting crucial image regions. Then, the CEntIoU metric, defined in Equation 5, combines classification error (CEnt) and the discrepancy between the ground truth mask and the saliency map (IoU). Intuitively, the ground truth mask and the saliency map serve as two different methods for identifying the same concept—the most significant region of an image pertinent to the prediction. λ balances the two components. This metric is used as a loss function during backpropagation.

$$CEntIoU = \lambda \cdot CEnt(a,p) - (1-\lambda) \cdot IoU(m,s) \qquad (5)$$

The classification evaluation includes accuracy on the overall dataset and the accuracy per class. Pixel accuracy, precision, recall, and specificity are used for the segmentation [17].

Between 2021 and 2024, photographic images of patients' oral cavities were captured during medical examinations at the Oral Medicine Unit of the P. Giaccone University Hospital in Palermo, Italy. Dental consultants took photos using both

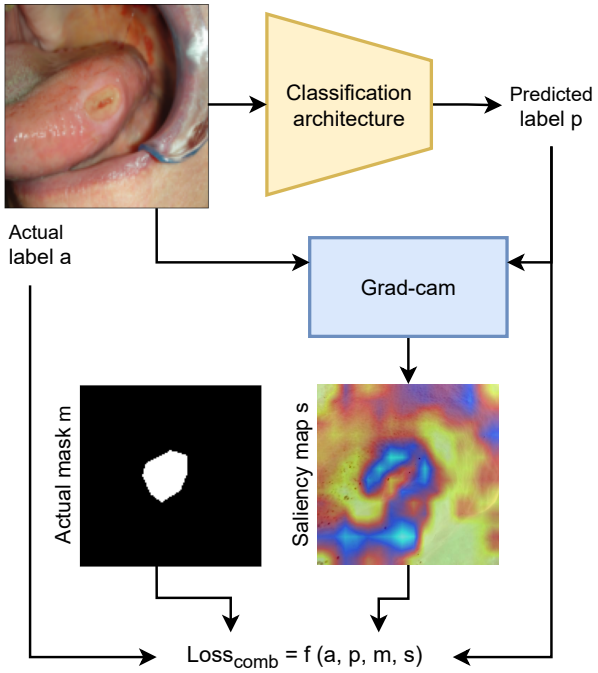

Fig. 3: Forward phase during the training to compute the CentIou loss based on the actual label a, the predicted label p, the actual mask m, and the saliency maps s.

standard cameras and smartphones. We collected 607 images of three distinct pathologies: aphthous (ap), traumatic (tr), and neoplastic (ne). The lesions were manually labeled by a dental team, associating each lesion with a border segment and a label. To promote collaborations and for the sake of reproducibility, publicly available dataset has been provided [18]. The dataset has been randomly split into three different sets: the validation set is used to perform early stopping during the training. The split has been performed allocating randomly 70% of the total images to the training set, 15% to the validation set, and 15% to the test set.

IV. Experiments and Results

The experiments were conducted on a VM powered by an RTX GPU. We made the source code publicly available to facilitate the replication of the experiments [19].

We conducted benchmark experimentation between SqueezeNet and ConvNext as convolutional models. For transformer architectures, we evaluated Vision Transformer (ViT) and Swin Transformer (Swint). All models are pre-trained on the `imagenet` dataset.

We present the performance of each model trained with the different approaches described in Section III. As the soft-mask approach requires a segmentation model for the mask generation, in Table II we propose a comparison between three segmentation models, FCN, DeepLab, and U-Net.

The experiments requiring a dedicated network for mask generation (both soft and hard) were conducted using U-Net, which outperformed the other segmentation models.

TABLE I: Classification metrics: global and individual class accuracies for aphthous(ap), traumatic(tr), and neoplastic(ne)

