ACTA OTORHINOLARYNGOLOGICA ITALICA 2023;43:283-290; doi: 10.14639/0392-100X-N2336

CLINICAL TECHNIQUES AND TECHNOLOGIES

Instance segmentation of upper aerodigestive tract cancer: site-specific outcomes

Instance segmentation nei tumori delle vie areo-digestive superiori

Alberto Paderno[1,2], Francesca Pia Villani[3], Milena Fior[2], Giulia Berretti[2], Francesca Gennarini[2], Gabriele Zigliani[2], Emanuela Ulaj[2], Claudia Montenegro[2], Alessandra Sordi[2], Claudio Sampieri[4], Giorgio Peretti[4], Sara Moccia[5,6], Cesare Piazza[1,2]

[1] Unit of Otorhinolaryngology, Head and Neck Surgery, ASST Spedali Civili of Brescia, Brescia, Italy; [2] Department of Medical and Surgical Specialties, Radiological Sciences, and Public Health, University of Brescia, School of Medicine, Brescia, Italy; [3] Department of Humanities, Università degli Studi di Macerata, Macerata, Italy; [4] Unit of Otorhinolaryngology, Head and Neck Surgery, IRCCS Ospedale Policlinico San Martino, Genoa, Italy; [5] The BioRobotics Institute, Scuola Superiore Sant'Anna, Pisa, Italy; [6] Department of Excellence in Robotics and AI, Scuola Superiore Sant'Anna, Pisa, Italy

SUMMARY

Objective. To achieve instance segmentation of upper aerodigestive tract (UADT) neoplasms using a deep learning (DL) algorithm, and to identify differences in its diagnostic performance in three different sites: larynx/hypopharynx, oral cavity and oropharynx.

Methods. A total of 1034 endoscopic images from 323 patients were examined under narrow band imaging (NBI). The Mask R-CNN algorithm was used for the analysis. The dataset split was: 935 training, 48 validation and 51 testing images. Dice Similarity Coefficient (Dsc) was the main outcome measure.

Results. Instance segmentation was effective in 76.5% of images. The mean Dsc was 0.90 ± 0.05. The algorithm correctly predicted 77.8%, 86.7% and 55.5% of lesions in the larynx/hypopharynx, oral cavity, and oropharynx, respectively. The mean Dsc was 0.90 ± 0.05 for the larynx/hypopharynx, 0.60 ± 0.26 for the oral cavity, and 0.81 ± 0.30 for the oropharynx. The analysis showed inferior diagnostic results in the oral cavity compared with the larynx/hypopharynx ($p < 0.001$).

Conclusions. The study confirms the feasibility of instance segmentation of UADT using DL algorithms and shows inferior diagnostic results in the oral cavity compared with other anatomic areas.

KEY WORDS: artificial intelligence, instance segmentation, deep learning, videomics

RIASSUNTO

Obiettivo. *Valutare l'instance segmentation utilizzando un algoritmo di intelligenza artificiale (IA) nei tumori delle vie aerodigestive superiori. Si è poi confrontata la performance diagnostica in tre diversi siti anatomici: laringe/ipofaringe, cavo orale e orofaringe.*

Metodi. *Sono state analizzate 1034 immagini NBI di 323 pazienti. Lo studio si è avvalso dell'algoritmo Mask-R-CNN. Il dataset è stato suddiviso in 935 immagini per il training, 48 per la validazione e 51 per il testing. Il principale outcome misurato è stato il Dice Similarity Coefficient (Dsc).*

Risultati. *L'algoritmo ha identificato le lesioni nel 76.5% delle immagini. Il valore totale del Dsc è stato 0,90 ± 0,05. Considerando le diverse sottosedi, sono state segmentate il 77,8% delle lesioni laringo-ipofaringee, l'86,7% delle orali e il 55,5% delle orofaringee. Il Dsc per le tre sottosedi è stato 0,90 ± 0,05, 0,60 ± 0,26 e 0,81 ± 0,30 rispettivamente per laringe/ipofaringe, cavo orale e orofaringe. L'analisi ha dimostrato risultati migliori per la laringe/ipofaringe e l'orofaringe rispetto a quelli del cavo orale ($p < 0,001$).*

Conclusioni. *Questo studio dimostra la fattibilità dell'instance segmentation nelle vie aerodigestive superiori utilizzando un algoritmo di IA, mostrando risultati inferiori nel cavo orale rispetto alle altre sottosedi.*

PAROLE CHIAVE: *intelligenza artificiale, instance segmentation, deep learning, videomica*

Received: October 14, 2022
Accepted: January 8, 2023

Correspondence
Alberto Paderno
Department of Otorhinolaryngology, Head and Neck Surgery, Spedali Civili di Brescia, piazzale Spedali Civili 1, Brescia, Italy
Tel. +39 030 3995319
E-mail: albpaderno@gmail.com

