2025 IEEE Symposium on Computational Intelligence in Health and Medicine Companion (CIHM Companion) | 979-8-3315-1978-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/CIHMCOMPANION65205.2025.11002690

Oral cancer recognition on photographic images via deep learning semantic segmentation

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

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

Irene Cantini
University of Pisa
i.cantini2@studenti.unipi.it
0000-0000-0000-0000

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*—Medical image segmentation is an important task supporting diagnosis and screening systems in several medical areas including oral cancer recognition. This paper explores the effectiveness of different deep learning (DL) architectures, including U-Net, LinkNet, PAN, and FPN for oral cavity lesion segmentation. Furthermore, we propose an ensemble model incorporating several decision fusion strategies to aggregate individual predictions, to improve the individual model performance. Our study employs a dataset acquired and manually labeled by the clinical subgroup of our team. On this dataset, we address two distinct segmentation problems: binary semantic segmentation to differentiate healthy tissue from diseased regions and multiclass semantic segmentation to identify three oral pathologies: aphthous, traumatic, and neoplastic lesions. We study the ensemble model's effectiveness in improving segmentation accuracy by combining different DL architectures' strengths. The results demonstrate that the ensemble strategy is highly effective for binary semantic segmentation, achieving a Dice score of 76.5%; while, for the multi-class problem of differentiating between multiple diseases, improvements are present but less marked.**

***Index Terms*—Oral cancer, Oral squamous cell carcinoma, Unet, Healthcare screening, Ensemble learning, Semantic segmentation.**

I. Introduction

Oral cancer, including Oral Squamous Cell Carcinoma (OSCC), is a major worldwide health issue. OSCC is the most common head and neck squamous cell carcinoma, accounting for 90% of all oral cancers. Despite advances in modern therapy, the mortality rate of oral cancer has not been optimally controlled. The estimated 5-year overall survival rate is approximately 50%, with more than 50% of patients presenting with advanced disease at the time of diagnosis [1, 2]. Treatment options for OSCC include surgery, radiation, and chemotherapy. However, the effectiveness of these treatments can vary based on many factors, such as the cancer's location and stage [3].

In this context, image segmentation emerges as a strategic image-processing component in healthcare, as it enables the delineation of diseased areas and patterns within medical images. By outlining anatomical structures and abnormalities, image segmentation helps radiologists and physicians identify and quantify various diseases and conditions [4, 5].

Deep Learning (DL) has emerged as the predominant method for image segmentation in computer vision, outperforming conventional techniques such as thresholding, edge or region-based methods[6]. Furthermore, among the various strategies to improve the performance of DL models, ensemble learning has proven effective. Ensemble is a technique relying on multiple basic learners to form an aggregate model to combine their strengths and limit the weaknesses of individual models, resulting in a better generalization of the learning system [7]. To tackle the semantic segmentation of OSSC, high-quality training data is essential. Photographic images, despite potential artifacts, provide a rapid and cost-effective medium for initial screening [8]. Unlike expensive techniques like histopathological or fluorescence imaging, which are better suited for detailed analysis, photographic methods are accessible and practical for broad use [9].

In this paper, we present an analysis of semantic segmentation within the oral cancer application domain on a newly collected set of photographic images of the oral cavity. We evaluate existing DL architectures. Then, we introduce different ensemble learning strategies to enhance the performance of individual segmentation models, for both binary and multiclass semantic segmentation tasks. The most noticeable improvement made by the ensemble model concerned binary segmentation, where we found an improvement of about 3% in the DICE metric. At the same time, it was less evident in the case of the multiclass problem.

