UNIVERSITY OF GENOVA

PHD PROGRAM IN BIOENGINEERING AND ROBOTICS

Cycle XXXVI (2020-2023)

Upper Aero Digestive Tract Cancer Diagnosis using Deep Learning Methods

by

Muhammad Adeel Azam

Thesis submitted for the degree of Doctor of Philosophy (36th cycle)

Dr. Leonardo De Mattos	Supervisor
Dr. Sara Moccia	Co-Supervisor
Prof. Paolo Massobrio	Head of the PhD program

Thesis Jury

Dr. Elena De Momi, Politecnico di Milano	External examiner
Dr. Maria Koskinopoulou, Heriot Watt University	External examiner

Biomedical Robotics Lab, Department of Advanced Robotics, Italian Institute of Technology.

Department of Informatics, Bioengineering, Robotics and Systems Engineering.

December 2023

Declaration

I hereby declare that the contents of this dissertation, except for references, are entirely original and have not been submitted for consideration for any other degree or institution. My dissertation is entirely my own work, with all the materials provided. This dissertation relies significantly from my published work including all the methodology and materials. I certify that there are no copyright concerns related to outside sources, and I have complete authority over the provided content and methods, both of which will be released with my consent.

Muhammad Adeel Azam
December 2023

Acknowledgements

I want to express deep appreciation to everyone who has assisted me in my PhD studies. First and foremost, I would like to thank Dr. Leonardo De Mattos, Head of the Biomedical Robotics Laboratory at the Italian Institute of Technology (IIT, Genoa), for giving me the opportunity to be a part of this distinguished group and for his constant encouragement throughout this vital research experience. I will always be grateful to my supervisor, Dr. Leonardo De Mattos and co-supervisor, Dr. Sara Moccia, for their essential advice and guidance during my doctoral studies. Achieving this milestone would not have been possible without their help in several areas, including paper writing, paper evaluations, and thesis preparation. I would like to express my gratitude to Prof. Giorgio Peretti for sharing his knowledge and experience as well as for his cooperative efforts in the field of endoscopy medical imaging. Drs. Claudio Sampieri and Alessandro Ioppi deserve special recognition for their assistance and for providing the datasets that I used in my research. Their special fusion of technical innovation and clinical experience made my three years of research much more pleasant. I would like to thank my collaborators, Dr. Veronica Penza, Chiara Baldini, Shunlei Li, and Ajay Gunalan, for their assistance in understanding the background and fundamentals of endoscopy image processing. I would like to express my gratitude to my father, Azam, for his consistent support while I pursued this ambition. I also want to thank Zareena, my mother, for her love and encouragement. Sana, my wife, deserves special thanks for her devotion and love during this entire process. I also want to thank Ali Adeel, my son, whose presence made me more energetic and productive. Throughout this journey, my brother Ashir, and sisters Urooj, Komal, and Fatima have given me support and encouragement. Their compassion and encouragement have been a source of strength for me. I'd like to thank all the ADVR members and friends for their support and sound suggestions. I've really liked our time together. Finally, I want to express my gratitude to everyone at IIT for their assistance and consideration.

Muhammad Adeel Azam
December 2023, Genova

Abstract

Objective: Narrow band imaging (NBI) and white light (WL) are endoscopic techniques to visualize upper aero digestive tract (UADT) cancers. However, these imaging techniques are less effective for diagnosing tumors in less competent centers since they depend on skilled medical experts. Recently, there has been evidence that deep learning (DL) has potential applications in UADT video endoscopy. This research aims to develop a DL for the automatic identification and delineation of UADT cancer.

Approach: In both WL and NBI frames, the YOLO DL model (YOLOv5s with YOLOv5m) ensemble, was used to diagnose laryngeal squamous cell carcinoma (LSCC). Six external LSCC laryngoscopy videos were tested in real-time for cancer detection. The *SegMENT* is a segmentation convolution neural networks (CNN), model proposed based on a modified *DeepLabV3+* model for precise UADT delineation using an in-domain transfer learning ensemble technique. Its accuracy was further validated on external datasets with NBI images of oral cavity SCC (OSCC) and oropharyngeal SCC (OPSCC). The *SegMENT-Plus* is the improved version of *SegMENT* model designed for large LSCC datasets. SegMENT-*Plus* used *EfficientNetB5* backbone as an encoder with a modified atrous spatial pyramid pooling (*m-ASPP*) block. The attentions blocks (SE and CBAM) were integrated into *m-ASPP* module to improve cancer segmentation. The *m-ASPP* was used to extract local and global LSCC features to overcome the limitation of conventional *ASPP* modules in literature. *SegMENT-Plus* was evaluated using a multi-center dataset from three hospitals (Genoa, Brescia, Seoul South Korea). The model was tested on LSCC frames, the delineation performance was compared with three otolaryngology experts. The unseen intraoperative laryngoscopy videos also validated for real-time performance. The *SegMENT-Plus* was compared with its predecessor *SegMENT* and other DL models (*UNET, ResUNET, DeepLabv3+, DoubleUET,*).

Main results: In the LSCC detection task, 219 patients from Genoa, Italy were enrolled, and were provided 624 LSCC video frames. YOLO models were trained using an 82.6% training set, an 8.2% validation set, and a 9.2% testing set. The ensemble algorithm (*YOLOv5s with YOLOv5m —Test Time Augmentation*) achieved top LSCC detection with 66% Precision, 62% Recall, and 63% mean Average Precision at 0.5 intersection over union (IoU). The average computation time per frame on laryngoscopy videos was 0.026 seconds. The *SegMENT* model for the UADT cancer delineation was developed using 219 patients (624 larynx frames), and external validation from Brescia, Italy for the OPSCC and OCSCC cohorts involved 116 and 102 NBI images, respectively. The *SegMENT* model achieved 0.68% IoU and 0.81% dice coefficient (DSC), outperforming other DL models. The DSC values in the OCSCC and OPSCC datasets improved significantly, with median DSC values

of 10.3% and 11.9%, respectively. This study includes 557 patients with 3933 laryngeal images from Genoa, Italy to the development of *SegMENT-Plus* to improve LDCC delineation. The optimal performance and generalization of the algorithm were confirmed by external testing cohorts from Seoul, South Korea, and Brescia, Italy. The external cohorts showed DSC between 81.4% and 84.9% and IoU between 81.8% and 85.7%.
Significance: The study identified a suitable CNN model for LSCC detection in WL and NBI video laryngoscopes. *SegMENT* outperformed previous results in external validation cohorts, showing promise for precise tumor segmentation. *SegMENT-Plus* holds the potential for improved early tumor detection and delineation, laying the foundation for a clinical system in LSCC margin delineation.

Keywords: Larynx cancer, deep learning, narrow band imaging, tumor segmentation, HNSCC

Table of contents

Contents

Upper Aero Digestive Tract Cancer Diagnosis using Deep Learning Methods i

Declaration..... ii

Acknowledgements iii

Table of contents vi

List of figures.....x

List of tables xv

Common Abbreviations xvii

Chapter 1..... 1

Introduction 1

1.1 Background 1

1.2 UADT Cancer Delineation 3

1.2.1 Limitation of traditional examination techniques 4

1.2.2 Baseline solutions and limitations..... 4

1.2.3 Proposed solution 5

1.3 Laryngeal Cancer Detection 6

1.3.2 Baseline solutions and limitations..... 6

1.3.3 Proposed solution 7

1.4 Laryngeal Cancer Delineation..... 8

1.4.2 Baseline solutions and limitations..... 9

1.4.3 Proposed solution 10

1.5 Contributions..... 11

1.6 Publications 11

1.6.1 Peer-reviewed journals 11

1.6.2 Conferences..... 12

1.7 Thesis organization..... 13

Chapter 2 14
Literature Review 14
2.1 Overview 14
2.2 UADT Articles Search Methods 14
2.3 UADT Classifications 15
2.4 UADT Detection and Delineation 19
2.5 UADT Delineation 20
2.6 Summary 23
Chapter 3 24
SegMENT: DL model for UADT cancer Delineation 24
3.1 Overview 24
3.2 Materials and methods 25
3.2.1 Data acquisition 25
3.2.2 *SegMENT* architecture 26
3.2.3 Baseline models 28
3.2.4 Ensemble technique 29
3.3 Experimental Setup 30
3.3.1 Training parameters 30
3.3.2 Validation on LSCC dataset 31
3.3.3 Validation on OCSCC and OPSCC datasets 31
3.3.4 Outcome analysis 31
3.3.5 Statistical analysis 33
3.4 Results 34
3.4.1 Datasets 34
3.4.2 Experimental Results 35
3.4.3 Statistical Results 38
3.5 Summary 39
Chapter 4 40
Laryngeal cancer detection with deep learning 40

4.1 Overview 40
4.2 Materials and methods 40
4.3 Results 46
4.3.1 Dataset 46
4.4 Summary 50
Chapter 5 51
SegMENT-Plus: DL model for LSCC cancer Delineation 51
5.1 Overview 51
5.2 Materials and methods 52
5.3 Results 61
5.4 Summary 70
Chapter 6 71
Towards Realtime LSCC Delineation 71
6.1 Overview 71
6.2 Materials and methods 71
6.3 Experimental Setup 76
6.4 Results 81
6.5 Summary 83
Chapter 7 84
Conclusions and Future Directions 84
7.1 Overview 84
7.2 UADT cancer delineation 84
7.2.1 Limitation and future direction 87
7.3 Automatic LSCC detection 87
7.3.1 Limitation and future direction 89
7.4 Precise LSCC delineation 89
7.4.1 Limitation and future direction 91
7.5 Automatic real-time LSCC delineation 91
7.5.1 Limitation and future direction 93

7.6 Conclusion.. 93
References 95

List of figures

Figure 1. 1 Development of a computer vision model. After training the model is evaluated on previously unseen images. If this does not fulfill the requirements, it can be retrained with more images or better set up. 2

Figure 1. 2 Computer vision tasks applied to upper aerodigestive tract endoscopy and laryngoscopy: (A) classification; (B) localization; (C) object detection; (D) semantic segmentation; and (E) instance segmentation. 3