Model	Training approach	Optim. Hyper-param		Evaluation metrics			
		lr	λ	Accuracy	Accuracy$_{ap}$	Accuracy$_{tr}$	Accuracy$_{ne}$
Squeezenet	hard-mask	1.34e-4	-	73.24 ± 0.16	83.02 ± 0.26	66.10 ± 0.14	70.24 ± 0.08
	soft-mask	2.88e-5	-	**78.16 ± 0.64**	83.92 ± 0.37	64.76 ± 0.66	87.14 ± 0.94
	CEnt	2.53e-5	-	70.21 ± 0.60	80.71 ± 0.78	54.77 ± 0.42	76.09 ± 0.61
	CEntIoU	3.08e-6	0.5	70.79 ± 0.31	66.87 ± 0.54	74.54 ± 0.12	70.96 ± 0.25
Convnext	hard-mask	2.13e-5	-	81.34 ± 0.26	77.42 ± 0.29	89.30 ± 0.20	76.61 ± 0.31
	soft-mask	2.64e-6	-	**84.52 ± 0.55**	89.61 ± 0.64	76.33 ± 0.52	88.18 ± 0.47
	CEnt	2.64e-5	-	79.76 ± 0.32	86.36 ± 0.21	77.97 ± 0.33	74.19 ± 0.45
	CEntIoU	4.75e-5	0.6	79.97 ± 0.71	76.10 ± 0.73	75.12 ± 0.79	90.17 ± 0.60
Vit	hard-mask	1.28e-5	-	79.43 ± 0.32	83.41 ± 0.15	76.77 ± 0.49	77.92 ± 0.33
	soft-mask	1.39e-6	-	**83.16 ± 0.45**	89.20 ± 0.51	81.39 ± 0.44	78.21 ± 0.39
	CEnt	1.88e-5		80.69 ± 0.72	83.34 ± 0.67	73.68 ± 0.80	85.82 ± 0.71
	CEntIoU	1.67e-6	0.6	81.56 ± 0.34	80.32 ± 0.39	84.13 ± 0.26	80.01 ± 0.37
SwinT	hard-mask	5.81e-5	-	81.24 ± 0.84	79.21 ± 0.89	75.26 ± 0.86	90.62 ± 0.77
	soft-mask	3.76e-6	-	**82.52 ± 0.41**	82.74 ± 0.52	90.82 ± 0.29	72.56 ± 0.42
	CEnt	5.28e-7	-	81.12 ± 0.58	84.22 ± 0.54	76.34 ± 0.67	83.13 ± 0.53
	CEntIoU	5.50e-5	0.7	81.73 ± 0.18	80.94 ± 0.19	92.31 ± 0.15	70.827 ± 0.21

The classification experiments were done by appending a NN consisting of a fully connected layer with 64 units with ReLU activation and a linear output layer producing the three classes. Regarding the hyper-parameters configuration, we fixed the *batchsize* at the maximum capacity of the GPU (32) and we performed a random search hyperparameter optimization to find the best lr value for each training strategy ranging between $1e-5$ and $1e-7$. About CEntIoU training, we also optimized the λ parameter values by searching between 0.4 and 0.9 with a step size of 0.05. We performed 20 runs for each model and each strategy to provide result confidence intervals at 95%.

In Table I we present the results achieved by each model trained using the different strategies with the corresponding optimized lr value: (i) a standard fine-tuning classification problem *CEnt*, (ii) the classification where the image has been hard masked *hard-mask*, (iii) the classification where the image has been soft masked *soft-mask*, and, (iv) the proposed training relying on the combined loss *CEntIoU*. Specifically, we can observe that the soft-mask training approach is the most effective in achieving the best classification results for all considered models.

In addition to the accuracy metrics, we propose a model's training computational complexity analysis of CEnt loss function against CEntIoU. Table III shows the average number of epochs, time per epoch, and total training time revealing for all models the CEntIoU loss function is more effective than CEnt. Despite saliency maps generation at the end of each epoch is computationally demanding, in all cases, fewer epochs are required to train a model using CEntIoU loss, reducing the number of epochs by 42% on average. This training speed-up has no accuracy performance implications as previously shown in Table I.

TABLE II: Segmentation evaluation on test set

Model	PixAcc	Prec	Recall	Spec
FCN	94.09	**80.05**	68.60	96.97
DeepLab	94.32	78.86	69.86	97.38
U-Net	**95.15**	79.86	**71.96**	**97.90**

TABLE III: Training convergence analysis. $^{\circ}$ s e+2. $^{\triangle}$ s e+4

Model	CEnt			CEntIoU		
	num. epochs	epoch time$^{\circ}$	train. time$^{\triangle}$	num. epochs	epoch time$^{\circ}$	train. time$^{\triangle}$
Squeezenet	381	0.85	3.24	214	1.02	**2.19**
Convnext	146	0.99	1.44	74	1.29	**0.95**
ViT	150	1.07	1.61	106	1.39	**1.47**
Swin	179	0.98	1.75	96	1.45	**1.39**

V. Conclusions and Future Work

Our study explored the application of DL architectures for oral cancer classification, focusing on improving performance through segmentation-driven strategies. Two segment-driven approaches were examined against traditional image classification and hard-masked classification [9]. The first involved training a dedicated model for soft-mask prediction and classifying the masked images, hiding irrelevant information to the classification. Then, we proposed CEntIoU, a combined loss function that considers both the classification accuracy and the mismatch between the activation map and the mask. This loss guides the NN to focus on specific areas of the image without explicitly masking redundant regions. The comparison revealed the superiority of segment-driven approaches in increasing classification accuracy or speeding up model training, making them suitable in oral cancer screening systems, depending on whether we want to prioritize model accuracy or face resource-limited contexts.