II. Related work

Semantic segmentation is the problem of partitioning an input image into one semantic region between a predefined number, where each pixel is assigned a class label indicating the category of the object it belongs to. Let x be an input image of shape $W \times H \times C$, representing the width, height, and number of channels respectively. The output of semantic image segmentation is a segmentation map y of shape $W \times$

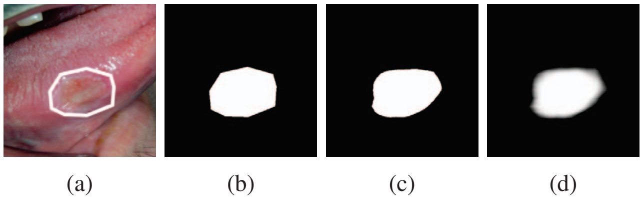

Fig. 1: Examples of predictions for soft and hard masks: (a) segmented image; (b) ground truth mask generated from the segment; (c) hard mask output; and (d) soft mask output.

$H \times L$, where L is the number of semantic classes. The goal of the training of a semantic segmentation model is to find a function $f_{theta} : \mathbb{R}^{W \times H \times C} \rightarrow \mathbb{R}^{W \times H \times L}$ able to assign a class label to each pixel in the image. θ represents the model parameters and, during the training process, the best parameters $theta^*$ are learned by solving the optimization problem shown in Equation 1.

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

Both hard and soft image segmentation techniques can be implemented, resulting in hard or soft masks. Soft masks consist of a probabilistic representation, in which each pixel is assigned a probability value indicating the chance of association with the target object. Hard masks, on the other hand, consist of binary segmentation maps, in which each pixel is assigned a discrete label indicating whether or not it belongs to the target object.

A. Deep learning architectures for image segmentation

Since 2015, U-Net introduced by Ronneberger et al. in the biomedical imaging context represents a major paradigm shift in semantic segmentation [10]. U-Net is a U-shaped encoder-decoder network architecture consisting of four encoder blocks and four decoder blocks. The encoder network serves as a feature extractor and reduces the dimensionality of the input image. Each encoding block consists of two convolutions followed by a ReLU. Skip connections provide additional information helping the decoder to generate better semantic features. In 2017, Chaurasia et al. proposed LinkNet (LN) [11]. This architecture has been designed to address the efficient segmentation challenge by employing a lightweight encoder-decoder architecture, that reduces computational costs and allows for rapid inference, making it suitable for real-time applications. Pyramid Attention Network (PAN) introduced by Li et al. in 2018 [12] enhances segmentation performance by combining spatial pyramid and attention mechanism to extract accurate and dense features for pixel labeling, rather than dilated convolutions and artificially designed decoding networks. The Feature Pyramid Attention module realizes a spatial attention pyramid structure on high-level information and combines global pooling to learn a better representation of features; then, a Global Attention Upsample module on each decoding level provides a global context as a guide for low-level features to select category position. The last model examined is Feature Pyramid Network (FPN) proposed in 2019 [13]. FPN was proposed to solve object and semantic segmentation tasks using a single model at the architectural level. This method utilizes the shared backbone of FPN to equip the well-known Mask R-CNN with a segmentation branch.

Ensemble learning enhances the effectiveness of DL models by combining multiple models to leverage their strengths. As noted in [14], ensemble strategies fall into two categories: homogeneous ensembles, where identical models are trained on different datasets, and heterogeneous ensembles, which combine diverse algorithms or architectures to capture varied data patterns.

III. Methodology

We used several state-of-the-art models to address two semantic segmentation tasks in medical imaging, proposing comprehensive benchmark experiments: (i) a binary semantic segmentation to distinguish healthy from diseased tissue and (ii) a multi-class semantic segmentation where different pathologies are distinguished.

DL segmentation models have been trained as suggested by Gros et al. [15] using *BCEWithLogitsLoss* considering the mismatch between ground truth and soft mask rather than hard mask, resulting in improved alignment accuracy. After training the models to predict soft masks, the output requires binarization to make it suitable to apply for the evaluation metrics. A binarization threshold optimization was performed by searching the optimal value ranging between 0.0 and 1.0 that maximizes the Dice score on the validation set [15].