Figure 1. 3 The anatomy of UADT with example of SCC cancer on three sites (oral cavity, oropharynx, and larynx). The light red mask indicates the delineation of SCC cancer regions. The LSCC images are taken from internal institution while OSCC and OPSCC taken from external health institutions. 5

Figure 1. 4 The anatomy of Larynx with example of SCC cancer with WL and NBI imaging techniques. The light bounding box indicates the detection of SCC cancer regions. The LSCC images taken at in-office are also known as pre-operative and during intraoperative procedure. 8

Figure 1. 5 The anatomy of Larynx with example of SCC cancer on three datasets population (Genova, Brescia, and Seoul, South Korea). The light red contour indicates the delineation of SCC cancer regions. The Genova has a large dataset and is internal cohort LSCC images while the other two datasets are external health institutions..... 10

Figure 2. 1 Flow diagram depicting the review's article selection process. The search terms are displayed. Figure (a) depicts the process of identifying and removing duplicate articles based on article titles and publisher information, whereas Figure (b) depicts the workflow used to search for articles using all keywords. 16

Figure 3. 1 Flow chart of data acquisition for the laryngeal squamous cell carcinoma (LSCC) dataset. WL, white light; NBI, narrow band imaging. 26

Figure 3. 2 Workflow diagram of the proposed SegMENT semantic segmentation architecture. The Xception backbone is customized for segmentation tasks by eliminating the dense layers for classification while maintaining the feature extraction layers. The model encoder uses two ASPP blocks designed, respectively, for high- and low-level image features. The decoder section concatenates the information coming from the two ASPP blocks, producing segmentation masks. 28

Figure 3. 3 Diagram describing the training and testing of the models assessed in this work. Initially, the models are initialized with weights obtained from training on the ImageNet dataset. Next, the models are trained on the specific datasets of interest. During testing, the trained models provide segmentation predictions, and ensemble used to generate the final segmentations. In addition, in-domain transfer learning (TL) can be used to enhance the segmentation performance on small datasets using trained weights from other anatomical subsites. LSCC: laryngeal squamous cell carcinoma. OCSCC: oral cavity squamous cell carcinoma. OPSCC: oropharyngeal squamous cell carcinoma. 30

Figure 3. 4 Graphical representation of the intersection over union (IoU) calculation on a white light right glottic cancer intraoperative videoframe. The boundary traced in light blue represents the ground truth segmentation provided by an expert clinician, while the red one represents the model prediction. The IoU is calculated by dividing the overlapping area (containing the true positive pixels) by the total area of union (encompassing the false negative, true positive, and false positive pixels)..... 33

Figure 3. 5 Overview of final configuration of datasets for experiments DL models. Left: Pie chart represents the LSCC dataset split before training SegMENT and other DL models, Middle: Pie chart indicates the split ratio of OCSCC dataset, while Right: Pie chart is the overview of OPSCC dataset division. The below bar plot figure indicates the detail distribution of large LSCC dataset. 34

Figure 3. 6 Boxplots of Intersection over Union (IoU) results from SegMENT ensemble and the other state-of-the-art segmentation models on the laryngeal squamous cell carcinoma (LSCC) dataset. The cross sign represents the mean value while the horizontal line inside the boxplot shows the median value. 35

Figure 3. 7 Boxplots of Dice similarity coefficient (DSC) result from SegMENT ensemble and the other state-of-the-art segmentation models on the laryngeal squamous cell carcinoma (LSCC) dataset. The cross sign represents the mean value, while the horizontal line inside the boxplot shows the median value. 36

Figure 3. 8 Examples of automatic segmentation results for the laryngeal squamous cell carcinoma dataset using SegMENT ensemble. DSC: dice similarity coefficient; IoU: intersection over union; WL: white light; NBI: Narrow Band Imaging. 38

Figure 3. 9 Examples of automatic segmentation results for the oral cavity and oropharyngeal squamous cell carcinoma datasets using SegMENT + ensemble in-domain transfer learning. DSC: Dice similarity coefficient; IoU: intersection over union; NBI: Narrow Band Imaging. 38

Figure 4. 1 Laryngeal cancer dataset sample images. The upper row contains the original Narrow Band Imaging (NBI) and white light (WL) laryngoscopy images, while in the lower row are reported the same frames after applying the Contrast Limited Adaptive Histogram

Equalization (CLAHE). Figure (a) represents an in-office WL view of infrahyoid epiglottic cancer. Figure (b) represents an in-office NBI videoframe of a left vocal fold cancer extending to the anterior commissure. Figure (c) represents an intraoperative WL videoframe of a right vocal fold cancer. While figure (d) represents an intraoperative NBI view of a left vocal fold cancer extending to the bottom of the ventricle. 42

Figure 4. 2 YOLOv5 architecture included backbone for features extraction. 42

Figure 4. 3 Graphical representation of the Intersection over Union (IoU) calculation on a Narrow Band Imaging (NBI) videoframe. The light blue rectangle represents the ground truth bounding box, while the red rectangle represents the model prediction. The IoU is calculated by dividing the overlap area for the total area of union. 46

Figure 4. 4 YOLO models performance metrics graphs on training and validation data. Figure A, B, and C respectively represent Recall, Precision, and mAP@0.5 curves while trained up to 100 epochs. 48

Figure 4. 5 Examples of automatic laryngeal cancer prediction provided by the ensemble model (YOLOv5s with YOLOv5m – TTA). The first two columns on the left contain images with ground truth bounding boxes, while the two columns on the right contain the same images with YOLO-predicted bounding boxes. Case A represents a carcinoma of the infrahyoid epiglottis; case B represents a carcinoma of the infrahyoid epiglottis; case C represents a carcinoma of the left vocal fold; case D represents a carcinoma of the right vocal fold involving the anterior commissure. 49

Figure 4. 6 Panel of testing videoframes extracted from 6 video laryngoscopies. Every row represents a different video: the first 4 pictures of every row are extracted from the original video, while the last 4 images are the same frames extracted after the prediction of the ensemble model (YOLOv5s with YOLOv5m – TTA). The first video (V1) reports a carcinoma of the left vocal fold; the second video (V2) reports a cancer of the right vocal fold; images extracted from the third video (V3) depict a carcinoma affecting the laryngeal surface of the suprahyoid epiglottis and a severe dysplasia of the right vocal fold. Lastly, V4 reports a carcinoma of the right vocal fold extending to the anterior commissure, V5 shows a tumor of the left vocal fold, while V6 reports frames extracted from a healthy larynx. .. 50

Figure 5. 1 SegMENT-Plus block diagram with integrated components. (a) SegMENT-Plus with skip connections (in dashed arrows) within the encoder-decoder architecture. (b) The m-ASPP module consists of five 3x3 convolutions with different dilated rates (1, 3, 6, 9 and 12), followed SE block and concatenated with the CBAM. 53

Figure 5. 2 CBAM architecture block diagram. The 'GAP' represents global averaging pooling while GMP denotes global max pooling. 55

Figure 5. 3 Comparison of various Atrous Spatial pyramid pooling modules. (a) ASPP (b) Dense ASPP, (c) WASPP and (d) Proposed m-ASPP. 57

Figure 5. 4 The loss convergence plot of each segmentation model, figure (a) shows the training convergence plot while figure (b) shows the validation convergence plots. 62

Figure 5. 5 The m-ASPP grad-Cam visualization sample images from Genova LSCC dataset. The grad-cam in third column shows m-ASPP module without SE and CBAM and only used GAP instead of CBAM. The fourth column indicates the grad-cam of m-ASPP with five parallel SE attention blocks. The fifth column indicates the grad-cam of m-ASPP module with SE and CBAM..... 64

Figure 5. 6 Boxplots showing the DSC and IoU results of different segmentation models obtained on the Genoa LSCC dataset..... 65

Figure 5. 7 Boxplots showing the DSC and IoU results of different segmentation models obtained on the Seoul, South LSCC dataset..... 65

Figure 5. 8 Boxplots showing the DSC and IoU results of different segmentation models obtained on the Brescia LSCC dataset. 66

Figure 5. 9 Sample LSCC segmentation results obtained with different segmentation models. The red contour in the images represent the ground truth, while the green contours are the segmentation results. From top to down rows: First three rows are from the Genoa LSCC test dataset; the following two rows are from the Brescia LSCC testing set; the last two rows are from the Seoul LSCC testing dataset. Images in column (a1) are the original images, while column (a2) shows the ground truth images. The images in columns (a3) – (a7) show the segmentations produced by the baseline segmentation models. The images in column (a8) are from SegMENT-Plus and in (a9) from SegMENT-Plus + TTA..... 66

Figure 5. 10 Kernel density estimation (KDE) curves on the three test datasets (Genoa, Seoul, and Brescia LSCC) for the baselined model with SegMENT-Plus and SegMENT_Plus + TTA. Figures (a1-a3) the KDE curves of the Dice Similarity Coefficient (DSC) metric, while (b1-b2) presents the KDE curves of the Intersection over Union (IoU) metric..... 69

Figure 6. 1 Workflow diagram of the study. IoU = Intersection over Union; DSC = dice similarity coefficient; FPS = frames per second. 73

Figure 6. 2 The figure illustrates the encoder-decoder with m-ASPP module in the middle architecture used in SegMENT-Plus. 74

Figure 6. 3 Testing video frames extracted from five video laryngoscopes. 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..... 76

Figure 6. 4 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. 76

Figure 6. 5 Examples of *SegMENT-Plus* automatic segmentation of laryngeal carcinoma. The panel on the left reports four examples from the University of Brescia dataset. Frames #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. 77

Figure 6. 6 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*. 78

Figure 6. 7 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. 79

Figure 6. 8 Graphical representation of the intersection over union (IoU) The boundary traced in light blue represents the ground truth segmentation, while the red one represents the model prediction. The IoU is calculated by dividing the overlapping area (containing the true positive pixels) by the total area of union (encompassing the false negative, true positive, and false positive pixels). 80

List of tables

Table 2. 1 A literature of research on the classification of UADT many subsites or the diagnosis of malignancies is compiled here, along with a description of the major authors' contributions and limitations. 17

Table 2. 2 The literature on the detection and delineation of UADT various subsites or cancer diagnosis has been summarized, along with key author contributions and limitations. 20

