The Laryngoscope

Real-Time Laryngeal Cancer Boundaries Delineation on White Light and Narrow-Band Imaging Laryngoscopy with Deep Learning

Claudio Sampieri, MD; Muhammad Adeel Azam, MSc; Alessandro Ioppi, MD; Chiara Baldini, MSc; Sara Moccia, PhD; Dahee Kim, PhD; Alessandro Tirrito, MD; Alberto Paderno, MD; Cesare Piazza, MD; Leonardo S. Mattos, PhD; Giorgio Peretti, MD

Objective: To investigate the potential of deep learning for automatically delineating (segmenting) laryngeal cancer superficial extent on endoscopic images and videos.

Methods: A retrospective study was conducted extracting and annotating white light (WL) and Narrow-Band Imaging (NBI) frames to train a segmentation model (*SegMENT-Plus*). Two external datasets were used for validation. The model's performances were compared with those of two otolaryngology residents. In addition, the model was tested on real intraoperative laryngoscopy videos.

Results: A total of 3933 images of laryngeal cancer from 557 patients were used. The model achieved the following median values (interquartile range): Dice Similarity Coefficient (DSC) = 0.83 (0.70–0.90), Intersection over Union (IoU) = 0.83 (0.73–0.90), Accuracy = 0.97 (0.95–0.99), Inference Speed = 25.6 (25.1–26.1) frames per second. The external testing cohorts comprised 156 and 200 images. *SegMENT-Plus* performed similarly on all three datasets for DSC ($p = 0.05$) and IoU ($p = 0.07$). No significant differences were noticed when separately analyzing WL and NBI test images on DSC ($p = 0.06$) and IoU ($p = 0.78$) and when analyzing the model versus the two residents on DSC ($p = 0.06$) and IoU (Senior vs. *SegMENT-Plus*, $p = 0.13$; Junior vs. *SegMENT-Plus*, $p = 1.00$). The model was then tested on real intraoperative laryngoscopy videos.

Conclusion: *SegMENT-Plus* can accurately delineate laryngeal cancer boundaries in endoscopic images, with performances equal to those of two otolaryngology residents. The results on the two external datasets demonstrate excellent generalization capabilities. The computation speed of the model allowed its application on videolaryngoscopies simulating real-time use. Clinical trials are needed to evaluate the role of this technology in surgical practice and resection margin improvement.

Key Words: artificial intelligence, laryngeal cancer, laryngoscopy, narrow-band imaging, segmentation, white light imaging.

Level of Evidence: III

***Laryngoscope*, 00:1–9, 2024**

INTRODUCTION

Transoral surgery nowadays represents the preferred approach to treat early-stage tumors of the upper aerodigestive tract (UADT), as it delivers sound oncological outcomes with minor morbidity.[1] This surgery is burdened by a high percentage of positive margins,[2,3] mainly due to the lack of precision of current methods to determine the full extent of tumors.

Narrow-band imaging (NBI) is an optical technique that enhances the submucosal vascularization using narrow bandwidth filters.[4] By analyzing the NBI vascular pattern, the otolaryngologist can distinguish normal from cancerous tissues and enhance the identification of tumor boundaries.[5] However, this technology is subjective and requires an extensive learning curve to be mastered.

From the Department of Experimental Medicine (DIMES) (C.S.), University of Genova, Genoa, Italy; Functional Unit of Head and Neck Tumors (C.S.), Hospital Clínic, Barcelona, Spain; Otorhinolaryngology Department (C.S.), Hospital Clínic, Barcelona, Spain; Department of Advanced Robotics (M.A.A., C.B., L.S.M.), Istituto Italiano di Tecnologia, Genoa, Italy; Dipartimento di Informatica, Bioingegneria, Robotica e Ingegneria dei Sistemi (DIBRIS) (M.A.A., C.B.), University of Genova, Genoa, Italy; Unit of Otorhinolaryngology – Head and Neck Surgery (A.I., A.T., G.P.), IRCCS Ospedale Policlinico San Martino, Genoa, Italy; Department of Surgical Sciences and Integrated Diagnostics (DISC) (A.I., A.T., G.P.), University of Genova, Genoa, Italy; Department of Otorhinolaryngology-Head and Neck Surgery (A.I.), "S. Chiara" Hospital, Azienda Provinciale per i Servizi Sanitari (APSS), Trento, Italy; The BioRobotics Institute and Department of Excellence in Robotics and AI (S.M.), Scuola Superiore Sant'Anna, Pisa, Italy; Department of Otorhinolaryngology (D.K.), Yonsei University College of Medicine, Seoul, Republic of Korea; Unit of Otorhinolaryngology – Head and Neck Surgery (A.P., C.P.), ASST Spedali Civili of Brescia, Brescia, Italy; and the Department of Medical and Surgical Specialties, Radiological Sciences, and Public Health (A.P., C.P.), University of Brescia, Brescia, Italy.

Additional supporting information may be found in the online version of this article.

Editor's Note: This Manuscript was accepted for publication on December 11, 2023.

Claudio Sampieri, Muhammad Adeel Azam, and Alessandro Ioppi have equally contributed to the manuscript.

The authors have no funding, financial relationships, or conflicts of interest to disclose.

Send correspondence to Alessandro Ioppi, Unit of Otorhinolaryngology – Head and Neck Surgery, IRCCS Ospedale Policlinico San Martino, Largo Rosanna Benzi 10, 16132 Genoa, Italy. Email: alessandroioppi@gmail.com

DOI: 10.1002/lary.31255

The application of data science to medicine is an emerging field that leverages artificial intelligence (AI), which can be used to automatically extract information from endoscopy images,[6] and has been applied for computer-aided detection,[7] image classification,[8] and tumor segmentation.[9] Currently, the most promising AI method for these tasks is Deep Learning (DL), which enables computers to automatically learn from data. In this field, semantic segmentation is a process that involves recognizing and labeling different categories of objects in the image, highlighting their contours with pixel-wise masks. DL-driven margins' delineation would represent a valuable support in clinical practice, potentially able to help clinicians better tailor the resections and reduce the amount of residual disease after surgery.

In a previous pilot study, we developed a DL model for UADT cancer segmentation on endoscopic images.[10] Our preliminary results were encouraging despite the limited dataset and the slow computing time of the model. In the present study, we developed a more reliable model (*SegMENT-Plus*) on an extended dataset of white light (WL) and NBI images, aiming to achieve high tumor delineation performance and fast processing time to implement it on real-time laryngeal cancer endoscopy. Furthermore, we tested the generalizability of the model on two external cohorts of laryngeal cancer endoscopic images.