How to cite this article: Paderno A, Villani FP, Fior M et al. Instance segmentation of upper aerodigestive tract cancer: site-specific outcomes. Acta Otorhinolaryngol Ital Acta Otorhinolaryngol Ital 2023;43:283-290. https://doi.org/10.14639/0392-100X-N2336

© Società Italiana di Otorinolaringoiatria
e Chirurgia Cervico-Facciale

OPEN ACCESS

This is an open access article distributed in accordance with the CC-BY-NC-ND (Creative Commons Attribution-Non-Commercial-NoDerivatives 4.0 International) license. The article can be used by giving appropriate credit and mentioning the license, but only for non-commercial purposes and only in the original version. For further information: https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en

Introduction

The application of computer vision techniques in diagnostic videoendoscopies (i.e. Videomics)[1,2] is a promising research field that is currently showing a fast rate of growth in many medical specialties. The recent refinement of deep learning (DL) algorithms for image processing and their application in the medical field opened novel possibilities in the management of endoscopic exams that, in the past, had only subjective value. In particular, videoendoscopy is a key component in the management of upper aerodigestive tract (UADT) tumours, influencing their entire diagnostic process, treatment and follow-up[3]. Notwithstanding, it remains a operator-dependent and time-consuming procedure, which is substantially limited by the variables of human experience and perception. This is especially true when endoscopy is applied in conjunction with optical biopsy techniques such as Narrow Band Imaging (NBI)[4], requiring even more specialised training and adding a further layer of complexity and subjectivity. Finally, no easily classifiable and structured data can be drawn from these examinations, significantly limiting their integration with other technologies (e.g., cross sectional imaging, ultrasound, genomic markers, and so on). This is also highlighted by initial attempts to standardise endoscopic evaluation and improve implementation of new analytic techniques[5].

Our study aimed to explore the potential of a novel DL algorithm, Mask R-CNN[6], in the diagnostic approach to UADT squamous cell carcinoma (SCC). The primary goal was to detect and classify neoplastic lesions and, at the same time, precisely define their margins, a task overall defined as "instance segmentation". In fact, Mask R-CNN provides a flexible and general framework for object instance segmentation that can also be potentially applied to medical images. This approach combines elements from the tasks of object detection (where the goal is to localise the lesion using a bounding box), object classification[7] (where the purpose is to classify each pixel into a set of categories – e.g., tumour vs. normal mucosa), and semantic segmentation (where the aim is to automatically delineate the lesion's margins). Finally, we included in our analysis three different areas of the UADT (oral cavity, oropharynx, larynx/hypopharynx) in order to identify potential site-related differences in the diagnostic capability of this DL algorithm, an information that is still lacking in the current literature. In fact, studies assessing the value of artificial intelligence in endoscopy are generally focused on a single site and are difficult to generalise in the context of UADT SCC, which can arise from a wide variety of anatomical structures, as well as epithelial and mucosal types.

Materials and methods

A retrospective study was performed including videoendoscopies performed between September 2009 and January 2021 in patients treated at the Unit of Otorhinolaryngology – Head and Neck Surgery, University of Brescia, Italy for SCC of the UADT. A total of 7,567 videoendoscopies were collected from a dedicated archive. All recordings were anonymised and associated with the corresponding histopathologic report.

The study primary endpoint was the definition of the diagnostic accuracy (in terms of Dice Similarity Coefficient [Dsc]) of the Mask R-CNN algorithm when applied to NBI UADT videoendoscopic frames. The secondary endpoint was the comparison of the algorithm's Dsc in the three different anatomical areas herein considered.

Inclusion criteria were as follows:

- primary or recurrent SCC of the UADT (distinguished between those occurring in the oral cavity, oropharynx, and larynx/hypopharynx);
- NBI evaluation with adequate quality (without pooling of saliva, blood spots, swallowing reflex, coughing or other technical issues);
- available histological examination obtained at the time of videoendoscopy or subsequent surgery.

All patients were examined both under white light (WL) and NBI through transnasal videolaryngoscopy (HD Video Rhino-laryngoscope Olympus ENF-VH, ENF-VQ, or ENF-V2, Olympus Medical System Corporation, Tokyo, Japan) or through transoral endoscopy by 0° rigid telescope coupled to an Evis Exera II HD camera connected to an Evis Exera II CLV-180B/III CV-190 light source (Olympus Medical Systems Corporation, Tokyo, Japan). Endoscopic videos were selected independently by two otolaryngologists with extensive experience (at least 4 years) in endoscopic assessment of UADT lesions by NBI and independently reviewed by an adjunctive expert. Images were then manually quality-controlled, with exclusion of those that were blurred, obscured by blood or secretions, or without adequate NBI evaluation.

Image processing

Three representative frames per video were selected for every lesion and saved in jpeg format. The most representative NBI videoframe was chosen and subsequent frames at 0.3 second time intervals were then automatically selected. Frame annotation was performed manually using the LabelMe application[8]. Annotations consisted of a variable number of key points marking the lesion margins in the videoendoscopic frame taking into account positive NBI patterns. The resulting masks were then saved in json format and stored in a dedicated folder. Two clinical experts concomitantly annotated the images and a further review was performed by a senior staff member. When an agreement regarding lesion margins was not reached, the frame was excluded from the analysis.