Table 3. 1 Performance evaluation of different semantic segmentation models during testing on the laryngeal squamous cell carcinoma (LSCC) dataset. The results represent the median scores from all the tests for each metric. Values in bold denote the best results. IoU: intersection over union. DSC: dice similarity coefficient. 33

Table 3. 2 Performance evaluation of models during testing on the oral cavity (OCSCC) and oropharynx squamous cell carcinoma (OPSCC) datasets. The results achieved by SegMENT trained only on the specific datasets and by the SegMENT ensemble model (i.e., incorporating features learned from the LSCC dataset) are compared to those reported employing UNet and ResNet in the previous study (Paderno, Piazza, et al., 2021a). The results represent the median scores from all the tests for each metric. Values in bold denote the best results. IoU: intersection over Union. DSC: dice similarity coefficient. TL: transfer learning. 39

Table 4. 1 Performance Evaluation of Various YOLO Models During Training-Validation Phase. mAP@.5 = mean Average Precision with an Intersection over Union threshold of 0.5; TP% = rate of true positive predicted bounding boxes among the total number of bounding boxes predicted; FP% = rate of false-positive predicted bounding boxes among the total number of bounding boxes predicted; TTA = test time augmentation. 43

Table 4. 2 Performance Evaluation of Various YOLO Models on Testing Phase. mAP@.5 = mean Average Precision with an Intersection over Union threshold of 0.5; TP% = rate of true positive predicted bounding boxes among the total number of bounding boxes predicted; FP% = rate of false-positive predicted bounding boxes among the total number of bounding boxes predicted; TTA = test time augmentation 44

Table 4. 3 Characteristics and Computation Times of the Testing Videos After Applying the Ensemble Model (YOLOv5s with YOLOv5m—TTA) for LSCC Detection. WHERE fps = frame per second; LSCC = laryngeal squamous cell carcinoma; Mb = megabytes; TTA = test time augmentation. 47

Table 5. 1 Genoa LSCC dataset distribution of images in subsets with type of examination and imaging technology used for training, validation, and testing. 58

Table 5. 2 SegMENT-Plus performance on genoa LSCC test dataset with different backbone models* 63

Table 5. 3 *SegMENT-Plus* performance with various ASPP modules on genoa LSCC testing dataset* 64

Table 5. 4 Genova LSCC median with (IQR) values as representative of test image findings. 67

Table 5. 5 Seoul LSCC median with (IQR) values statistical results on external test images. 68

Table 5. 6 Brescia LSCC median with (IQR) statistical results on external test images. ... 68

Table 6. 1 Composition of the University of Genova LC image dataset. The subdivision of frames in training-validation and test datasets is reported. The acquisition setting (in-office vs. intra-operative) and imaging modality (white light = WL vs. Narrow Band Imaging = NBI) are described. 81

Table 6. 2 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). IoU = Intersection over Union; DSC = dice similarity coefficient; FPS = frames per second. 82

Table 6. 3 Testing videos characteristics and computation time expressed as median (Q1-Q3) after applying SegMENT-Plus. MB = megabyte; s = seconds. 83

Common Abbreviations

AI	Artificial Intelligence
ANOVA	Analysis of Variance
ASPP	Atrous Spatial Pyramid Pooling
AG	Attention Gate
AUC	Area Under the Curve
CAD	Computer Aided Detection
CBAM	Channel Block Attention Module
CNN	Convolutional Neural Network
CVAT	Computer Vision Annotation Tool
CLAHE	Contrast Limited Adaptive Histogram Equalization
DL	Deep Learning
DSC	Dice Similarity Coefficient
EEF	Endoscopy Enhancing Filter
FCNN	Fully Convolutional Neural Network
FPS	Frames Per Second
FP	False Positive
FN	False Negative
FPS	Frame Per Second
FoV	Fields of View
GAP	Global Average Pooling
HD	High Definition
HSD	Honestly Significant Differences
IoU	Intersection over Union
KDE	Kernal Density Estimation

LSCC	Laryngeal Squamous Cell Carcinoma
ML	Machine Learning
m-ASPP	Modified Atrous Spatial Pyramid Pooling
mAP	Mean Average Precision
NBI	Narrow Band Imaging
OPSCC	Oropharyngeal Squamous Cell Carcinoma
OCSCC	Oral Cavity Squamous Cell Carcinoma
ROC	Receiver Operating Characteristic
SPIES	Storz Professional Image Enhancement System
SE	Squeeze Excitation
SD	Standard Deviation
SegMENT	Segmentation Model for Ear Nose Throat
TTA	Test Time Augmentations
TL	Transfer Learning
TP	True Positive
TN	True Negative
UADT	Upper Aero Digestive Tract
VL	Video Laryngoscopy
VIA	VGG Image Annotator
WASPP	Water Atrous Spatial Pyramid Pooling
WL	White Light
YOLO	You Only Look Once

Chapter 1

Introduction

1.1 Background

Artificial intelligence (AI) is a field of computer science that attempts to develop automatic ways for solving issues that normally require human intelligence. Computer vision (CV) is a field of study that uses AI methodologies to enable computers to derive meaningful information from digital images, videos, and other visual inputs and then conduct actions or make ideas based on that information. Machine learning (ML) is the most popular AI technology used in the medical industry. Traditional ML models are distinguished by the fact that they require expert guidance to determine which attributes are relevant for data interpretation. If a researcher wants an ML model to recognize a cancer tumor in an image, he must extract and organize a set of handcrafted features that better define texture, size, form, and color. The model will then compute the probability of a cancer tumor being in that image based on these data. This, in addition to being time demanding, restricts the model's potential because complex features may be impossible to capture using human instructions. Recent advances in processing power have enabled tremendous development in deep learning (DL), overcoming the limits of classical ML techniques. DL models are made up of complicated multilayer artificial neural networks (ANNs), the most common of which are convoluted neural networks (CNNs) in image processing (Chai et al., 2021). CNNs do not need to be instructed what attributes define an object; they can learn how to recognize it by features and anticipate the target object automatically. The DL are currently being used in the medical field with impressive outcomes (Rajkomar et al., 2019). The DL has been used specifically to endoscopy, with commercially available in market gastrointestinal solutions that assist clinicians in spotting challenging lesions during the inspection (Van Der Sommen et al., 2020). DL has recently been investigated in the field of otorhinolaryngology. More and more findings are being published in this field, particularly for upper aero digestive tract (UADT) endoscopy and laryngoscopy, encouraging researchers to designate an emerging field of research known as "videomics." Several DL algorithms are used systematically in videomics to organize, process, and model the unstructured data derived from UADT endoscopic recordings (Paderno, Holsinger, et al., 2021). Cancers of the UADT, including the mouth, tonsils, oropharynx, hypopharynx, nasopharynx, nasal cavity, and larynx, were caused primarily by tobacco and alcohol use, as well as HPV infection (Fritz & Rajasekaran,

2023). DL algorithms can eventually be trained to detect images without the assistance of a person with a large enough training dataset. These models learn to identify complex images like UADT mucosal lesions and potentially outperform a human otolaryngologist (Fiz et al., 2017a). The flowchart of a DL model for image analysis is shown in Figure 1.1. The original image dataset is divided into two parts: a training validation set for learning features of objects and a test set for assessing the model's abilities. The performance of DL models improves as the number of training images increases. So far, the best method to train DL models to recognize medical images such as UADT lesions has been to ask professionals to label and annotate the images so that the algorithm may learn how to recreate the original annotation. This is referred to as supervised training, and it requires professional otolaryngologists to annotate the images for the purpose of transferring their ability to recognize lesions.

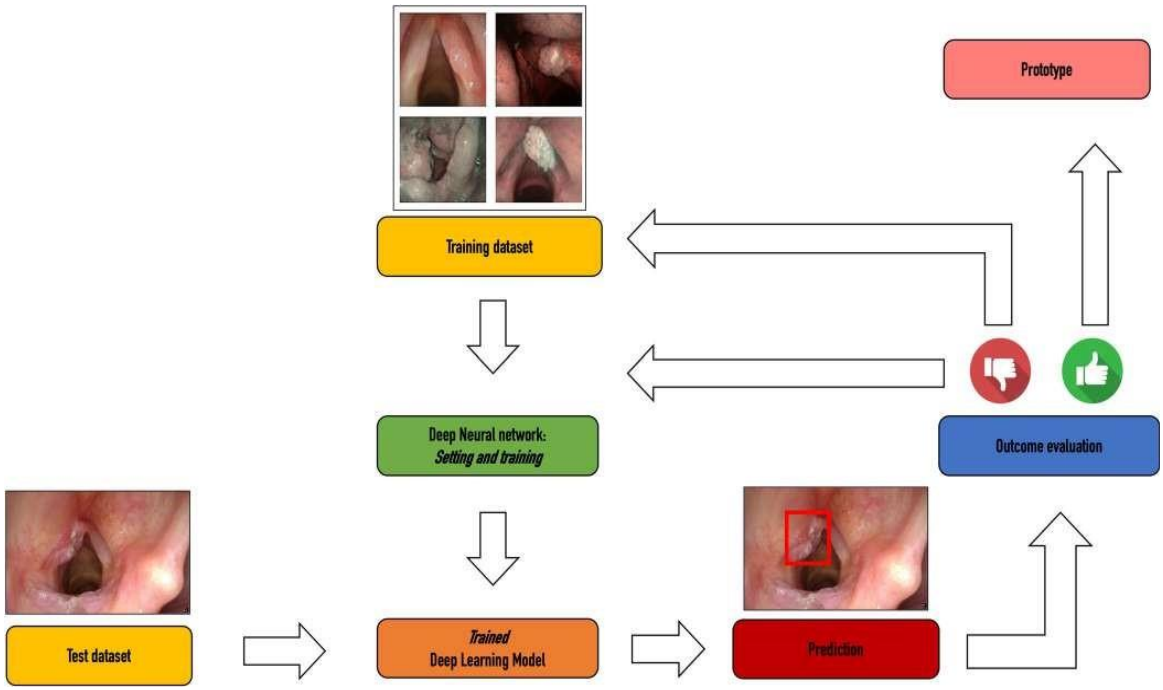


Figure 1. 1 Development of a computer vision model. After training the model is evaluated on previously unseen images. If this does not fulfill the requirements, it can be retrained with more images or better set up.