MATERIALS AND METHODS

Data Acquisition

Videolaryngoscopies of laryngeal cancer patients performed between 2014 and 2020 at the Unit of Otolaryngology and Head and Neck Surgery of San Martino Hospital-University of Genova, Italy, were retrieved. The local Institutional Review Board approval was obtained (CER Liguria: 169/2022). Inclusion criteria were: (1) patients with a biopsy-proven laryngeal squamous cell carcinoma; (2) availability of the pre-treatment recorded videolaryngoscopy. Exclusion criteria were: (1) low quality of the images in terms of unclear view of the lesion boundaries due to the presence of blur, altered exposition, or saliva artifacts; (2) patient age less than 18 years old.

The videos were captured both in the office and in the operating room. In the office, they were performed using a flexible nasopharyngo-laryngoscope (HD Video Rhino-laryngoscope ENF-VH, Olympus Medical System Corporation, Tokyo, Japan) through a transnasal route. Patients undergoing transoral laser microsurgery were examined pre-operatively under general anesthesia with 0°, 30°, or 70° rigid telescopes (HD camera head connected to a Visera Elite CLV-S190 light source, Olympus Medical System Corporation, Tokyo, Japan). Both types of videolaryngoscopies were performed using WL and NBI. From each videolaryngoscopy, when available, five WL and five NBI frames with different view angles were extracted. These frames were selected to be the most representative of the lesion and to offer a clear view of its boundaries. Overall, the frames obtained from microlaryngoscopy generally appeared sharper and of higher definition. The collected images varied in terms of resolution, with widths and heights ranging from 768 to 1920 pixels and 576 to 1072 pixels, respectively.

These frames were annotated using CVAT annotation software (Intel Corporation, Santa Clara, United States)[11] by three expert physicians who had a specific background of at least 5 years of training in laryngology and NBI interpretation. The annotations resulted in a careful delimitation of the tumors' borders creating a pixel-by-pixel mask and labeling it "squamous cell carcinoma": from now on these will be referred to as ground-truth segmentations. If multiple lesions were visible, multiple segmentations were performed to select all the laryngeal cancer pixels in the image. If one physician was not sure about the margins' annotation, the images were reviewed collectively by the three experts. Finally, the frame was annotated based on the consensus of the majority of the experts. This image dataset (from now on referred to as the University of Genova dataset) was finally split into a training-validation set (90% of the images) and a test set (10% of the images) with a patient-wise distribution method so that the images from the same patients used for the training were not included in the testing set.

Two external datasets containing laryngeal cancer images both in WL and NBI were used to verify the generalizability of the model (not for training). They were provided by the unit of otolaryngology of the Spedali Civili—University of Brescia (Italy) and the department of otolaryngology of Severance Hospital—Yonsei University, Seoul (Republic of Korea). Both centers complied with the abovementioned annotation policy and the two datasets were annotated by a single clinician from each center, with a personal experience in laryngology and NBI image interpretation of at least 5 years.

Lastly, five unedited preoperative videolaryngoscopies of five different patients (not used for training) from the University of Genova were selected for testing the computational speed of the model and simulating a real-time laryngeal cancer segmentation during an examination. Figure 1 synthetizes the workflow of the study.

Deep Learning Model Development and Validation

The architecture of our DL laryngeal cancer segmentation model (*SegMENT-Plus*) is fully described in the supplemental material. The model was programmed using Python (version 3.9) in Keras (version 2.11.0) and Tensorflow (version 2.11.0). The experiments were conducted on a workstation with a Dual Intel Xeon Gold 5222 (3.8 GHz) CPU, 128 GB RAM, and an NVIDIA RTX A6000 GPU with 48 GB. The outcomes of *SegMENT-Plus* were evaluated by comparing the predicted segmentations with the ground-truth segmentations. Standard evaluation metrics for semantic segmentation were used as reported in the literature for this topic.[6,12,13] A classification of each pixel in the images as a true positive (TP), true negative (TN), false positive (FP), or false negative (FN) was used to derive the evaluation metrics below.

Accuracy: the percentage of pixels in the image that is correctly classified by the model.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

Precision (also known as positive predictive value) measures the percentage of pixels correctly recognized as carcinoma among all the pixels that the model has predicted as carcinoma.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall (also known as sensitivity) measures the percentage of pixels correctly recognized as carcinoma among all the pixels that should have been recognized as carcinoma.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Intersection over Union (IoU): is a metric that evaluates the similarity between the predicted segmentation and the ground truth segmentation (Supplemental Figure 1). It measures the proportion of the intersection of the predicted and ground truth segmentations with respect to their union:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}}$$

Dice similarity coefficient (DSC) is calculated as the ratio of twice the intersection of the predicted and ground truth segmentation and the sum of the areas of the predicted and ground truth segmentations.

$$\mathrm{DSC} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FN} + \mathrm{FP}}$$

As the most comprehensive representatives of a segmentation performance, DSC and IoU were selected as the primary outcome for statistical comparisons.

A standard metric for evaluating the inference speed of a DL model is frames per second (fps), which is defined as the number of frames of a video processed within a second. Since videolaryngoscopies are generally acquired with an image rate of 25 fps, an algorithm with a median processing time of around 25 fps is considered to be compatible with real-time implementation.[14,15]

Comparison with Human Physicians

To better evaluate the efficacy of *SegMENT-Plus*, we compared its segmentation performances with those of human physicians. A subset of 100 images from the University of Genova testing dataset was selected to be particularly representative of the task. The annotations were reviewed collectively by three experts. This subset was used to compare the performance of *SegMENT-Plus* with that of two human physicians, who were a second-year otolaryngology resident (junior resident) and a fourth-year resident (senior resident). These physicians were asked to independently annotate every laryngeal cancer image of the subset. Both of them were previously trained in the use of the annotation software. They were also familiar with endoscopies of the UADT and with NBI. Their annotations were analyzed with respect to the ground truth and the resulting IoU and DSC were computed. Finally, these results were compared with those from *SegMENT-Plus*.