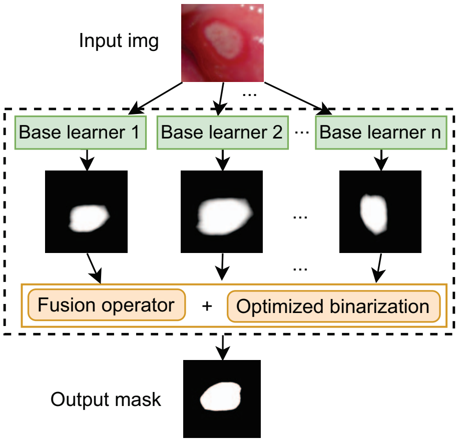

Fig. 2: Heterogeneous ensemble schema.

We employ the models previously discussed in the state-of-the-art section: U-Net, LN, PAN, FPN. Then, we explore the design of different heterogeneous ensemble models by exploiting different decision fusion strategies, including *minimum*, *maximum*, *average*, and *median* operations. Figure 2 shows an overview of the heterogeneous ensemble schema.

To measure the effectiveness of our approach in addressing the segmentation task, we use well-known evaluation metrics: pixel accuracy, dice, precision, and recall [16].

Between 2021 and 2024, images were taken of patients' oral cavities during medical examinations at the Oral Medicine Unit of the P. Giaccone University Hospital in Palermo, Italy. Oral consultants took photographic images using both standard cameras and smartphones. We collected images of three different pathologies: neoplastic, aphthous, and traumatic. The lesions on the photos 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 datasets have been provided. The Oral-AI dataset is available at [17].

IV. Experiments and results

In this section, we present our experiments and results conducted to address two distinct segmentation problems: binary semantic segmentation, which distinguishes between healthy and injured tissue, and multi-class semantic segmentation, which distinguishes between three different pathologies.

The hyperparameters used for training the models were selected by setting the batch size to the maximum capacity of 32 samples of `NVIDIA a100` gpu adopted during the experiments and varying the learning rate between the following values $lr = [5e-4, 1e-5, 5e-6, 1e-6]$. After identifying the best hyperparameters for each model, we ran the training for 150 epochs 10 times to have statistical results. Source code available at [18].

A. Binary semantic segmentation

Table I presents the results for the binary semantic segmentation problem. U-Net emerged as the top-performing individual model with a DICE score of 73.5%. However, the most effective ensemble model was initially based on the max operation. To enhance performance, we optimized the binarization threshold on the validation set. Table II illustrates the DICE metric differences between ensemble models using the default binarization value of 0.5 and the optimized value that improved the DICE scores by 2% to 4%.

TABLE I: Binary segmentation on the test set. * median.

Model	Pixel Acc.	Dice	Prec	Recall
U-Net	94.5 ± .02	73.5 ± .15	79.9 ± .26	68.1 ± .32
LN	93.9 ± .01	69.1 ± .11	80.5 ± .27	60.5 ± .33
PAN	93.5 ± .04	70.9 ± .05	74.5 ± .33	67.8 ± .22
FPN	94.0 ± .07	70.3 ± .04	80.0 ± .13	63.3 ± .23
Ens.				
mean	94.7	73.7	83.4	61.4
med*	94.4	71.1	84.1	61.6
max	93.9	**74.9**	69.0	**82.0**
min	93.6	63.0	88.4	48.9

In Figure 3, we show prediction examples of different models. In (a), we can observe an ideal case in which each individual model, as well as the ensemble model, produces accurate and consistent masks. Case (b) illustrates the noise impact in the input images. In this case, the presence of saliva bubbles, due to a non-standardized acquisition process, leads to misclassification by the U-Net and PAN models. In (c), we present a neoplastic lesion characterized by a non-uniform color, where different models identify different parts of the lesion, reflecting the challenge of reaching a consensus on segmentation. Finally, (d) presents a case in which none of the models successfully recognizes the lesion, indicating the limitations of current methods in some complex scenarios, e.g. the lesion is very small.