Task-specific metrics are used to evaluate performance on a testing set. If the results are unsatisfactory, the process can be repeated by modifying the method, choosing a different model, or expanding the training dataset. Notably, whereas DL models may perform well within a narrow training cohort, they suffer with varied populations. External validation datasets from various institutions are critical for testing generalization ability, considering differences in acquisition processes and population demographics. Figure 1.2 depicts a DL

model used in UADT to accomplish various tasks such as classification, detection, semantic segmentation, and instance segmentation.

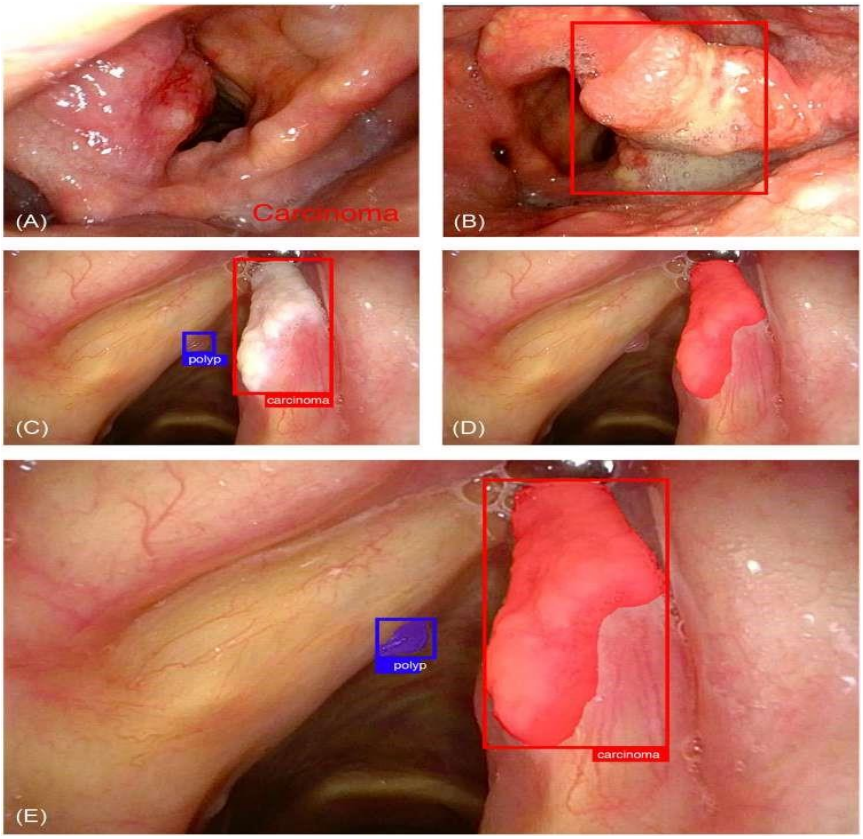


Figure 1. 2 Computer vision tasks applied to upper aerodigestive tract endoscopy and laryngoscopy: (A) classification; (B) localization; (C) object detection; (D) semantic segmentation; and (E) instance segmentation.

Classification comprises labeling images, whereas detection determines the location and class of objects and provides bounding box coordinates. Semantic segmentation assigns labels to pixels to distinguish item categories and delineate their contours using pixel-by-pixel masks. In contrast, instance segmentation extends beyond object detection by producing segmentation masks for each detected object, allowing for a more in-depth understanding of individual instances in an image.

1.2 UADT Cancer Delineation

Europe has seen a rise in the death rate from UADT malignancies (oral cavity, oropharynx, and larynx) (La Vecchia et al., 2004), with some of the biggest increases in incidence rates occurring in central and eastern nations. Endoscopy is used to perform the UADT cancer examination at an otolaryngological clinical evaluation. Endoscopy, whether performed through the nose with flexible instrumentation or transoral with rigid telescopes, provides a

detailed, magnified, and fully enhanced image of the UADT, especially when combined with high-definition (HD) technology. Endoscopy enhancing filters (EEFs), such as Narrow Band Imaging (NBI) or the Storz Professional Image Enhancement System (SPIES), served an important role in the last decade, improving conventional white light (WL) endoscopy by highlighting the submucosal and subepithelial neoangiogenic network associated with malignant transformation (Piazza et al., 2011). Endoscopy with enhanced filters outperforms standard WL endoscopy in the diagnosis of UADT carcinomas by improved imaging of cancer-related abnormal intrapapillary capillary loops (Carta et al., 2016; Piazza, Cocco, De Benedetto, et al., 2010a; Vilaseca et al., 2017a). Nowadays, EEFs like NBI are widely used in various head and neck subsites such as the larynx/ hypopharynx (Arens et al., 2016; Ni et al., 2011) oropharynx (Filauro et al., 2018; Piazza, Cocco, Del Bon, et al., 2010), nasopharynx (Madana et al., 2015; Vlantis et al., 2016), and oral cavity (Deganello et al., 2021; Vlantis et al., 2019),where they play a fundamental role in detection, classification, and delineation of superficial margins of malignant lesions.

1.2.1 Limitation of traditional examination techniques

The analysis and interpretation of UADT video endoscopies might be difficult, especially in less experienced infrastructure. Even using EEFs, the detection and evaluation of vascular anomalies is limited due to the significant heterogeneity in the presentation of squamous cell carcinomas (SCCs) in this part of the body. Furthermore, delineating boundaries might be difficult when mucosal vascularization is affected by other factors such as inflammatory disease or recent (chemo) radiation (W. H. Wang et al., 2012). Finally, several aspects hinder the large-scale implementation of EEFs during routine UADT endoscopic assessment, such as its intrinsic operator-dependent nature and the relatively steep learning curve needed to master this technique.

1.2.2 Baseline solutions and limitations

The advancement in DL technique overcomes the diagnostic challenges of the UADT oncologic diseases (Paderno, Holsinger, et al., 2021). The DL methods are systematically used to process the unstructured video data obtained from diagnostic endoscopy to convert subjective assessment into objective findings. Parallelly, the use of DL in video endoscopy, especially in the gastrointestinal field, has already become relevant in the literature and even on the market (Kudo et al., 2021). When moving to the specific field of UADT, however, only a few studies have been published in the current literature, with most focusing on laryngeal endoscopy (Yao et al., 2022). Among all AI-powered methods, DL techniques based on CNNs are increasingly used in UADT video endoscopy analysis for automatic disease detection (Tamashiro et al., 2020), classification (Cho et al., 2021), and segmentation (Fehling et al., 2020). There is only a limited dataset related to UADT that is publicly

available. In segmentation task there is still limited work that focuses on different sites of UADT. According to our knowledge, there is no public dataset available with ground truth tumor contour related to Oral Cavity Squamous Cell Carcinoma (OCSCC), oropharyngeal Squamous Cell Carcinoma (OPSCC) and Laryngeal Squamous Cell Carcinoma (LSCC).

1.2.3 Proposed solution

In this thesis, to overcome the public dataset limitation, a dataset of WL and NBI video frames of LSCC was collected and manually annotated by expert otolaryngologist. 219 LSCC patients were retrospectively included in the study. A total of 683 videoframes composed the LSCC dataset, while the external validation cohorts collected for OPSCC and OCSCC contained 116 and 102 images, respectively. As seen in Figure 1.3, the general anatomy of UADT with SCC cancer varies across three different locations. A new CNN-based semantic segmentation model for video endoscopy of the UADT, named "*SegMENT"* developed. This model was specifically developed for the identification and segmentation of UADT cancer in endoscopic video frames, with particular attention to LSCC, OPSCC and OCSCC. The impact of in-domain transfer learning through an ensemble technique was also evaluated on the small external datasets to improve delineation. The detailed description of dataset and methodology used to develop "*SegMENT*" model and in-domain transfer learning approach described in chapter 3.

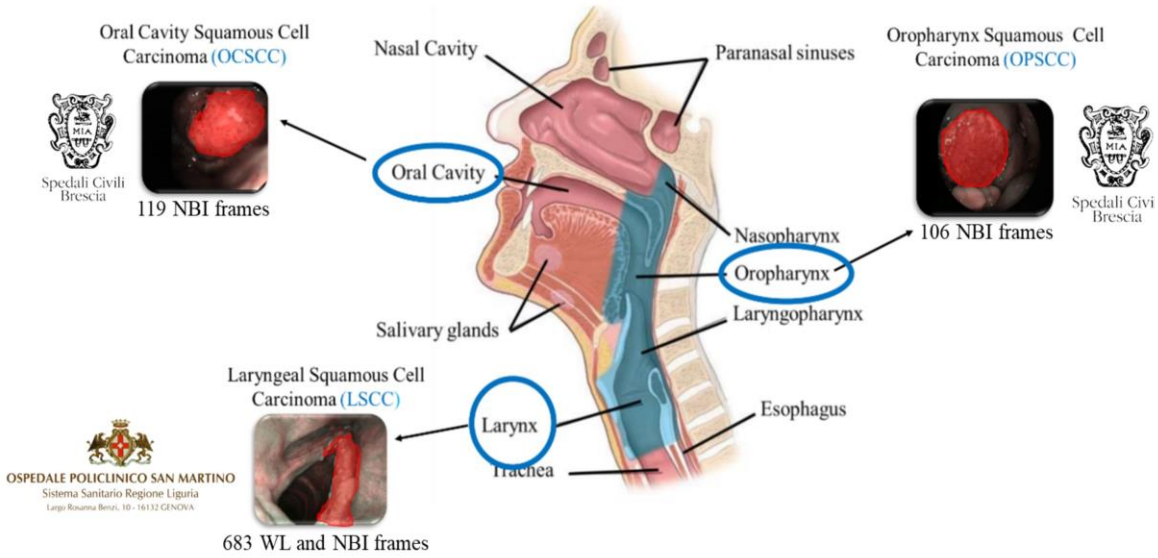


Figure 1. 3 The anatomy of UADT with example of SCC cancer on three sites (oral cavity, oropharynx, and larynx). The light red mask indicates the delineation of SCC cancer regions. The LSCC images are taken from internal institution while OSCC and OPSCC taken from external health institutions.