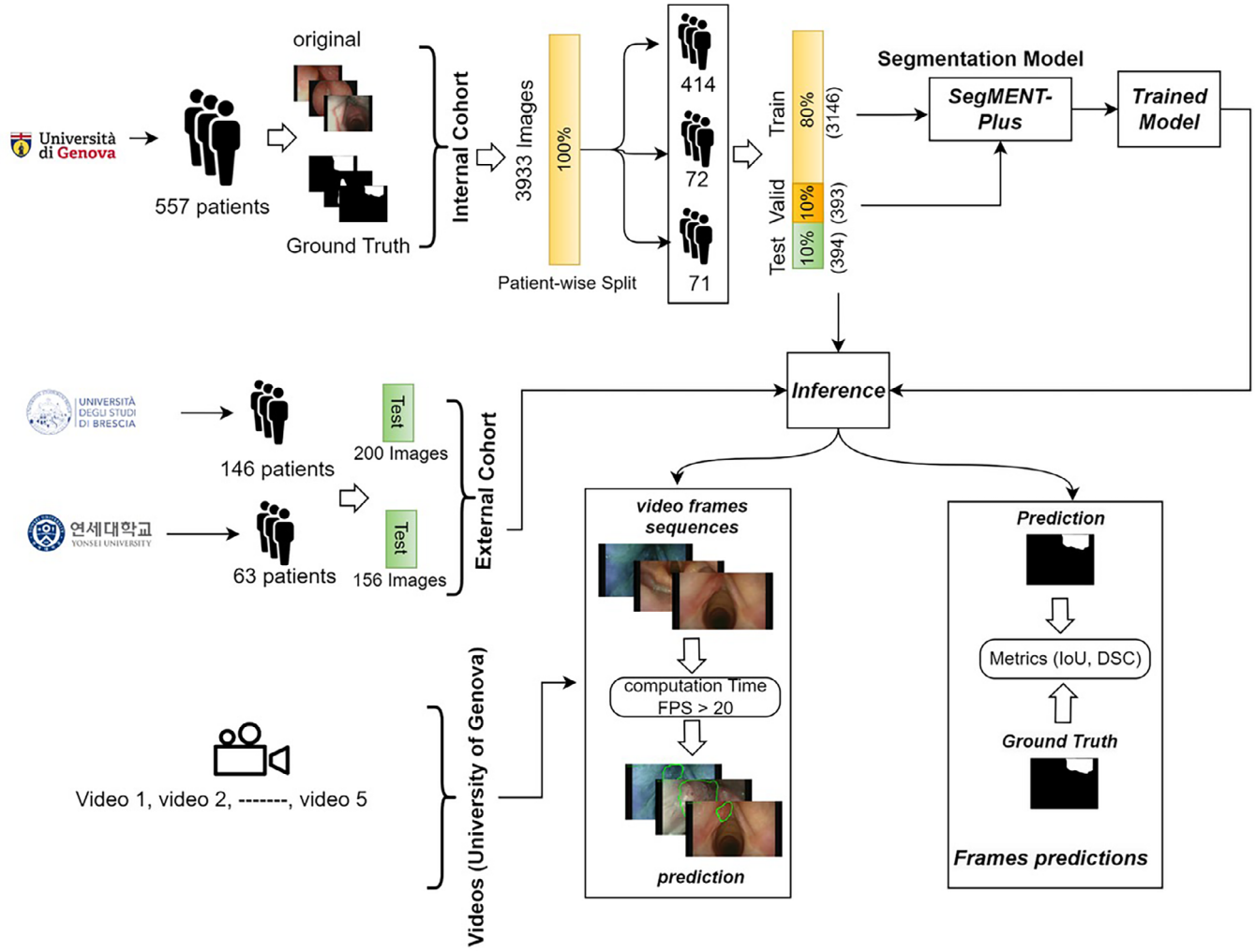

Fig. 1. Workflow diagram of the study. IoU, Intersection over Union; DSC, dice similarity coefficient; FPS, frames per second; valid, validation. [Color figure can be viewed in the online issue, which is available at www.laryngoscope.com.]

Statistical Analysis

Categorical variables (type of sets, gender) were summarized by counts and percentages, while continuous variables (age, DSC, IoU, Recall, Precision, Accuracy) were reported as medians ± interquartile range (IQR: 75th Q–25th Q). After performing the Shapiro–Wilk normality test, differences in the distribution of continuous variables between two independent groups were tested using the Mann–Whitney *U* test. Similarly, differences in distributions of continuous variables among more than two independent groups were assessed with Kruskal–Wallis

TABLE I.
Composition of the University of Genova LC Image Dataset. The Subdivision of Frames in Training-Validation and Test Datasets is Reported.

Dataset distribution	Type of examination	No. patients	No. images	Imaging	No. images	No. images
Training-validation	In-office	435	1208	WL	1884	3539
	Intra-operative	250	676			
	In-office	411	1153	NBI	1655	
	Intra-operative	206	502			
Test	In-office	45	129	WL	208	394
	Intra-operative	26	79			
	In-office	41	122	NBI	186	
	Intra-operative	20	64			

Note: The Acquisition setting (in-office vs. intra-operative) and imaging modality (white light = WL vs. Narrow-Band Imaging = NBI) are described.