After this selection process, a total of 1034 endoscopic images were obtained. Three different sub-datasets were generated according to the lesion primary site: oral cavity, oropharynx, and larynx/hypopharynx. In this way, the total frames analysed were 653 for the larynx/hypopharynx, 246 for the oral cavity, and 135 for the oropharynx.

Dataset

The dataset included 1034 images from 323 patients. For algorithm training and testing the dataset was split over patients and balancing the three classes into three sets: 935 images from 290 subjects for training, 48 images from 16 subjects for validation, and 51 images from 17 subjects for testing. All images were resized to the same dimension of 480 x 640 pixels.

DL analysis

In this work, Mask R-CNN [9] was used to segment the tumour in endoscopic frames. This convolutional neural network (CNN) consists of backbone, Region Proposal Network (RPN), and three heads for classification, bounding-box regression and segmentation (Fig. 1).

As backbone, we used the ResNet50 [10] combined with the Feature Pyramid Network (FPN) [11] to extract features from the input frame at multiple scales. Starting from the features computed with the backbone, the RPN identifies candidate regions containing the tumour. For each of the proposed regions, the final bounding box containing the tumour and the tumour segmentation are obtained from the three heads. To cope with the relatively limited size of the dataset, we used the weights computed on the COCO dataset [12] to initialise the layers of Mask R-CNN. To reduce the risk of overfitting, we performed on-the-fly data augmentation during training by applying: random brightness changes in the range (0.5, 1.1), random contrast changes in the range (0.8, 3) and random rotation in the range (-20, 20).

The model was trained for 100 epochs, using the Stochastic Gradient Descent (SGD) as optimiser with an initial learning rate of 0.001 and momentum of 0.9. We used a loss which is the combination of different contributions:

$$\boxed{L = L_{cls} + L_{box\_reg} + L_{rpn\_cls} + L_{rpn\_loc} + L_{mask}}$$

where $\boxed{L_{cls}}$ is the loss in the classification head, $\boxed{L_{box\_reg}}$ is the loss in bounding-box regression head, $\boxed{L_{rpn\_cls}}$ is the classification loss in the RPN, $\boxed{L_{rpn\_loc}}$ is the localisation loss in the RPN, and $\boxed{L_{mask}}$ is the loss in segmentation head. The loss equations can be found in the original Mask R-CNN paper [9].

Performance metrics and statistical analysis

As a primary endpoint the segmentation performance was evaluated using the Dsc, which is a statistical validation