1.3 Laryngeal Cancer Detection

Laryngeal squamous cell carcinoma (LSCC) is an epithelial cancer arising from the respiratory mucosal lining of the larynx. LSCC is the second most prevalent type of head and neck cancer, accounting for approximately one-fifth of all cancers among UADT (Q. Zhang et al., 2021). It primarily affects men and has a greater incidence and mortality rate. According to Global Cancer Observatory projections, there will be 94,000 deaths and 177,000 new cases worldwide in 2018 (Gamez et al., 2020). The most common type, accounting for around 95% of occurrences, is squamous cell carcinoma (SCC), which is mostly associated with smoking (Bray et al., 2018). Early detection of LSCC is mandatory to increase survival rates and reduce the morbidity caused by treatments that are needed for advanced-stage disease. Identification of LSCC usually starts in the otolaryngologist's office by flexible trans nasal fiberoptic endoscopy. High definition (HD) video laryngoscopy has recently replaced standard fiberoptic endoscopy, as it offers better accuracy in detection and videorecording of laryngeal lesions (Scholman et al., 2020). The WL endoscopic appearance of LSCCs may be nonspecific, and preoperative clinical assessment is not always in agreement with the final histopathologic diagnosis (Odell et al., 2021). The NBI optical technique improves diagnosis of malignant lesions by enhancing submucosal neoangiogenic changes, designing thick dark spots that can be observed within and surrounding the malignant lesion itself (Arens et al., 2016) . Consequently, the use of HD video laryngoscopy in combination with NBI allows more effective and earlier detection of LSCC (Vilaseca et al., 2017a).

1.3.1 Limitation of traditional examination techniques

The current exploitation of this technology in less experienced centers is burdened by its operator dependent nature, being influenced by the frequency of its use and related expertise in endoscopy. Surgeons often encounter a challenge during the short pre-operative evaluation since the surgery is performed on an awake patient, making it difficult to clearly visualize the laryngeal site. At this moment, there is no automatic AI help for surgeons in typical LSCC detection systems, increasing the possibility of overlooking early-stage carcinoma during the examination (Esmaeili et al., 2020b). Moreover, traditional techniques suffer from a relatively long learning curve (Irjala et al., 2011) and are hampered by intrinsic limitations such as subjectivity in interpretation, attention, and visual inspection capabilities.

1.3.2 Baseline solutions and limitations

The DL techniques have been demonstrated to be particularly suitable for automatic laryngeal cancer detection and interpretation (Chai et al., 2021; Tama et al., 2020; W. H. Wang et al., 2012). The impact of DL laryngeal detection when coupled with NBI improves

LSCC screenings (Inaba et al., 2020). However, most studies in literature are related to static frame detection and do not include the possibility of applying DL for real-time detection of LSCC on WL and NBI video laryngoscopies. In real-time video detection the DL model encoders the challenges related to moving endoscope, blurriness, and uneven illumination of videos frames. The computation time constraints are also an important challenge for real time LSCC detection. There is no publicly available dataset available with ground truth related to detection work.

1.3.3 Proposed solution

Recorded videos of LSCC were retrospectively collected from in-office trans nasal video endoscopies and intraoperative rigid endoscopies. Two hundred and nineteen patients were retrospectively enrolled. A total of 624 LSCC videoframes were extracted using WL and NBI imaging techniques. A few samples imaging of WL/NBI from in-office and intraoperative examination seen in Figure 1.4. To assess a new DL application for real-time LSCC detection in both WL and NBI video laryngoscopies, the You-Only-Look-Once (YOLO) CNN model implemented. The mosaic augmentation applied to enhance the number of images during training. The proposed ensemble algorithm (YOLOv5s with YOLOv5m—TTA) achieved the best LSCC detection results on static frames while YOLOv5m was used to assess the automatic detection of LSCC in six video laryngoscopies. This study identified a suitable CNN model for LSCC detection in WL and NBI video laryngoscopies. Detection performances are highly promising. The limited complexity and quick computational times for LSCC detection make this model ideal for real-time processing. The summary of detection model and data collection described in chapter 4.

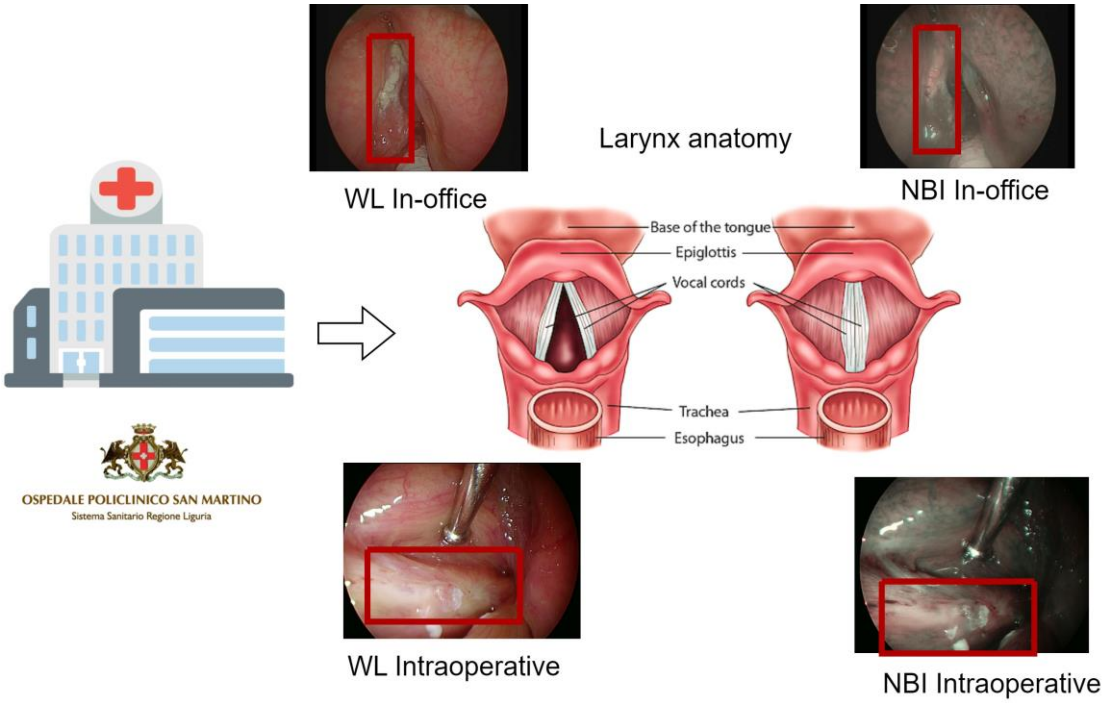


Figure 1. 4 The anatomy of Larynx with example of SCC cancer with WL and NBI imaging techniques. The light bounding box indicates the detection of SCC cancer regions. The LSCC images taken at in-office are also known as pre-operative and during intraoperative procedure.

1.4 Laryngeal Cancer Delineation

Segmentation is the process of locating and contouring the anatomical features of laryngeal malignancies. The goal of early cancer delineation is to prevent the disease from spreading to later stages. When a tumor is under staged, there may be cancerous tissue that remains after treatment. On the other hand, an over staged tumor may cause the loss of healthy tissue, which would limit laryngeal function instead of protecting the organ (B. Li et al., 2011). The traditional method for examining LSCC involves a flexible endoscope for real-time visual inspection during pre-examination and intraoperative surgery. The use of NBI enhances LSCC detection by emphasizing the laryngeal submucosal vasculature (C et al., 2010).

1.4.1 Limitation of traditional examination techniques

A significant challenge arises due to the considerable variation in the endoscopic appearance of LSCC among different patients during examination. Nonetheless, the efficacy of NBI is limited outside expert clinical centers due to its reliance on operator expertise. As a result, computer-assisted endoscopic image processing could potentially offer a solution to this challenge (Hamad et al., n.d.; Ji et al., 2020a; Pan et al., 2022).

1.4.2 Baseline solutions and limitations

DL-based segmentation models have also been used for delineating regions of interest in endoscopic laryngeal images, such as the vocal folds, the epiglottis, and the glottic space. This includes models with the potential for assessing the respiratory space and vocal fold angles (Ding et al., 2022; Hamad et al., n.d.). In addition, some authors explored CNNs for segmenting laryngeal lesions (Ji et al., 2020). The combination of NBI with automatic cancer boundaries assessment is of particular interest in UADT carcinoma treatment. Numerous authors have reported that NBI allows for a better definition of excisional margins with consequently improved resections of head and neck mucosal cancer (Farah et al., 2016; Garofolo et al., 2015). However, even with the advent of NBI, the rate of positive surgical margins still represents an open issue(Fiz et al., 2017a; Gorphe & Simon, 2019a; Sampieri et al., 2022a). AI-based automatic segmentation is the one reported by Adamian et al (Adamian et al., 2021) who implemented a CNN for vocal fold tracking in video endoscopies. The model could estimate the anterior glottic angle by tracking the vocal folds' free borders, which may allow us to automatically identify patients with vocal fold movement disorders. In the literature, there is only one publicly available dataset composed of two patients sequence used in (Laves et al., 2019) study. The dataset's shortcoming is that it has a smaller number of patients with the lowest variety of images. The public dataset is about laryngeal subsite segmentation rather than disease. A further limitation in the literature is that only a few studies focus on external validation but with different imaging modalities and other UADT sites (Shephard et al., 2023; Ye et al., n.d.). External validation is important for demonstrating the generalization ability of the model to predict LSCC. In study (Y. Li et al., 2023), the external validation performs specific to LSCC with WL and NBI imaging. This study implemented segmentation task prior to classification and did not include major evaluation metrics of segmentation task since the final output task focuses on classification.