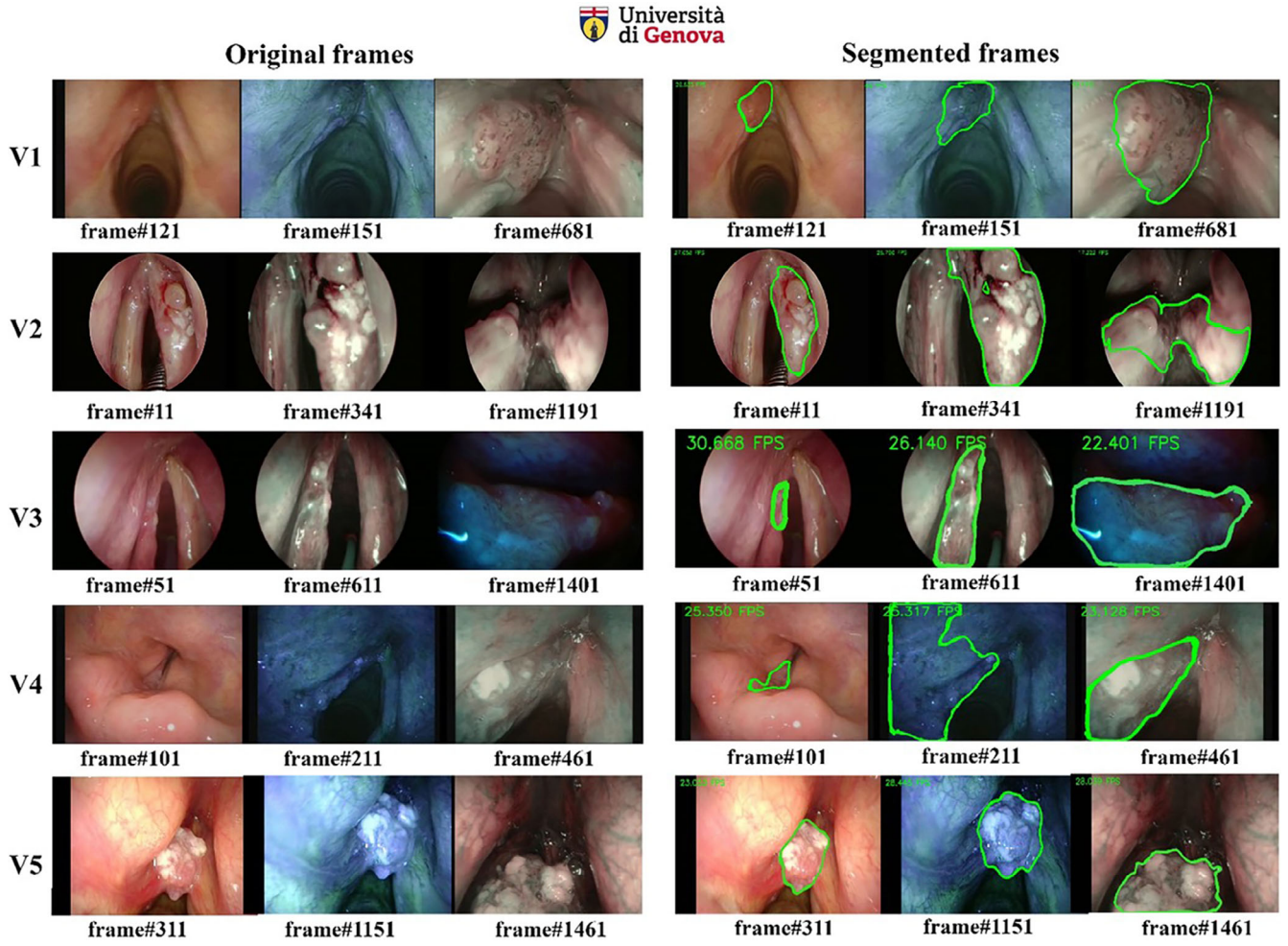

Fig. 2. Testing video frames extracted from five videolaryngoscopies. Each row represents a different video: the pictures in the panel on the left are white light or narrow-band imaging frames belonging to the original video, while the pictures in the panel on the right are the same frames after the elaboration with the deep learning model. [Color figure can be viewed in the online issue, which is available at www.laryngoscope.com.]

TABLE II.
Results of *SegMENT-Plus* on the University of Genova testing dataset and on the two external testing datasets (University of Brescia and Yonsei University) are reported as median (first and third quartiles).

Dataset	DSC	IoU	Recall	Precision	Accuracy	FPS
University of Genova	0.83 (0.70–0.90)	0.83 (0.73–0.90)	0.88 (0.74–0.96)	0.85 (0.70–0.94)	0.97 (0.95–0.99)	25.6 (25.1–26.1)
University of Brescia	0.81 (0.68–0.88)	0.81 (0.70–0.87)	0.90 (0.70–0.98)	0.82 (0.91–0.66)	0.96 (0.92–0.98)	26.2 (25.0–26.1)
Yonsei University	0.81 (0.55–0.89)	0.84 (0.68–0.89)	0.92 (0.72–0.98)	0.76 (0.49–0.86)	0.99 (0.97–0.99)	25.5 (25.4–27.0)

Abbreviations: DSC, dice similarity coefficient; FPS, frames per second; IoU, Intersection over Union.

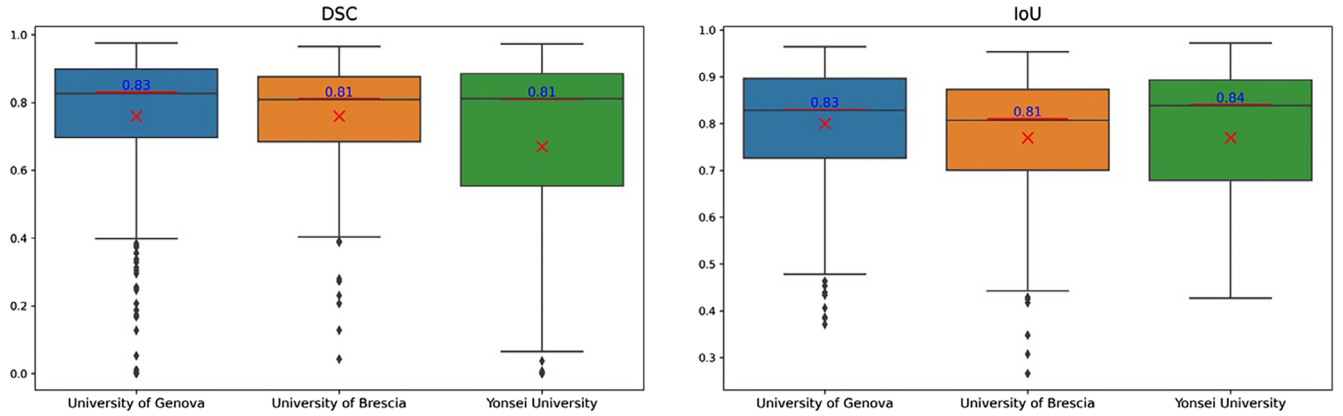

Fig. 3. Boxplots representing the segmentation performance of *SegMENT-Plus* on the three testing datasets. The horizontal bar inside the box represents the median value (also reported in numbers), the "x" represents the mean, and the box represents 50% of the distribution within the 1st and the 3rd quartiles. DSC, dice similarity coefficient; IoU, Intersection over union. [Color figure can be viewed in the online issue, which is available at www.laryngoscope.com.]