Acknowledgment

Work supported by the MIUR in the framework of the FoReLab project (Dept of Excellence).

REFERENCES

[1] Y.-J. Kim and J. H. Kim, "Increasing incidence and improving survival of oral tongue squamous cell carcinoma," *Scientific reports*, vol. 10, no. 1, p. 7877, 2020.

[2] J.-Y. Zhou, W.-J. Wang, C.-Y. Zhang, Y.-Y. Ling, X.-J. Hong, Q. Su, W.-G. Li, Z.-W. Mao, B. Cheng, C.-P. Tan, *et al.*, "Ru (ii)-modified tio2 nanoparticles for hypoxia-adaptive photoimmunotherapy of oral squamous cell carcinoma," *Biomaterials*, vol. 289, p. 121757, 2022.

[3] C. Bramati, S. Abati, S. Bondi, A. Lissoni, G. Arrigoni, F. Filipello, and M. Trimarchi, "Early diagnosis of oral squamous cell carcinoma may ensure better prognosis: A case series," *Clinical Case Reports*, vol. 9, no. 10, 2021.

[4] B. Hunt, A. J. Ruiz, and B. W. Pogue, "Smartphone-based imaging systems for medical applications: a critical review," *Journal of Biomedical Optics*, vol. 26, no. 4, pp. 040902–040902, 2021.

[5] P. R. Jeyaraj and E. R. Samuel Nadar, "Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm," *Journal of cancer research and clinical oncology*, vol. 145, pp. 829–837, 2019.

[6] M. Parola, G. L. Mantia, F. Galatolo, M. G. Cimino, G. Campisi, and O. Di Fede, "Image-based screening of oral cancer via deep ensemble architecture," in *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1572–1578, 2023.

[7] B. Song, C. Zhang, S. Sunny, D. R. Kc, S. Li, K. Gurushanth, P. Mendonca, N. Mukhia, S. Patrick, S. Gurudath, *et al.*, "Interpretable and reliable oral cancer classifier with attention mechanism and expert knowledge embedding via attention map," *Cancers*, vol. 15, no. 5, p. 1421, 2023.

[8] M. Parola, F. A. Galatolo, G. La Mantia, M. G. Cimino, G. Campisi, and O. Di Fede, "Towards explainable oral cancer recognition: Screening on imperfect images via informed deep learning and case-based reasoning," *Computerized Medical Imaging and Graphics*, vol. 117, p. 102433, 2024.

[9] V. Anand, S. Gupta, D. Koundal, and K. Singh, "Fusion of u-net and cnn model for segmentation and classification of skin lesion from dermoscopy images," *Expert Systems with Applications*, vol. 213, p. 119230, 2023.

[10] H. Azimi, J. Zhang, P. Xi, H. Asad, A. Ebadi, S. Tremblay, and A. Wong, "Improving classification model performance on chest x-rays through lung segmentation," 2022.

[11] L. Wang, X. Ye, L. Ju, W. He, D. Zhang, X. Wang, Y. Huang, W. Feng, K. Song, and Z. Ge, "Medical matting: Medical image segmentation with uncertainty from the matting perspective," *Computers in Biology and Medicine*, vol. 158, p. 106714, 2023.

[12] K. Islam, "Recent advances in vision transformer: A survey and outlook of recent work," *arXiv preprint arXiv:2203.01536*, 2022.

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241, Springer, 2015.

[16] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that?," *arXiv preprint arXiv:1611.07450*, 2016.

[17] T. Schlosser, M. Friedrich, T. Meyer, and D. Kowerko, "A consolidated overview of evaluation and performance metrics for machine learning and computer vision," *Tobias Schlosser, Michael Friedrich, Trixy Meyer, and Danny Kowerko–Junior Professorship of Media Computing, Chemnitz University of Technology*, vol. 9107, 2023.

[18] M. Parola, "Poci dataset - photographic oral cancer imaging dataset, kaggle." https://www.kaggle.com/datasets/marcoparola7/poci-photographic-oral-cancer-imaging-dataset, 2025.

[19] E. Malaspina, "Github improve-classifier-via-segment code repository, https://github.com/marcoparola/improve_classifier_via_segment," 2023.