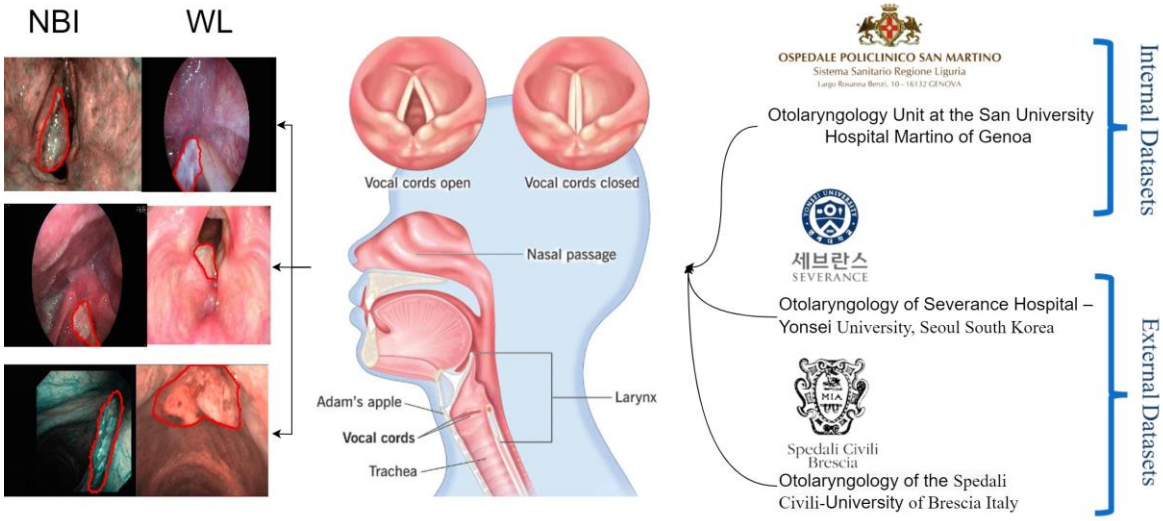


Figure 1. 5 The anatomy of Larynx with example of SCC cancer on three datasets population (Genova, Brescia, and Seoul, South Korea). The light red contour indicates the delineation of SCC cancer regions. The Genova has a large dataset and is internal cohort LSCC images while the other two datasets are external health institutions.

1.4.3 Proposed solution

A total of 4289 annotated laryngeal images from 766 patients were included in this study. The multi-center dataset from three different hospitals was included in this study also shown in Figure 1.5. SegMENT-*Plus*, a DL segmentation specifically developed and optimized for accurate delineation of LSCC. *SegMENT-Plus* uses *EfficienstNetB5* as encoder with a new modified Atrous Spatial Pyramid Pooling (*m-ASPP*) block that integrates channel block attention module (CBAM) and squeeze excitation (SE). In this new architecture, CBAM extracts local and global LSCC features from the encoder, while the SE block refines cancer segmentation on each dilated convolution output within *m-ASPP* module. The proposed segmentation architecture also demonstrated statistically significant improvement in dice coefficient (DSC) and intersection over union (IoU) compared to other state of the art architectures, showing that this work is a concrete foundation towards a clinical system for the automatic delineation of LSCC margins in endoscopic images. The brief description of novel *SegMENT-Plus* architecture, datasets and real-time delineation described in chapter 5 and 6.

1.5 Contributions

- Development of a Novel DL segmentation network named *SegMENT* to precisely delineate LSCC, OPSCC, and OSCC using a domain transfer learning approach **(Chapter 3).**
- DL-based detection of LSCC on pre-operative and intraoperative static frames and video laryngoscope **(Chapter 4).**
- A large laryngeal WL and NBI imaging dataset was collected and annotated at the IRCSS San Martino Hospital's Unit of Otorhinolaryngology-Head and Neck Surgery in Genoa, Italy. The dataset is now private but will become open source in the future for future research. **(Chapter 5).**
- External cohort dataset collected and annotated in Brescia, Italy and Seoul, South Korea for a multi-centric study; the dataset is confidential but will be publish for research purposes in future. **(Chapter 5 and 6).**
- Novel advanced segmentation model named *SegMENT-Plus* for enhancing LSCC delineation on a large dataset comprising WL and NBI with pre- and intra-operative frames using a multicentric study approach **(Chapter 5 and 6).**
- Investigate performance improvements of the semantic segmentation model to state of art DL methods on LSCC endoscopic images. External cohort performance validation to demonstrate model generalization ability **(Chapter 5 and 6).**
- Preliminary delineation results on real-time LSCC video laryngoscopy, which could transform clinical practice by minimizing tumor miss-detection and margin errors **(Chapter 6).**
- Comprehensive comparison of the novel DL model with senior and junior laryngologist **experts (Chapter 6).**

1.6 Publications

1.6.1 Peer-reviewed journals

1. **Muhammad Adeel Azam**, Sampieri C, Ioppi A, Azam MA, Baldini C, Li S, Moccia S, Peretti G, Mattos LS. "**Automatic delineation of laryngeal squamous cell carcinoma during endoscopy**". Biomedical Signal Processing and Control. 2024 Feb 1; 88:105666. https://doi.org/10.1016/j.bspc.2023.105666.
2. Sampieri C, Baldini C, **Muhammad Adeel Azam**, Moccia S, Mattos LS, Vilaseca I, Peretti G, Ioppi A. "**Artificial Intelligence for Upper Aerodigestive Tract Endoscopy and Laryngoscopy: A Guide for Physicians and State-of-the-Art Review**". Otolaryngology–Head and Neck Surgery. 2023 Apr 13. https://doi.org/10.1002/ohn.343.

3. **Muhammad Adeel Azam**, Sampieri C, Ioppi A, Benzi P, Giordano GG, De Vecchi M, Campagnari V, Li S, Guastini L, Paderno A, Moccia S. "**Videomics of the upper aero-digestive tract cancer: Deep learning applied to white light and Narrow Band Imaging for automatic segmentation of endoscopic images**". Frontiers in Oncology. 2022 Jun 1; 12:900451. https://doi.org/10.3389/fonc.2022.900451.
4. **Muhammad Adeel Azam**, Sampieri C, Ioppi A, Africano S, Vallin A, Mocellin D, Fragale M, Guastini L, Moccia S, Piazza C, Mattos LS. "**Deep learning applied to white light and narrow band imaging video laryngoscopy: toward real-time laryngeal cancer detection**". The Laryngoscope. 2022 Sep;132(9):1798-806. https://doi.org/10.1002/lary.29960.
5. Sampieri C, **Muhammad Adeel Azam,** Ioppi, A., Baldini, C., Moccia, S., Kim, D., Tirrito, A., Paderno, A., Piazza, C., Mattos, L.S. and Peretti, G. (2024), **"Real-Time Laryngeal Cancer Boundaries Delineation on White Light and Narrow-Band Imaging Laryngoscopy with Deep Learning".** The Laryngoscope. https://doi.org/10.1002/lary.31255.

1.6.2 Conferences

1. **Muhammad Adeel Azam**, Baldini C, Li S, Sampieri C, Ioppi A, Moccia S, Peretti G and Mattos LS, "**Automatic Segmentation of Laryngeal Cancer Using Deep Learning: A Multicentric Dataset Study**," Proceedings of the 12th Conference on New Technologies for Computer/Robot Assisted Surgery (CRAS 2023), ISBN: 9789491857089, pp. 50-51, Paris, France, September 11 – 13, 2023.
2. Baldini C, **Muhammad Adeel Azam**, Sampieri C, Ioppi A, Moccia S, Peretti G and Mattos LS, "**An Automated Approach for Informative Frame Selection in Laryngeal Cancer Screening**," Proceedings of the 12th Conference on New Technologies for Computer/Robot Assisted Surgery (CRAS 2023), ISBN: 9789491857089, pp. 60-61, Paris, France, September 11 – 13, 2023.
3. **Muhammad Adeel Azam**, Baldini C, Li S, Sampieri C, Ioppi A, Moccia S, Peretti G and Mattos LS, "**Real-Time Detection of Laryngeal Squamous Cell Carcinoma via Video Laryngoscopy with WL and NBI using YOLO**," Proceedings of the 11th Conference on New Technologies for Computer/Robot Assisted Surgery (CRAS 2022), pp. 92-93, Napoli, Italy, April 25 – 27, 2022.

1.7 Thesis organization

The thesis is organized as follows.

Chapter 2 The literature on existing techniques is reported. The chapter is divided into three sections. First, we discuss approaches for automatic UADT diagnosis and their limitations; second, we present existing traditional and DL techniques for UADT cancer detection with significant contributions and limitations of the work; and third, we provide details of the segmentation literature and key contributions.

Chapter 3 The UADT delineation is presented using the novel "*SegMENT*" semantic segmentation method. The chapter included datasets used as internal and external cohorts for experimental purposes. The chapter also detailed the in-domain transfer learning approach that was used to test the dataset of small external cohorts.

Chapter 4 reports on the LSCC identification task utilizing a DL technique, with a brief discussion of the ensemble DL (YOLO) algorithm. The chapter also covers the dataset utilized in this study as well as the capability of the DL model for detecting LSCC in real-time on WL and NBI laryngoscopy surgical preoperative and intraoperative videos.

Chapter 5 describes the development of large LSCC datasets and collection of two external datasets for automatic delineation task using DL model. This chapter also covers the new *SegMENT-Plus* semantic DL architecture in depth. The ablation study and comparison with other state of art techniques are also merged in this chapter.

Chapter 6 demonstrates an extension of the *SegMENT-Plus* model for comparison with otolaryngologists' experts, as well as preliminary quantitative and qualitative results from real-time intraoperative laryngeal videos.

Chapter 7 concludes the thesis by providing a short discussion of every chapter. It reports the problem tackled in the conducted research, highlights their limitations, and presents the possible future directions.

Chapter 2

Literature Review

2.1 Overview

This chapter presents works that are closely related to these study directions listed as follows: (1) UADT cancer classification, detection, and delineation on WL and NBI imaging using DL; (2) Previous UADT datasets and their limitations. (3) DL technique limitations and clinical challenges in the automatic diagnosis of UADT cancers.

Related Publications

1. Sampieri C, Baldini C, **Muhammad Adeel Azam**, Moccia S, Mattos LS, Vilaseca I, Peretti G, Ioppi A. "**Artificial Intelligence for Upper Aerodigestive Tract Endoscopy and Laryngoscopy: A Guide for Physicians and State-of-the-Art Review**". Otolaryngology–Head and Neck Surgery. 2023 Apr 13. https://doi.org/10.1002/ohn.343.
2. Baldini C, **Muhammad Adeel Azam**, Sampieri C, Ioppi A, Moccia S, Peretti G and Mattos LS, "**An Automated Approach for Informative Frame Selection in Laryngeal Cancer Screening**," Proceedings of the 12th Conference on New Technologies for Computer/Robot Assisted Surgery (CRAS 2023), ISBN: 9789491857089, pp. 60-61, Paris, France, September 11 – 13, 2023.