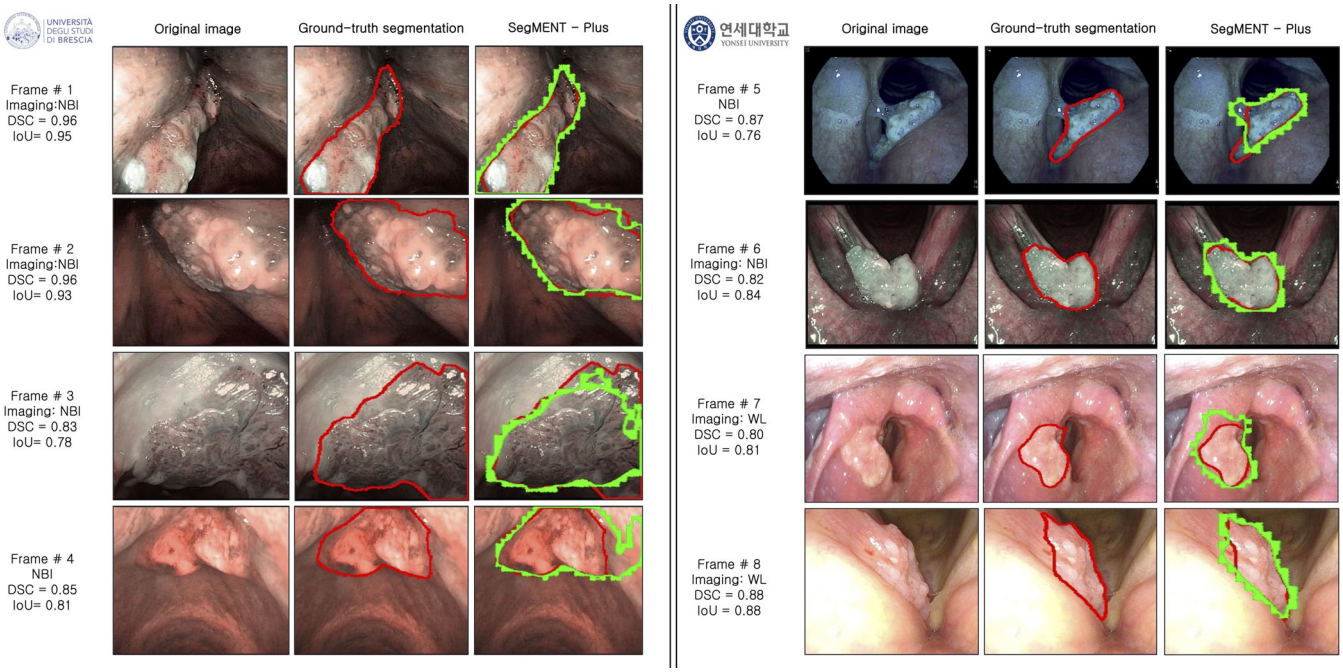

Fig. 4. Examples of *SegMENT-Plus* automatic segmentation of laryngeal carcinoma. The panel on the left reports four examples from the University of Brescia dataset. Frame #1 and #2 show a left and a right vocal cord carcinoma, respectively, while Frames #3 and #4 represent two different anterior commissure carcinomas. The panel on the right reports four examples from the Yonsei University dataset. Frame #5 represents a left vocal cord cancer, Frame #6 shows a commissural carcinoma, Frame #7 is about a right false vocal fold carcinoma while Frame #8 represents a right true vocal cord carcinoma. Red segmentations represent the ground truth provided by the experts, while green annotations are the areas predicted by the model. DSC, dice similarity coefficient; IoU, intersection over union; WL, white light; NBI, Narrow-Band Imaging. [Color figure can be viewed in the online issue, which is available at www.laryngoscope.com.]

test. In case of significant differences in the latter test, post-hoc multiple comparisons using Dunn's test were performed adjusting according to Bonferroni's method to control for the inflated Type I error. A two-sided $p < 0.05$ was considered significant. Statistical analysis was carried out using Python version 3.9 (packages scipy.stat and statmodels version 0.13.2).

RESULTS

The videolaryngoscopies of 557 patients examined at the unit of otolaryngology of IRCSS San Martino hospital, University of Genova were retrospectively retrieved. There were 485 (87.1%) males and 72 (12.9%) females with a median age of 67.0 ± 7. From these videolaryngoscopies, a total of 3933 images of laryngeal cancer were extracted to compose the image dataset for *SegMENT-Plus*. These frames were divided as follows: 3539 (89.9%) frames composed the training-validation set, while 394 (10.1%) frames were allocated to the test set. The characteristics and distribution of the University of Genova dataset are reported in Table I.

On the University of Genova test set, the model achieved median values of DSC = 0.83 ± 0.20, IoU = 0.83 ± 0.17, Recall = 0.88 ± 0.22, Precision = 0.85 ± 0.24, Accuracy = 0.97 ± 0.04 and median inference speed of 25.6 fps. A further comparison was made by analyzing separately the WL and NBI images of the testing set. Although the number of frames in the WL dataset was larger (2092 frames vs. 1841 frames), no significant differences in the model performance were noticed between WL and NBI images in terms of DSC = 0.81 ± 0.20 and 0.81 ± 0.19 respectively, $p = 0.06$; and IoU = 0.83 ± 0.14 and 0.82 ± 0.16 respectively, $p = 0.78$.

Finally, *SegMENT-Plus* was tested on five previously unseen videolaryngoscopies. The characteristics of the five testing videos and the computational time of the algorithm are reported in Supplemental Table 1. Examples of the original video frames and the corresponding ones processed by the DL model are shown in Figure 2 and two representative videos are available (Videos S1 and S2).

External Cohort Datasets

Two external datasets were used to test the generalizability of *SegMENT-Plus*. The University of Brescia dataset consisted of a total of 200 images (WL = 48