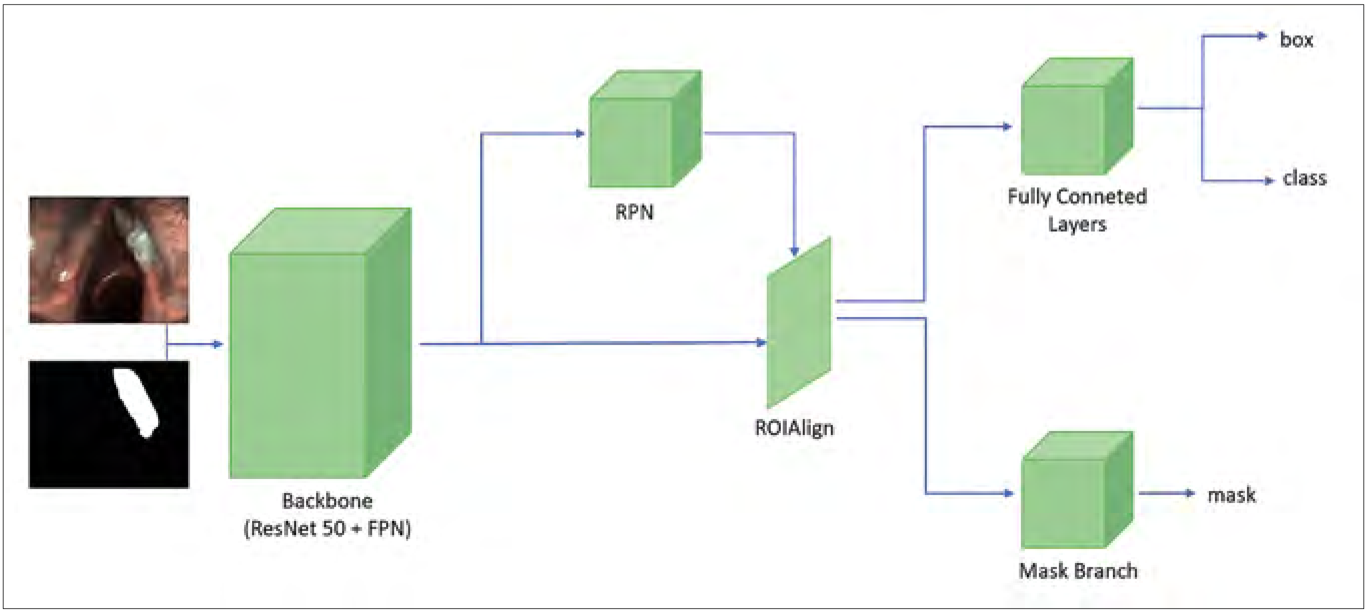

Figure 1. Schematic representation of the proposed architecture. The Mask R-CNN is made of a backbone (composed by a ResNet50 and a feature pyramid network), a region proposal network (RPN), ROIAlign, and three heads, for classification, bounding-box regression, and segmentation.

metric based on the spatial overlap between the predicted (A_{mask}) and ground-truth (A_{gt}) segmentation:

$$DSC = \frac{2 \text{ x } |A_{gt} \cap A_{mask}|}{|A_{gt}| + |A_{mask}|}$$

Dsc can assume values in a range from 0, indicating no overlap, to 1, indicating complete overlap.

Furthermore, outcomes were also evaluated using the following spatial overlap-based metrics:

Pixel accuracy (Acc) represents the percent of pixels in the image which are correctly classified.

It is defined as: $Acc = \frac{TP + TN}{TP + TN + FP + FN}$

where TP, TN, FP, FN denote the true positives, true negatives, false positives and false negatives, respectively.

Recall (Rec), also known as Sensitivity or True Positive Rate, defines the portion of positive pixels in the ground-truth which are also identified as positive in the predicted segmentation.

It is defined as: $Rec = \frac{TP}{TP + FN}$

Specificity (Spec), or True Negative Rate, measures the portion of negative pixels (background) in the ground-truth that are also identified as negative in the predicted segmentation.

It is defined as: $Spec = \frac{TN}{TN + FP}$

Precision (Prec), or Positive Predictive Value, measures how accurate the predictions are, i.e. the percentage of correct predictions.

It is defined as: $Prec = \frac{TP}{TP + FP}$

F1-score is a balance between precision and recall, also known as harmonic mean.

It is defined as: $F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}$

Intersection over Union (IoU), also referred to as Jaccard index, represents the area of overlap between the predicted segmentation and the ground truth divided by the area of union between the predicted segmentation and the ground truth.

It is defined as: $\text{IoU} = \frac{TP}{TP + FP + FN}$

Mean Average Precision (mAP), which represents the average of the area under the Recall-Precision curve, was also computed.

Outcomes were compared between the different subsites analysed using non-parametric statistics. The Kruskal-Wallis H-test was used for the overall comparison and the Mann-Whitney U rank test for pair comparisons. Statistical analysis was performed using Jupiter Notebook 6.4.5 with pandas 1.4.1 and ScyPy 1.8.0 libraries.

Results

Overall performance

The total number of images used for the test set was 51. The algorithm demonstrated the ability to correctly predict 39 of 51 images (76.5%). The average Dsc score was 0.79 (range, 0.26-0.97; standard deviation (SD), 0.22). Overall and site-specific performance metrics are summarised in Table I and Figure 2. Samples of the segmentation results are presented in Figure 3.

Laryngeal/hypopharyngeal lesions

The total number of laryngeal and hypopharyngeal lesions in the test set were 27 (52.9% of the test dataset). Out of that number, our algorithm correctly predicted 21 lesions (77.8%). The mean Dsc score was 0.90 ± 0.05, the first quartile was 0.90 and the third quartile 0.94 (Tab. I).

Oral lesions

The oral lesions comprised in the test set were 15 (29.4% of the total). The algorithm performed a correct prediction in 13 cases (86.7%). The mean Dsc score was 0.60 ± 0.26, the first quartile was 0.34 and the third quartile 0.84 (Tab. I).

Oropharyngeal lesions

In the test set, the oropharyngeal lesions were 9 of 51 images (17.6%). The algorithm predicted 5 images (55.5%). The mean value of Dsc score was 0.81 ± 0.30, the first quartile was 0.92 and the third quartile 0.95 (Tab. I).

Comparison between three different UADT sites

Results for each site are summarised in Table I. The overall diagnostic performance, defined by the Dsc score, was significantly different between the different sites ($p = 0.002$). Pairwise analysis showed that the difference was related to significantly inferior results in the oral cavity when compared with larynx/hypopharynx ($p < 0.001$).

Diagnostic results proved to be significantly correlated with the site analysed also considering other performance metrics: accuracy ($p < 0.001$), specificity ($p = 0.02$), IoU ($p = 0.002$), and F1 score ($p = 0.002$). As above, this difference is related to inferior results in the oral cavity *vs* larynx/hypopharynx. However, when considering accuracy, it is also possible to evidence a significant difference between oral cavity and oropharynx ($p = 0.03$).