2.2 UADT Articles Search Methods

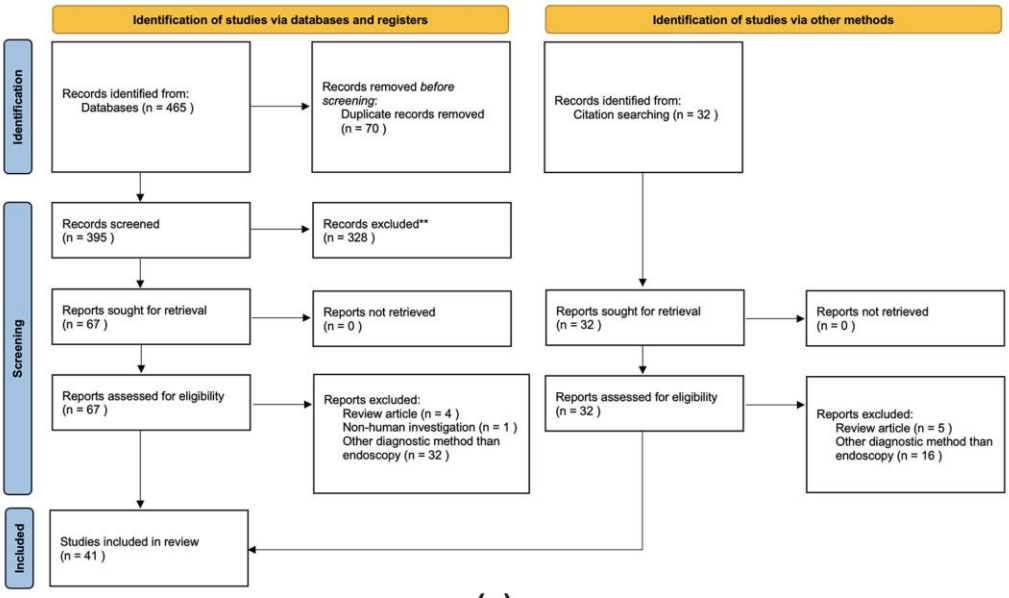
A literature was carried out to find the relevant articles identified by searching through Medline, Google Scholar, Cochrane, and Embase databases. MeSH terms—including UADT districts, AI techniques, and endoscopic examinations were used. Pertinent articles identified in the references after full-text reading were also included for comprehensiveness. Inclusion criteria: (1) application of AI-based models (machine learning [ML] and deep learning [DL]) to endoscopy of the UADT; (2) reported outcome measures of the algorithm performance. Exclusion criteria were: (1) use of other diagnostic methods rather than endoscopy; (2) nonstandard endoscopy techniques (i.e., high-speed endoscopy, contact

endoscopy, etc.); (3) review articles, systematic reviews, and letters to the editor; (4) nonhuman investigations; (5) papers published in non-peer-reviewed journals; and (6) other languages than English. The search strings and the selection steps are reported in the flowchart in Figure 2.1. The relevant information of the included articles such as date of publication, first author, AI algorithm, anatomical site of application, AI-task investigated, main features, and outcomes are detailed in Tables 2.1 - 2.2.

2.3 UADT Classifications

Regarding the classification task, AI-based algorithms can be applied to video endoscopy in different ways. The important classification task in UADT is that of predicting the histology of the various lesions encountered during the examination, the so-called “optical biopsy.”. Despite being challenging, this task has the potential to create computer-assisted diagnosis technologies able to support clinicians in deciding the optimal patient's therapeutic pathway. The first attempts were made using handcrafted feature extraction and traditional ML classifier by focusing on the tissue texture (especially from NBI-enhanced images) Araujo (Araújo et al., 2019). Nevertheless, this approach was limited by the number and the quality of the features selected by the investigators, and it was less efficient compared to modern DL models. The CNN classifiers are almost exclusively used to analyze endoscopic images and they have been widely investigated to classify OCSCC (Heo et al., 2022) nasopharyngeal carcinomas (Xu et al., 2022) on both WL and NBI images, and even laryngopharyngeal reflux lesions (Witt et al., 2014). Interestingly the studies of Zhao et al, (Q. Zhao et al., 2022) showed that DL models can obtain better results in a binary-classification network, while they lose diagnostic power in a multiclass operation. Intuitively, as happens for humans, the more complicated the prediction task, the lower the accuracy. In fact, Zhao and colleagues showed a loss of accuracy (0.80 vs 0.93) when training their model to recognize 4 classes (normal, polyp, keratinization, and carcinoma) instead of malignant/premalignant lesions from benign ones. However, in most cases, these analyses are burdened by the lack of external validation cohorts while, to date, no implementation of a videoendoscopic system was attempted in clinical practice, thus making it difficult to assess the real implications of this technology. Possibly this is also due to a diffuse distrust in this technology as it is particularly difficult to understand the decision-making process of the DL algorithms considering the impossibility to evaluate which features are used to make the lesions' classification tasks and confirm the need for very large and diverse datasets for training. That said, the sole introduction of the binary classification tool for screening which lesions are at high risk of malignancy would be already a turning point in clinical practice. In this light, Ren et al used one of the biggest datasets of laryngoscope images to develop a DL model able to reach an overall classification accuracy above 95% even prediction. This “black box” nature of DL models, apart from hampering

the model structure optimization, creates open ethical issues regarding their reliability. The overall summary from classification task summarized in Table 2.1.



(a)

Database	Search strategy
Medline	("Artificial Intelligence"[Mesh] OR "Machine Learning" [Mesh] OR "Deep Learning" [Mesh] OR "neural network*"[tw] OR "convoluted neural network*"[tw] OR "convolutional neural network*"[tw] OR "computer vision" [tw] OR "videomics*"[tw]) AND ("Larynx" [Mesh] OR "Pharynx" [Mesh] OR "oropharynx" [tw] OR" oral cavity" [tw] OR "mouth" [tw] OR "tongue" [tw] OR "nose" [tw] OR "nasal" [tw] OR "paranasal sinus*"[tw] OR "nasopharynx" [tw] OR "hypopharynx" [tw] OR "airway*"[tw] OR "laryngeal neoplasm*"[tw] OR "oral cavity neoplasm*"[tw] OR "oropharyngeal neoplasm*"[tw]) AND ("Laryngoscopy" [Mesh] OR "Endoscopy" [Mesh] OR "videoendoscop*"[tw] OR "fibroscop*"[tw] OR "video*" [tw] OR "photo*"[tw] OR "picture*" [tw] OR "frame*" [tw] OR "videolaryngoscopy"[tw] OR "videolaryngoscope*"[tw]) NOT ("pathomic*" [tw] OR "phatology" [tw] OR "genetic*"[tw] OR "protein*"[tw] OR "histology"[tw] OR "Magnetic Resonance Imaging"[Mesh] OR "Tomography, X-Ray Computed"[Mesh] OR "Positron-Emission Tomography" [Mesh] OR "radiology"[tw] OR "orthopantomogram*"[tW] OR "cone beam computed tomograph*"[tw] OR "radiomic*" [tw])
EMBASE	('Artificial Intelligence'/exp OR 'Machine Learning'/exp OR 'Deep Learning'/exp OR 'convoluted neural network':ab,ti OR 'computer vision':ab,ti) AND ('Larynx'/exp OR 'Pharynx'/exp OR 'oropharynx'/exp OR 'oral cavity'/exp OR 'nose'/exp OR 'nasal':ab,ti OR 'paranasal sinus':ab,ti OR 'nasopharynx'/exp OR 'hypopharynx'/exp OR 'airway':ab,ti OR 'laryngeal neoplasm':ab,ti OR 'oral cavity neoplasm':ab,ti OR 'oropharyngeal neoplasm':ab,ti) AND ('Laryngoscopy'/exp OR 'Endoscopy'/exp OR 'videoendoscopy':ab,ti OR 'videolaryngoscopy':ab,ti) NOT ('pathomic':ab,ti OR 'phatology':ab,ti OR 'genetic':ab,ti OR 'Magnetic Resonance Imaging':ab,ti OR 'Positron-Emission Tomography':ab,ti OR 'radiomic':ab,ti)
Cochrane library	(Artificial intelligence OR Machine learning OR Deep learning OR convoluted neural network OR computer vision) AND (Larynx OR pharynx OR oral cavity OR tongue OR Nasal OR paranasal sinuses OR hypopharynx OR airway OR laryngeal neoplasm OR oral cavity neoplasm OR oropharyngeal neoplasm) AND (Laryngoscopy OR endoscopy OR videoendoscopy OR video OR videolaryngoscopy) NOT (pathomic OR radiomic)
Google Scholar	(Artificial intelligence OR Machine learning OR Deep learning OR convolution neural network OR computer vision) AND (Larynx OR pharynx OR oral cavity OR tongue OR Nasal OR paranasal sinuses OR hypopharynx OR airway OR laryngeal neoplasm OR oral cavity neoplasm OR oropharyngeal neoplasm) AND (Laryngoscopy OR endoscopy OR videoendoscopy OR video OR videolaryngoscopy) NOT (pathomic OR radiomic OR genomic OR Magnetic Resonance Imaging OR Computed Tomography OR Positron Emission Tomography)

(b)

Figure 2. 1 Flow diagram depicting the review's article selection process. The search terms are displayed. Figure (a) depicts the process of identifying and removing duplicate articles

based on article titles and publisher information, whereas Figure (b) depicts the workflow used to search for articles using all keywords.

Table 2. 1 A literature of research on the classification of UADT many subsites or the diagnosis of malignancies is compiled here, along with a description of the major authors' contributions and limitations.