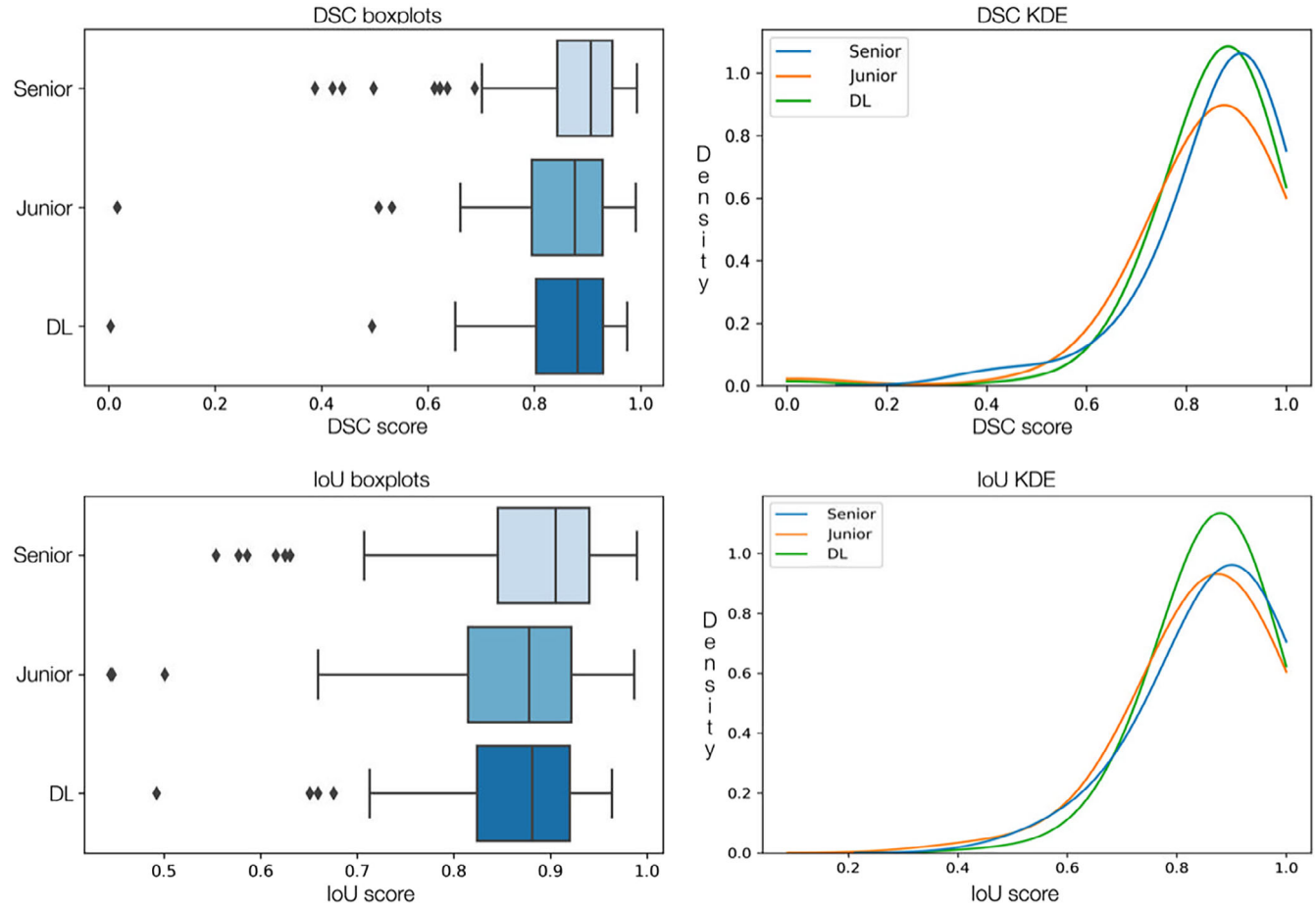

Fig. 5. Resident physicians' and *SegMENT-Plus* segmentation performances are represented with boxplots and kernel density estimates (KDE) curves. In the left figures, the boxplots are presented: the vertical bar inside the boxes represents the median value, while the box represents 50% of the distribution within the 1st and the 3rd quartiles. KDE curves, which are depicted on the right-hand side, visually illustrate how the outcomes scores of each rater tend to concentrate around specific values within the distribution. DSC, dice similarity coefficient; IoU, Intersection over union; DL, deep learning model *SegMENT-Plus*. [Color figure can be viewed in the online issue, which is available at www.laryngoscope.com.]

NBI = 152) from 146 patients (age = 65.0; males = 122), while the Yonsei University dataset comprised 156 frames (WL = 119; NBI = 37) from 63 patients (age = 64.0; males = 58). No significant differences were seen among the three sets for DSC ($p = 0.051$), and for IoU ($p = 0.066$). Median real-time inference speed was maintained for both cohorts (26.2 and 25.4 FPS, respectively). Table II summarizes the segmentation outcomes on the three test datasets. Figure 3 shows the boxplots of *SegMENT-Plus* performances on the three test sets. Figure 4 shows some automatic segmentation examples for the external validation datasets.

Comparison with Human Physicians

The result of the junior resident, senior resident, and *SegMENT-Plus* on a subcohort of 100 images, in comparisons with the ground truth annotations, were the following: senior resident, DSC = 0.91 ± 0.11, IoU = 0.91 ± 0.14; junior resident, DSC = 0.88 ± 0.14, IoU = 0.88 ± 0.11; *SegMENT-Plus* DSC = 0.89 ± 0.12, IoU = 0.89 ± 0.09. Based on the Kruskal–Wallis test, no significant differences were observed between the segmentation performances of the three groups for DSC, $p = 0.057$. For IoU, the Kruskal–Wallis test resulted in a $p = 0.046$. Pairwise comparisons with the post-hoc Dunn's test corrected with Bonferroni's method were not able to find any statistical differences in the three groups (Senior vs. Junior, $p = 0.07$; Senior vs. *SegMENT-Plus*, $p = 0.13$; Junior vs. *SegMENT-Plus*, $p = 1.00$). Overall, the closest p value to a statistical difference was among the two residents, while *SegMENT-Plus* did not differ significantly from both physicians. The Kernel density estimation