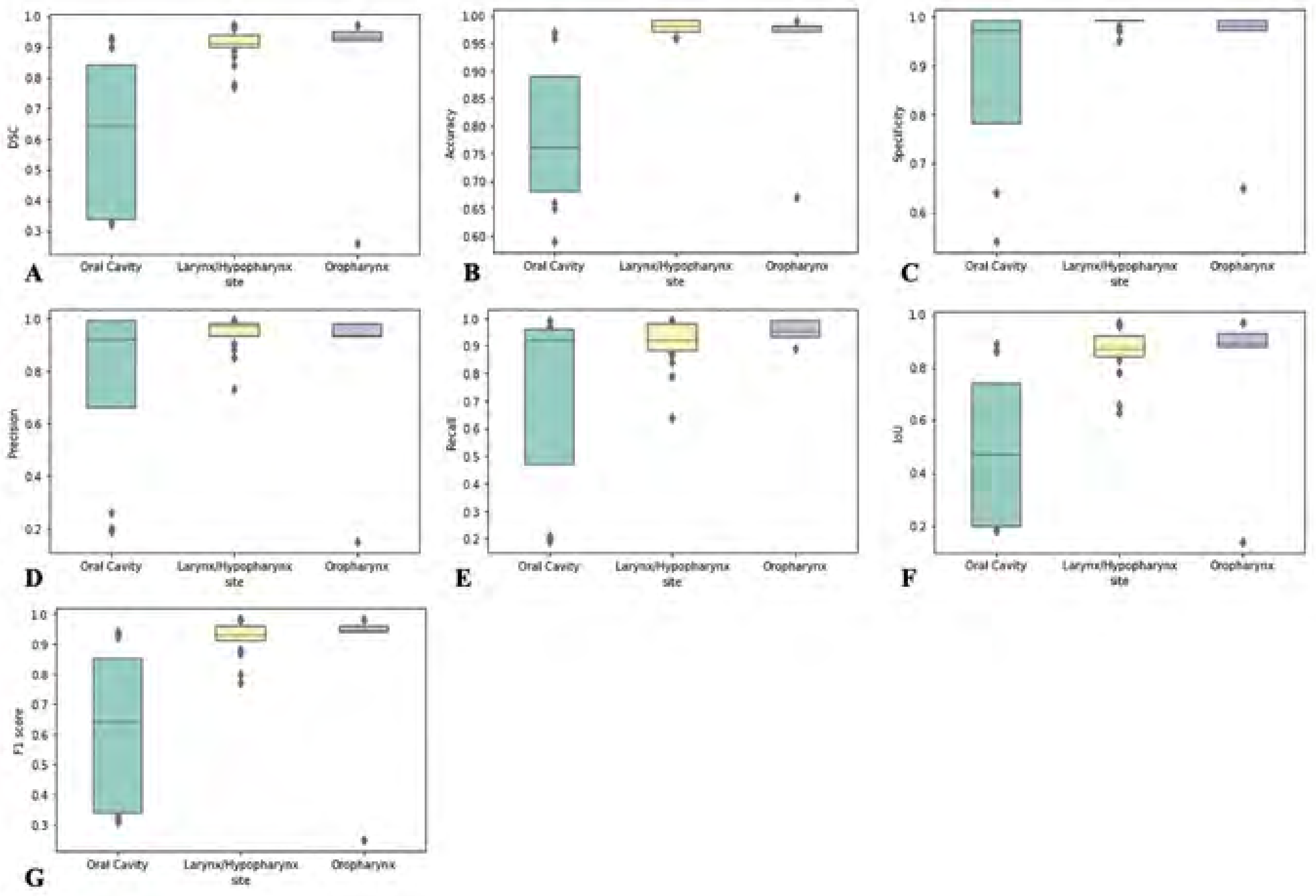

Figure 2. Box plots detailing the diagnostic accuracy of the algorithm in different sites according to various metrics. (**A**) Dice similarity coefficient (Dsc); (**B**) Accuracy; (**C**) Specificity; (**D**) Precision; (**E**) Recall; (**F**) Intersection over Union (IoU); (**G**) F1 score.

Table I. Summary of the diagnostic performance according to different metrics.

Mean value (SD)	Overall	Larynx/hypopharynx	Oral cavity	Oropharynx
Dsc	0.79 ± 0.23	0.90 ± 0.05	0.60± 0.26	0.80 ± 0.30
Accuracy	0.91 ± 0.12	0.98 ± 0.01	0.79 ± 0.13	0.92 ± 0.14
Specificity	0.93 ± 0.12	0.98 ± 0.01	0.86 ± 0.16	0.92 ± 0.15
Precision	0.85 ± 0.24	0.94 ± 0.06	0.73 ± 0.32	0.79 ± 0.36
Recall	0.86 ± 0.22	0.91 ± 0.08	0.73 ± 0.33	0.95 ± 0.04
IoU	0.73 ± 0.27	0.87 ± 0.09	0.49 ± 0.30	0.76 ± 0.14
F1 score	0.80 ± 0.23	0.92 ± 0.05	0.61 ± 0.27	0.81 ± 0.31

Dsc: Dice similarity coefficient; IoU: Intersection over Union; SD: standard deviation.