B. Multi-class semantic segmentation.

Table III shows the results of the multiclass semantic segmentation problem. Dice score and Recall metrics are

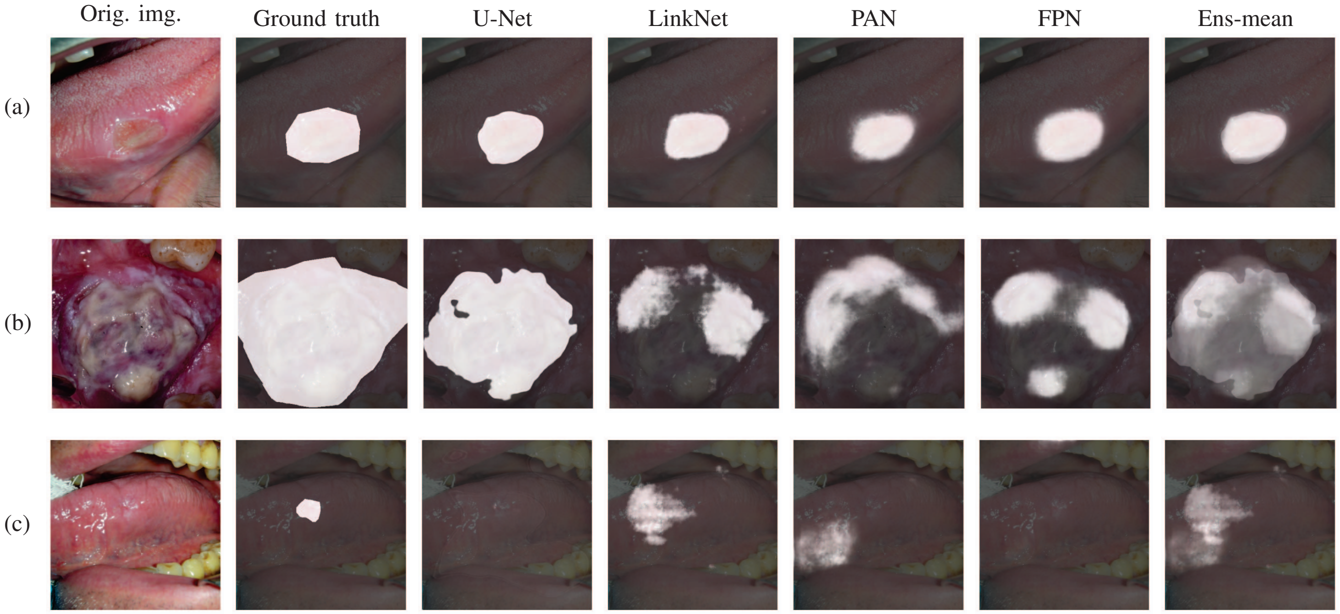

Fig. 3: Binary segmentation predictions of the four single models and the mean-based ensemble against the ground truth.

TABLE II: Threshold optimization for ensemble technique.

Ensemble	**Opt. Th.**	**Dice@0.5)**	**Dice@opt. th**
mean	0.27	73.7	**76.5**
median	0.07	71.1	75.0
max	0.80	74.9	76.0

presented both globally and by class/pathology. Recall is more important than Precision in the medical field as it is considered better to have false alarms than to miss a real case of disease. Failure to identify a tumor could cause it to grow and spread, endangering the patient's life.

From the results, we can observe that, as with the binary problem, the best single model is U-Net, which outperforms the other models on the Dice metric by about 3%. The marginal benefits of the ensemble relying on the max operation suggest that the individual models may already capture the essential features needed for accurate segmentation.

Figure 4 shows prediction examples. In particular, Figure 4(a) shows a traumatic injury in which each of the four individual models identified parts of it. The ensemble method aggregates these individual predictions, producing a more consistent segment approximating the ground truth. Figure 4(b) illustrates a neoplastic lesion well segmented by all models; however, the irregular shape of the lesion is a challenge, causing a degree of inaccuracy in all the predictions.