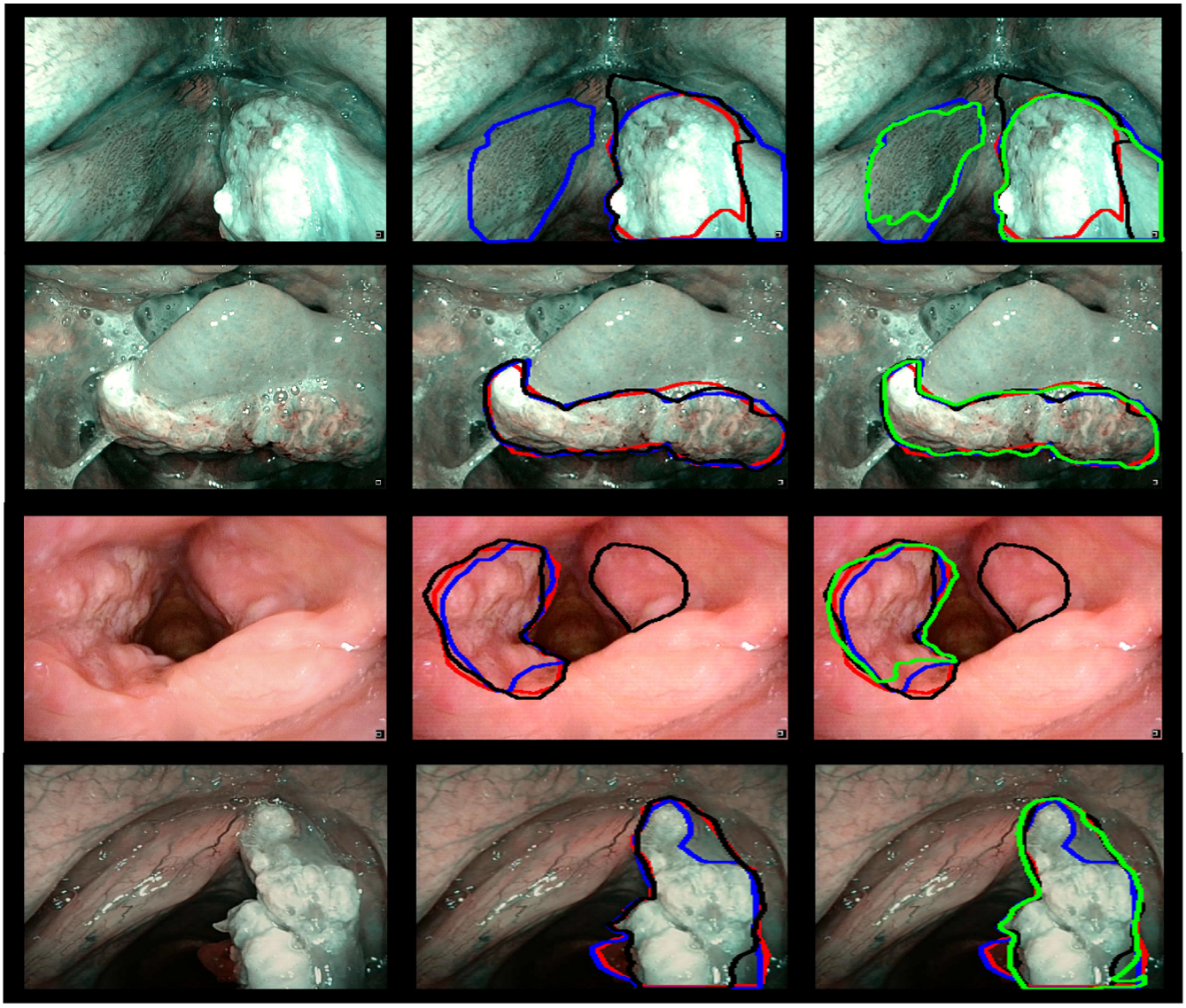

Fig. 6. Examples of the segmentations performed by the human physicians and *SegMENT-Plus* on the University of Genova test dataset. The column on the left shows the original frames, the column in the middle reports the annotation performed by the human physicians, and in the column on the right the DL model prediction is added. Blue = ground truth; Red = senior resident; black = junior resident; green = *SegMENT-Plus*. [Color figure can be viewed in the online issue, which is available at www.laryngoscope.com.]

curves and box plots are reported in Figure 5. Some graphical examples of the model's segmentation outputs compared to residents' and the ground truth ones are reported in Figure 6.

DISCUSSION

The precise delineation of surgical margins, especially for transoral approaches, represents a critical point in the treatment of laryngeal cancer. Indeed, the persistence of tumoral residue after this kind of procedure still represents a frequent issue.[3,16,17] Several technologies have been employed to enhance the visibility of malignant tissue on the UADT mucosa (i.e., bioendoscopy, autofluorescence, contact endoscopy), yet each of these tools suffers from specific drawbacks that have limited their use.[18]

Among those, NBI represents the most used and studied.[4,19,20] During surgery, NBI can provide accurate real-time tumor margin information and has been proven to reduce the rate of positive margins in laryngeal transoral surgery.[21] However, its reliability is hampered by conditions that alter the vascularization of the larynx, and its widespread adoption has been limited as the interpretation of the vascular patterns requires specific skills and a dedicated learning process.[22] Consequently, the introduction of computer-aided systems based on AI models able to detect cancerous tissue and identify its extension may represent a solution to this issue.

In the present study, we describe the application of a new semantic segmentation DL model able to automatically delineate laryngeal cancer from endoscopic images and videos. The model called *SegMENT-Plus*, is an improved version of our previous algorithm that demonstrated high segmentation capabilities on limited datasets of laryngeal, oral cavity, and oropharyngeal carcinomas.[10] *SegMENT-Plus* architecture was modified and trained on a larger dataset of laryngeal endoscopic images, to increase segmentation accuracy without compromising the computing time. When compared to our previous experiment,[10] the current DL model showed increased segmentation outcomes (DSC = 0.827 vs. 0.814; IoU = 0.826 vs. 0.686; accuracy = 0.972 vs. 0.969, respectively). Moreover, the training on a larger dataset (3146 vs. 547 images) allowed it to obtain robust performances even when tested on external datasets. Furthermore, with the current experiment hardware, *SegMENT-Plus* achieved a computing speed feasible for real-time implementation.

A possible evolution of semantic segmentation is instance segmentation, which generates a segmentation mask for every single object detected in the image distinguishing among different classes. Our group tried to explore this task discovering that the laryngeal and hypopharyngeal subsites are easier to process by the instance segmentation model compared to oral cavity and oropharynx.[23] Nevertheless, contrary to the present work, the number of images was limited (test set $n = 27$), and further studies are needed to corroborate those findings. As this is a new research field, the existing literature in this specific area is quite limited. To the best of our knowledge, the only similar work is the one by Ji et al.[24] In this article, the authors developed a DL model able to delineate laryngeal leukoplakia from laryngoscopic images, achieving an average DSC of 0.78 and an IoU of 0.83 on a smaller dataset of 649 WL laryngeal frames. That said, leucoplakias appear as more heterogeneous lesions when compared to laryngeal cancer, as they are characterized by high contrast margins, thus preventing a precise comparison.

Interestingly, in the present article, when *SegMENT-Plus* was tested separately on WL and NBI frames, no difference in the segmentation performance was noted. This shows how the model could overcome the difficulties in analyzing the complex NBI vascular pattern, approaching human experts' performances (ground truth). Therefore, AI might enhance the use of NBI even in less experienced centers by improving the accuracy of lesion detection and margin identification, regardless of the operator's familiarity with this technique.

The implementation of AI for automatic segmentation of laryngeal cancer in endoscopy has the potential to significantly impact clinical practice in the future. First, these models can improve the accuracy of tumor boundaries delineation, reducing human error and discrepancies in positive margins rates among different centers. Secondly, automatic segmentation can enhance both the surgical and non-surgical treatment of laryngeal cancer. In the first case, it can aid in surgery planning, allowing for a more precise surgical resection and reducing the risk of complications. In the second scenario, it can be used to objectively monitor treatment response over time, enabling a more precise assessment of treatment efficacy combining information obtained by radiology assessments and endoscopic evaluations. Finally, it can be used to automatically collect large volumes of data from multiple endoscopies, enabling more comprehensive research on laryngeal cancer.