Discussion

In this study, we evaluated for the first time the specific task of instance segmentation in clinical endoscopy for head and neck SCC. The analysis included three sites of the UADT to allow comparison of the algorithm's diagnostic performance in different anatomical areas. The algorithm was able to identify and segment the lesion in 76.5% of cases, and showed remarkable diagnostic accuracy, especially in consideration of the complex task to be performed. Interestingly, results were significantly inferior in the oral cavity, where all outcome measures underperformed when compared with larynx/hypopharynx and, in some cases

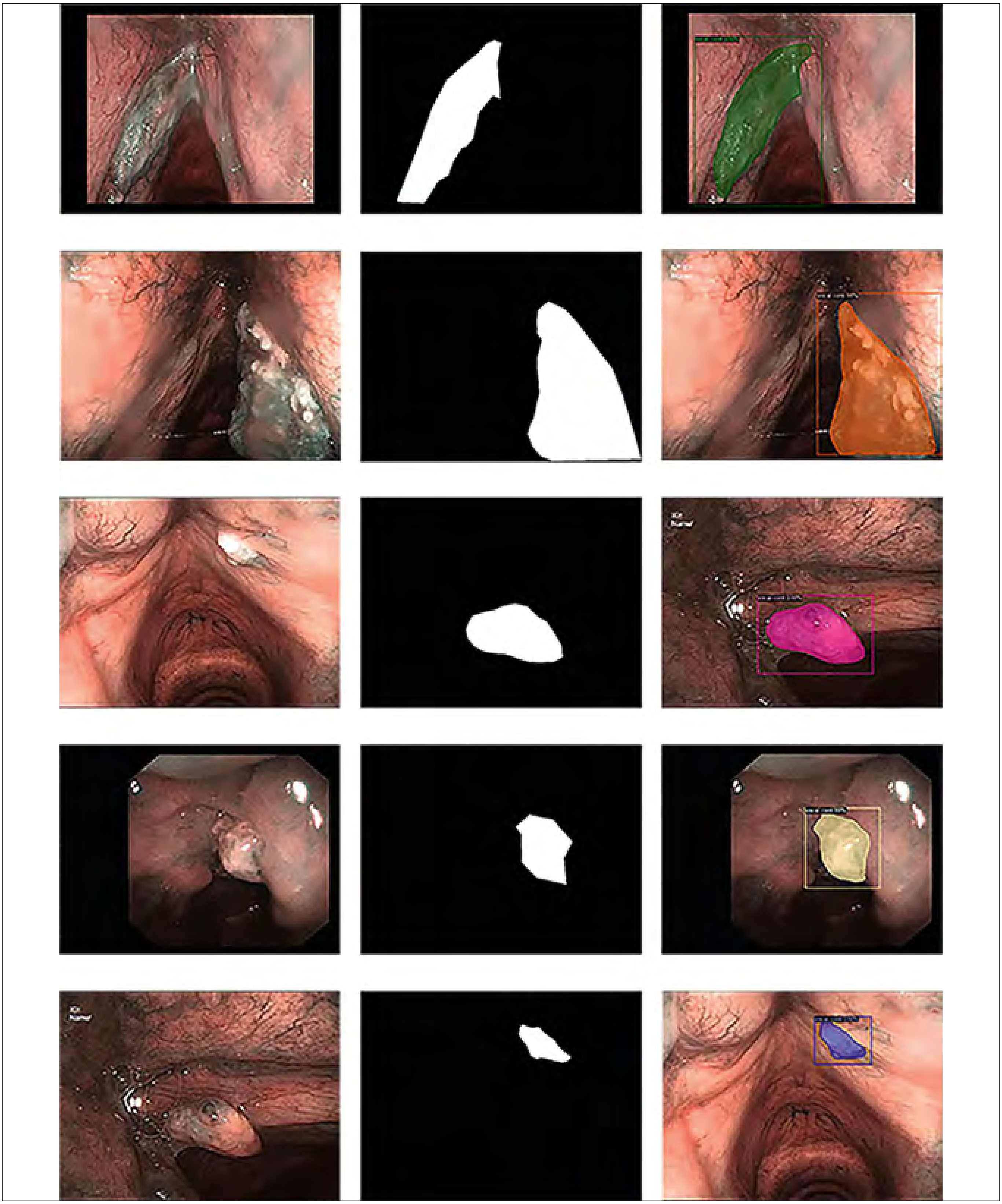

Figure 3. Visual samples of the segmentation results. From left to right: raw endoscopic frames, ground truth annotation, and predictions obtained with the proposed method.