V. Conclusion

In this study, we proposed a medical image segmentation analysis for oral cancer. We tested different DL segmentation architectures including U-Net, LinkNet, PAN, and FPN; we also proposed an ensemble exploiting various fusion strategies on soft masks. The experiments have been conducted on a photographic dataset manually labeled by our clinical team to address two main segmentation tasks: distinguishing healthy tissue from diseased areas in a binary semantic segmentation task and identifying specific oral pathologies such as aphthous, traumatic, and neoplastic lesions by addressing multi-class segmentation. With a Dice score of 76.5%, the results show that the ensemble technique performs very well for binary semantic segmentation; additionally, improvements are evident but less significant for the multi-class problem, where the ensemble model overcomes the best single model of 0.3% on the Dice score.

Acknowledgment

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

TABLE III: Multi-class semantic segmentation evaluation on the test set.

	Dice				**Recall**			
Model	Avg.	Neoplastic	Aphthous	Traumatic	Avg.	Neoplastic	Aphthous	Traumatic
U-Net	71.2 ± .017	62.6 ± .015	60.4 ± .017	46.7 ± .019	62.9 ± .040	59.0 ± .037	49.2 ± .045	38.4 ± .044
LN	71.0 ± .014	61.3 ± .013	57.7 ± .011	43.3 ± .013	64.8 ± .039	60.0 ± .039	57.6 ± .032	33.2 ± .040
PAN	67.4 ± .008	57.1 ± .008	57.2 ± .010	38.4 ± .009	59.6 ± .028	52.8 ± .029	62.4 ± .026	27.6 ± .030
FPN	67.4 ± .010	57.8 ± .009	50.5 ± .011	48.0 ± .011	58.1 ± .031	52.5 ± .027	41.6 ± .028	38.9 ± .035
Ensemble								
mean	71.1	61.3	65.0	50.2	61.7	55.7	55.0	41.3
median	69.7	61.0	61.7	46.7	59.7	54.5	54.8	36.5
max	**71.5**	61.1	64.8	48.2	62.6	56.3	54.5	39.7
min	68.1	62.0	60.7	42.9	57.0	56.2	51.8	30.8

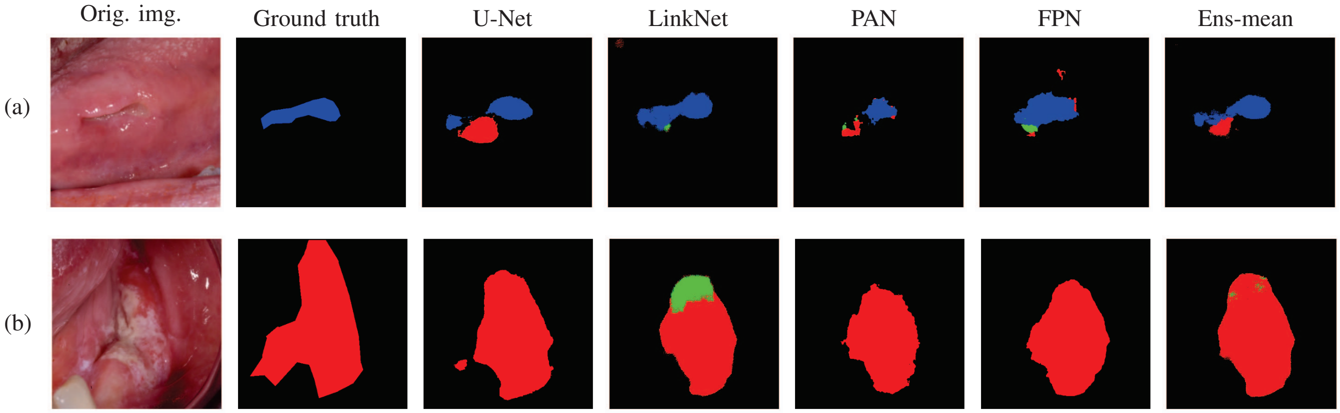

Fig. 4: Multi-class segmentation predictions of the four single models and the mean-based ensemble aggregation against the ground truth. In red neoplastic, in green aphthous, and in blue traumatic.