Indeed, models like *SegMENT-Plus* are not intended to replace the clinician but rather to represent a support in decision-making, offering potential for standardization toward more equitable treatment. At the moment, as shown by the results obtained from the comparison with human physicians, the model did not achieve an identical segmentation as the expert clinicians, but its results were rather similar to those of a last year resident, suggesting that further improvements in the model's architecture and in the training dataset are necessary.

This study has limitations that should be acknowledged. First, the training and testing datasets are still limited. They should be increased and enriched with other external images, such as frames gathered from different video sources or with different enhancing filter modality (other than NBI). Second, the ground truth annotations lack external cross-validation while inter- and intra-annotators' reliability was not assessed in this article. This aspect is particularly relevant in this field and should be addressed in future studies in view of clinical trials. Finally, the retrospective nature of the work based on a selection of high-quality images represents an intrinsic bias that should be corrected in future works with a prospective evaluation of the technology.

CONCLUSION

The proposed DL segmentation model was able to accurately delineate laryngeal cancer boundaries in endoscopic images and videos. The high diagnostic performances were also maintained when tested on two different external datasets demonstrating robust generalization capabilities. The fast computation speed of the model allowed us to successfully apply it on videolaryngoscopies showing potential for real-time use. Based on this data, a clinical implementation is feasible for testing the model's benefit in a real-life setting.

ACKNOWLEDGMENTS

This work was carried out within the framework of the project "RAISE-Robotics and AI for Socioeconomic Empowerment" and has been partially supported by European Union-NextGenerationEU. The views and opinions expressed herein are those of the authors alone and do not necessarily reflect those of the European Union or the European Commission.

BIBLIOGRAPHY

1. Baird BJ, Sung CK, Beadle BM, Divi V. Treatment of early-stage laryngeal cancer: a comparison of treatment options. *Oral Oncol*. 2018;87:8-16.
2. Fiz I, Koelmel JC, Sittel C. Nature and role of surgical margins in transoral laser microsurgery for early and intermediate glottic cancer. *Curr Opin Otolaryngol Head Neck Surg*. 2018;26(2):78-83.
3. Gorphe P, Simon C. A systematic review and meta-analysis of margins in transoral surgery for oropharyngeal carcinoma. *Oral Oncol*. 2019;98:69-77.
4. Vilaseca I, Valls-Mateus M, Nogués A, et al. Usefulness of office examination with narrow band imaging for the diagnosis of head and neck squamous cell carcinoma and follow-up of premalignant lesions. *Head Neck*. 2017;39(9):1854-1863.
5. Garofolo S, Piazza C, del Bon F, et al. Intraoperative narrow band imaging better delineates superficial resection margins during transoral laser microsurgery for early glottic cancer. *Ann Otol Rhinol Laryngol*. 2015; 124(4):294-298.
6. Sampieri C, Baldini C, Azam MA, et al. State of the art review artificial intelligence for upper aerodigestive tract endoscopy and laryngoscopy: a guide for physicians and state-of-the-art review. *Otolaryngol Head Neck Surg*. 2023;2023(00):1-19.
7. 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(9):1798-1806.
8. Dunham ME, Kong KA, Mcwhorter AJ, Adkins LK. Optical biopsy: automated classification of airway endoscopic findings using a convolutional neural network. Published online 2020.
9. 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;11:11.
10. 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; 12:12.
11. Computer Vision Annotation Tool.
12. Yue G, Zhuo G, Li S, et al. Benchmarking polyp segmentation methods in narrow-band imaging colonoscopy images. *IEEE J Biomed Health Inform* Published Online July 1. 2023;27:3360-3371.
13. Ali S, Zhou F, Braden B, et al. An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. *Sci Rep*. 2020; 10(1):1-15.
14. Yang XX, Li Z, Shao XJ, et al. Real-time artificial intelligence for endoscopic diagnosis of early esophageal squamous cell cancer (with video). *Dig Endosc*. 2021;33(7):1075-1084.
15. Wang P, Xiao X, Glissen Brown JR, et al. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nat Biomed Eng*. 2018;2(10):741-748.
16. Fiz I, Mazzola F, Fiz F, et al. Impact of close and positive margins in transoral laser microsurgery for TIS-T2 glottic cancer. *Front Oncol*. 2017; 7(OCT):1-9.
17. Sampieri C, Costantino A, Spriano G, Peretti G, de Virgilio A, Kim SH. Role of surgical margins in transoral robotic surgery: a question yet to be answered. *Oral Oncol*. 2022;133:106043.
18. de Kleijn BJ, Heldens GTN, Herruer JM, et al. Intraoperative imaging techniques to improve surgical resection margins of oropharyngeal squamous cell cancer: a comprehensive review of current literature. *Cancers (Basel)*. 2023;15(3):1-25.
19. Lin YC, Watanabe A, Chen WC, Lee KF, Lee IL, Wang WH. Narrowband imaging for early detection of malignant tumors and radiation effect after treatment of head and neck cancer. *Arch Otolaryngol Head Neck Surg*. 2010;136(3):234-239.
20. Rosenthal EL. Optical imaging of head and neck cancer: opportunities and challenges. *JAMA Otolaryngol Head Neck Surg*. 2014;140(2):93-94.
21. Zwakenberg MA, Westra JM, Halmos GB, Wedman J, van der Laan BFAM, Plaat BEC. Narrow-band imaging in transoral laser surgery for early glottic cancer: a randomized controlled trial. *Otolaryngol Head Neck Surg (United States)*. Published online. 2023;39(7):1343-1348.
22. Valls-Mateus M, Nogués-Sabaté A, Blanch JL, Bernal-Sprekelsen M, Avilés-Jurado FX, Vilaseca I. Narrow band imaging for head and neck malignancies: lessons learned from mistakes. *Head Neck*. 2018;40(6): 1164-1173.
23. Paderno A, Villani FP, Fior M, et al. Instance segmentation of upper aerodigestive tract cancer: site-specific outcomes. *Acta Otorhinolaryngol Ital*. 2023;43(4):283-290.
24. Ji B, Ren J, Zheng X, et al. A multi-scale recurrent fully convolution neural network for laryngeal leukoplakia segmentation. *Biomed Signal Process Control*. 2020;59:101913.