(i.e., accuracy), oropharynx. This is in line with what previously observed by Piazza and coworkers [13] when applying bioendoscopic tools such as NBI. This result is possibly related to the wide array of epithelial subtypes observed in the oral cavity, adjunctive limits specifically correlated with oral examination (e.g., presence of light artifacts), and confounding factors (e.g., tongue blade, teeth, or dentures) that the ML software must learn to take into account.

Instance segmentation represents the ultimate step in video analysis since it allows at the same time detection, classification, and segmentation of multiple elements in each single frame, which is possible thanks to the integration of different analytic components in the same general algorithm. This approach is particularly suited to the context of UADT endoscopy since different alterations (e.g., concomitant inflammatory or benign lesions) can be frequently encountered in the field of view together with the target lesion, and due to the fact that patients with head and neck SCC can develop distinct islands of neoplastic or dysplastic mucosa (i.e., field of cancerisation) that might involve various portions of the videoframe, even without continuity.

In general, recent CNN-based methods have demonstrated remarkable results in segmentation of the UADT and proved to be well-suited for such a complex task. Laves et al. [14] first demonstrated that a weighted average ensemble network of UNet and ErfNet were the best suited for laryngeal segmentation of intra-operative images under direct laryngoscopy, with a mean IoU of 84.7%. However, different authors subsequently strived toward development of diagnostic algorithms that could be applied in real time in office-based and intra-operative endoscopy. Paderno et al. [15] explored the use of fully CNNs for real-time segmentation of SCC in the oral cavity and oropharynx. In this work, different architectures were compared detailing their diagnostic performance and inference time, demonstrating the possibility to achieve real-time segmentation. In accordance with previous findings in literature, the present study confirms that the oral cavity may have inferior diagnostic results due to the high variability of subsites when compared with other areas of the UADT (i.e., oropharynx, larynx, and hypopharynx). When dealing with normal laryngeal anatomy, Fehling et al. [16] explored the possibility to achieve a fully automated segmentation of the glottic area using a CNN in high-speed laryngeal videos. The algorithm obtained a Dsc over 0.85 for all subsites analysed. Finally, Li et al. [17] proposed a method to segment nasopharyngeal malignancies in endoscopic images based on DL, reaching an accuracy of 88.0%. However, progressive advances in automatic segmentation of the UADT can be observed thanks to a recent article by Azam et al. [2], in which SegMENT, a novel CNN-based segmentation model, outperformed previously published results on the external validation cohorts. The model was initially trained on WL and NBI endoscopic frames of laryngeal SCC, but also showed to be effective in the segmentation of independent frames of oral and oropharyngeal cancer. The authors stated that the model demonstrated potential for improved detection of early tumours, more precise biopsies and better selection of resection margins.

In general, results of automatic segmentation are inferior to those obtained in more straightforward tasks such as frame classification [18-20] or lesion detection [21,22] since a more in-depth conceptual model of UADT lesions is required to allow accurate definition of margins. However, semantic segmentation is a key objective when striving towards more complex tasks involving computer vision and human-machine interaction. In fact, other than providing a purely diagnostic tool, a comprehensive understanding of all UADT alterations and suspicious lesions may grant significant aid in intra-operative management. This is even more true when considering instance segmentation, which epitomises in itself all the needs and requirements of the visual examination of endoscopic images, allowing a full automatic understanding of complex endoscopic scenarios, even those involving more than one lesion and/or more than one pathology.

Potential issues have been addressed to limit biases related to the analysis technique:

- patients (and their related frames) in the training, validation, and test sets have been distinguished into separated groups to avoid overfitting;
- frames were annotated and reviewed by 3 experts to limit subjective errors;
- frame selection and data augmentation were performed to reduce the impact of artifacts or technical biases.

However, intrinsic limits should be acknowledged. In particular, the gold standard over which the algorithm has been trained (i.e., the "ground truth") is represented by an expert opinion of the tumor margins and not by the histopathological definition per se. In fact, as of today, it is not technically possible to provide a direct *in situ*, *in vivo* morphologic correlation between endoscopic images and their histopathological specimen.

Conflict of interest statement

The authors declare no conflict of interest.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

AP, FPV, FG, MF, GB, GZ, EU, CM, AS, SM, CP: contributed to data collection and analysis; AP, CP: performed manuscript preparation; AP, FPV, FG, SM, CP: performed final edits and revisions; AP, FPV, MF, GB, FG, GZ, EU, CM, AS, CS, GP, SM, CP: reviewed contributed conceptually to the article and approved the submitted version.

Ethical consideration

This study was approved by the Institutional Ethics Committee (please specify name of the Institution University of Brescia) (protocol number 4267).

The research was conducted ethically, with all study procedures being performed in accordance with the requirements of the World Medical Association's Declaration of Helsinki.

Written informed consent was obtained from each participant/patient for study participation and data publication.