References

[1] J. Pillai, T. Chincholkar, R. Dixit, and M. Pandey, “A systematic review of proteomic biomarkers in oral squamous cell cancer,” *World Journal of Surgical Oncology*, vol. 19, pp. 1–28, 2021.

[2] S. He, R. Chakraborty, and S. Ranganathan, “Proliferation and apoptosis pathways and factors in oral squamous cell carcinoma,” *International journal of molecular sciences*, vol. 23, no. 3, p. 1562, 2022.

[3] S. Dey, A. K. Singh, A. K. Singh, K. Rawat, J. Banerjee, V. Agnihotri, and D. Upadhaya, “Critical pathways of oral squamous cell carcinoma: molecular biomarker and therapeutic intervention,” *Medical Oncology*, vol. 39, no. 3, p. 30, 2022.

[4] N. Salpea, P. Tzouveli, and D. Kollias, “Medical image segmentation: A review of modern architectures,” in *Computer Vision – ECCV 2022 Workshops* (L. Karlinsky, T. Michaeli, and K. Nishino, eds.), (Cham), pp. 691–708, Springer Nature Switzerland, 2023.

[5] P. Malhotra, S. Gupta, D. Koundal, A. Zaguia, W. Enbeyle, *et al.*, “Deep neural networks for medical image segmentation,” *Journal of Healthcare Engineering*, vol. 2022, 2022.

[6] H.-J. He, C. Zheng, and D.-W. Sun, “Chapter 2 - image segmentation techniques,” in *Computer Vision Technology for Food Quality Evaluation (Second Edition)* (D.-W. Sun, ed.), pp. 45–63, San Diego: Academic Press, second edition ed., 2016.

[7] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, “A survey on ensemble learning,” *Frontiers of Computer Science*, vol. 14, pp. 241–258, 2020.

[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] P. M. Speight, J. Epstein, O. Kujan, M. W. Lingen, T. Nagao, K. Ranganathan, and P. Vargas, “Screening for oral cancer—a perspective from the global oral cancer forum,” *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, vol. 123, no. 6, pp. 680–687, 2017.

[10] 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* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), (Cham), pp. 234–241, Springer International Publishing, 2015.

[11] A. Chaurasia and E. Culurciello, “Linknet: Exploiting encoder representations for efficient semantic segmentation,” in *2017 IEEE Visual Communications and Image Processing (VCIP)*, pp. 1–4, 2017.

[12] N. Khosravan, A. Mortazi, M. Wallace, and U. Bagci, “Pan: Projective adversarial network for medical image segmentation,” in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pp. 68–76, Springer, 2019.

[13] A. Kirillov, R. Girshick, K. He, and P. Dollár, “Panoptic feature pyramid networks,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6399–6408, 2019.

[14] M. Ganaie, M. Hu, A. Malik, M. Tanveer, and P. Suganthan, “Ensemble deep learning: A review,” *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105151, 2022.

[15] C. Gros, A. Lemay, and J. Cohen-Adad, “Softseg: Advantages of soft versus binary training for image segmentation,” *Medical Image Analysis*, vol. 71, p. 102038, 2021.

[16] D. Müller, I. Soto-Rey, and F. Kramer, “Towards a guideline for evaluation metrics in medical image segmentation,” *BMC Research Notes*, vol. 15, no. 1, p. 210, 2022.

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

[18] I. Cantini and M. Parola, “Github oral_segmentation code repository, github.com/marcoparola/oral_segmentation,” 2024.