References

[1] Paderno A, Holsinger FC, Piazza C. Videomics: bringing deep learning to diagnostic endoscopy. Curr Opin Otolaryngol Head Neck Surg 2021;29:143-148. https://doi.org/10.1097/MOO.0000000000000697

[2] Azam MA, Sampieri C, Ioppi A, et al. Videomics of the upper aero-digestive tract cancer: deep learning applied to white light and narrow band imaging for automatic segmentation of endoscopic images. Front Oncol 2022;12:900451. https://doi.org/10.3389/fonc.2022.900451

[3] Paderno A, Gennarini F, Sordi A, et al. Artificial intelligence in clinical endoscopy: insights in the field of videomics. Front Surg 2022;9:933297. https://doi.org/10.3389/fsurg.2022.933297

[4] Piazza C, Del Bon F, Paderno A, et al. The diagnostic value of narrow band imaging in different oral and oropharyngeal subsites. Eur Arch Otorhinolaryngol 2016;273:3347-3353. https://doi.org/10.1007/s00405-016-3925-5

[5] Nogal P, Buchwald M, Staśkiewicz M, et al. Endoluminal larynx anatomy model – towards facilitating deep learning and defining standards for medical images evaluation with artificial intelligence algorithms. Otolaryngol Pol 2022;76:1-9. https://doi.org/10.5604/01.3001.0015.9501

[6] He K, Gkioxari G, Dollár P, et al. Mask R-CNN. IEEE Trans Pattern Anal Mach Intell 2020;42:386-397. https://doi.org/10.1109/TPAMI.2018.2844175

[7] Cho WK, Lee YJ, Joo HA, et al. Diagnostic accuracies of laryngeal diseases using a convolutional neural network-based image classification system. Laryngoscope 2021;131:2558-2566. https://doi.org/10.1002/lary.29595

[8] Russell BC, Torralba A, Murphy KP et al. LabelMe: a database and web-based tool for image annotation. Int J Comput Vis 2008;77:57-173. https://doi.org/10.1007/s11263-007-0090-8

[9] He K, Gkioxari G, Dollár P, et al. Mask R-CNN, 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy; 2017. pp. 2980-2988. https://doi.org/10.1109/ICCV.2017.322

[10] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA; 2016. pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[11] Lin T, Dollár P, Girshick RB, et al. Feature pyramid networks for object detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA; 2017. pp. 936-944, https://doi.org/10.1109/CVPR.2017.106

[12] Lin T, Maire M, Belongie S, et al. Microsoft COCO: common objects in context. European Conference on Computer Vision. Cham: Springer; 2014. pp. 740-755.

[13] Piazza C, Del Bon F, Peretti G, et al. "Biologic endoscopy": optimization of upper aerodigestive tract cancer evaluation. Curr Opin Otolaryngol Head Neck Surg 2011;19:67-76. https://doi.org/10.1097/MOO.0b013e328344b3ed

[14] Laves MH, Bicker J, Kahrs LA, et al. A dataset of laryngeal endoscopic images with comparative study on convolution neural network-based semantic segmentation. Int J Comput Assist Radiol Surg 2019;14:483-492. https://doi.org/10.1007/s11548-018-01910-0

[15] Paderno A, Piazza C, Del Bon F, et al. Deep learning for automatic segmentation of oral and oropharyngeal cancer using narrow band imaging: preliminary experience in a clinical perspective. Front Oncol 2021;11:626602. https://doi.org/10.3389/fonc.2021.626602

[16] Fehling MK, Grosch F, Schuster ME et al. Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep Convolutional LSTM Network. PloS One 2020;15:E0227791. https://doi.org/10.1371/journal.pone.0227791

[17] Li C, Jing B, Ke L, et al. Development and validation of an endoscopic images-based deep learning model for detection with nasopharyngeal malignancies. Cancer Commun 2018;38:59. https://doi.org/10.1186/s40880-018-0325-9

[18] Song B, Sunny S, Uthoff RD, et al. Automatic classification of dual-modalilty, smartphone-based oral dysplasia and malignancy images using deep learning. Biomed Opt Express 2018;9:5318. https://doi.org/10.1364/BOE.9.005318

[19] Esmaeili N, Sharaf E, Gomes Ataide EJ, et al. Deep convolution neural network for laryngeal cancer classification on contact endoscopy-narrow band imaging. Sensors 2021;21:8157. https://doi.org/10.3390/s21238157

[20] Dunham ME, Kong KA, McWhorter AJ, et al. Optical biopsy: automated classification of airway endoscopic findings using a convolutional neural network. Laryngoscope 2022;132(Suppl. 4):S1-S8. https://doi.org/10.1002/lary.28708

[21] Inaba A, Hori K, Yoda Y, et al. Artificial intelligence system for detecting superficial laryngopharyngeal cancer with high efficiency of deep learning. Head Neck 2020;42:2581-2592. https://doi.org/10.1002/hed.26313

[22] Azam MA, Sampieri C, Ioppi A, et al. Deep learning applied to white light and narrow band imaging videolaryngoscopy: toward real-time laryngeal cancer detection. Laryngoscope 2022;132:1798-1806. https://doi.org/10.1002/lary.29960