First author	Work	Anatomy site	Task	Methods	Datasets	Outcomes	Limitations
Wang (Y. Y. Wang et al., 2022)	Orthogonal region selection for vocal folds (VFs) closure	Larynx	Classification	LARNet-STC (custom model)	videos=58 laryngoscopy frames=9102 Patients=20 (private dataset)	classification in open and closed vocal cords. Rec.=0.96 Prc. = 0.90 F1 = 0.93	Instead of classifying and localizing malignancies, simply focus on vocal cord movements.
Araujo (Araújo et al., 2019)	Early-stage laryngeal SCC diagnosis using handcrafted features.	Larynx	Classification	ML vs DL features obtained with ResNet and inception + ML (SVM).	1320 Images patches,4 tissue classes (healthy, hypertrophic, leukoplakia, IPCL vessels). (public dataset)	Early classification of SCC. Rec.=0.98 Prc.= 0.98 F1 = 0.98 time = 41 s	Patches are too small; preprocessing steps are required to calculate patch area in real-time implementation.
Irem Turkmen (Irem Turkmen et al., 2015)	Classification of laryngeal disorders based on shape and vascular defects of vocal folds.	Larynx	Classification + segmentation	Handcrafted features extraction + ML.	Patients=70 (Healthy 28 Polyp 17 Nodule 20 Laryngitis 30 Sulcus vocalis 29). (private dataset)	Segmentation of the vessels and classification of laryngeal benign lesions five classes. Prec.= 0.69-0.94 Rec. = 0.73-0.94 Spec.= 0.89-0.99	This study has a limited number of images per class and no segmentation performance metrics are calculated.
Moccia (Moccia et al., 2017)	Texture-based laryngeal tissue classification for early-stage diagnosis.	Larynx	Classification	Handcrafted features extraction + ML.	NBI images = 330. 1320 Images patches,4 tissue classes (healthy, hypertrophic, leukoplakia, IPCL vessels). (public dataset)	ML on textural information to classify laryngeal tissues. Prec.= 0.93 Rec. = 0.94 F1 = 0.92	The performance of ML metrics is inadequate. DL performed better on this dataset. Patches are too small; in real-time implementation, techniques for preprocessing are required.
Barbalata (Barbalata	Laryngeal tumor detection and	Larynx	Classification	Linear discriminant analysis and matched filtering	120 NBI laryngeal SCC images.	Segmentation of the (RoI, vessels), LSCC classification.	The dataset is limited. Statistical techniques are less adaptable because

& Mattos, 2016)	classification in endoscopic video				(private dataset)	RoI segmentation (DSC = 0.69); vessel segmentation (DSC = 0.76); classification (accuracy = 0.84).	of heterogeneous UADT anatomy.
Witt (Witt et al., 2014)	Detection of chronic laryngitis due to laryngopharyngeal reflux using color and texture analysis of laryngoscopy images	Larynx and / hypopharynx	Classification	Hue and texture feature extraction + multilayer perceptron.	Images from 62 videos. (private dataset)	Classification of laryngopharyngeal reflux. Results comparison across different regions. Accuracy = 0.80; AUC = 0.89	The imaging modalities and disease diagnosis are not clearly defined. The standard performance metrics for clinical application not presented.
Mascharak (Mascharak et al., 2018)	Detecting oropharyngeal carcinoma using multispectral, narrow-band imaging and ML.	Oropharynx	Classification	Color and texture feature extraction + naive Bayesian classifier	NBI and WL images of 30 patients. (private dataset)	Identification of oropharyngeal carcinoma. Accuracy = 0.52-0.70. sensitivity = 0.47-0.71. specificity = 0.58-0.69. AUC = 0.51-0.92	The performance of ML metrics is inadequate. The dataset is limited. techniques are less adaptable because of uneven illumination color of UADT during endoscopy.
Heo (Heo et al., 2022)	DL model for tongue cancer diagnosis using endoscopic images classification.	Oral cavity	Classification	CNN, VGG, DenseNet, MobileNet, ResNet, and EfficientNetB3	Total images=5576 (malignant= 1941, non-malignant=3635). External validation alos performed. (private dataset)	Classification of oral cancer. comparison with human experts. Accuracy = 0.64-0.84. Prec. = 0.52-0.79. F1= 0.57-0.78. AUC = 0.71-0.89	The attention mechanism was not performed to show the predictions. Qualitative results were not performed in this study. The sub benign classes still an open challenge in this research.
He (Y. He et al., 2021)	CNN-based method for LSCC diagnosis	Larynx	Classification	InceptionV3	4591 NBI images. External validation. Comparison with human experts. (private dataset)	Accuracy = 0.91. Sens. = 0.90. spec. = 0.91; AUC = 0.96. time per frame = 0.010 s	This study only classifies benign vs malignant images. The sub benign classes still an open challenge in this research.

Zhao (Q. Zhao et al., 2022)	Vocal cord lesions classification based on CNN and transfer learning	Larynx	Classification	MobileNetV2	4-class (normal, polyp, keratinization, carcinoma) vs 2-class (urgent, nonurgent) in 5122images. Comparison with human experts. (private dataset)	4-class: Accuracy = 0.80; F1= 0.78; AUC = 0.96. 2-class: Accuracy =0.99. AUC = 0.98. time = 0.011 s	The attention mechanism was not performed to show the predictions. The external validation no presented for model generalization ability.
Xu (Xu et al., 2022)	DL for nasopharyngeal carcinoma identification using both WL and NBI endoscopy	Nasopharynx	Classification	Siamese DCNN	4783 NBI images. 2,898 WLI images and 1,885 NBI images. (private dataset)	Evaluation both at an image-level and patient-level. Image-level: Accuracy = 0.95. AUC = 0.99; patient-level: accuracy = 0.96. time = 0.039 s	The study is limited to classifying only two classes (nasopharyngeal carcinoma vs normal). No external validation performed.

2.4 UADT Detection and Delineation

AI can further help clinicians in localized regions of interest during the endoscopic examination, highlighting eventual abnormalities, and even predicting the nature of those. Yet, the detection of lesions in UADT endoscopic images has been the focus of a small number of investigations so far. A representative work is that of Inaba et al who trained a DL model (Inaba et al., 2020) to localize eventual mucosal cancerous lesions in the pharynx and larynx during the transoral passage of esophagogastroduodenoscopy. They reported an accuracy of 0.97 and an AUC of 0.98 in laryngeal (mainly supraglottic) and pharyngeal cancer detection by analyzing 1200 NBI mucosal images. Tamashiro et al also used esophagus gastroduodenoscopy (Tamashiro et al., 2020) WL and NBI images for screening potential pharyngeal lesions. They reached an accuracy of 67.0% and a sensitivity of 80.0% after having trained the model with 5403 images of healthy pharynxes and pharyngeal cancer. Moreover, even if the authors used the model to analyze only still frames and not video endoscopies, they reported inference times of 0.020 seconds, thus compatible with real-time processing. Nevertheless, the studies mentioned above did not report F1 scores nor mAP values, thus lacking a more comprehensive performance assessment. Similarly, Kim et al tried to achieve the detection of laryngeal masses by using an instance segmentation CNN (Kim et al., 2021), achieving a variable F1 score ranging from 0.67 to 0.83. However, the model did not achieve real-time performances (5-40 seconds per frame).

2.5 UADT Delineation

AI-based models can also be used for automatically delineating the boundaries of structures and lesions of the UADT. This task could have interesting clinical implications, especially when applied to cancer lesions. These models could potentially help to increase the accuracy of incisional biopsies and improve the definition of the resection margins in the operating room. Attempts to automate the segmentation of cancer tissue in these regions have been made with good results by Li et al (DSC = 0.78) (C. Li et al., 2018) for the nasopharynx and by Paderno et al for the oropharynx (DSC = 0.76) (Paderno, Piazza, et al., 2021a). The latter took advantage of the NBI technique which is known to be especially useful in the case of unknown primary. In these contexts, AI-based segmentation models have been also used for delineating regions of interest in endoscopic laryngeal images, such as the vocal folds, the epiglottis, and the glottic space (Hamad et al., n.d.). This includes models with the potential for assessing the respiratory space and vocal fold angles (Ding et al., 2022). The combination of NBI with automatic cancer boundaries assessment is of particular interest in UADT carcinoma treatment. Numerous authors have reported that NBI allows for a better definition of excisional margins with consequently improved resections of head and neck mucosal cancer (Farah et al., 2016; Garofolo et al., 2015). Finally, another use for AI-based automatic segmentation is the one reported by Adamian et al34 who implemented a CNN for vocal fold tracking in video endoscopies (Adamian et al., 2021). The model could estimate the anterior glottic angle by tracking the vocal folds' free borders, which may allow us to automatically identify patients with vocal fold palsy.

Table 2. 2 The literature on the detection and delineation of UADT various subsites or cancer diagnosis has been summarized, along with key author contributions and limitations.

First author	Work	Anatomy site	Task	Methods	Datasets	Outcomes	Limitations
Kim (Kim et al., 2021)	Automated laryngeal mass detection algorithm for home-based self-screening test based on CNN.	Larynx	Detection + Segmentation	Mask RCNN	1224 laryngeal images (1153 mass cases and 71 for no-mass). (private dataset)	Detection of laryngeal masses and vocal folds: Rec.= 0.80, Prec. = 0.73, Accuracy = 0.62, F1 = 0.76	The study did not clearly mention the imaging modalities. The external validation no included in this study. The major segmentation performance metrics not evaluated

Matava (Matava et al., 2020)	A CNN for real time classification, identification, and labelling of vocal cord.	Larynx + trachea	Detection	MobileNetV1, Inception, ResNet	457 bronchoscopy and laryngoscopy images. (private dataset)	Identification and classification of vocal cords or tracheal rings. Sens. = 0.56-0.86. Spec. = 0.85-0.97. time = 0.200-0.067 s	The description and distribution of dataset not clearly described. No external validation performed. Lack of detection evaluations (IoU, mAP).
Inaba (Inaba et al., 2020)	AI system for detecting superficial laryngopharyngeal cancer with high efficiency of DL	Larynx + hypopharynx	Detection	RetinaNet	1200 NBI mucosal images. (private dataset)	Accuracy = 0.97. Sens. = 0.95. Spec. = 0.98	No external validation included. Lack of detection evaluations (IoU, mAP). AI vs more than one senior expert comparison not included. WL imaging was not included.
Tamashiro (Tamashiro et al., 2020)	AI based detection of pharyngeal cancer using. CNN.	Pharynx	Detection	Single Shot MultiBox Detector	5403 NBI and WL image training images. validation set of 1912 with and without cancer images. (private dataset)	Accuracy = 0.67. Sens. = 0.80. spec. = 0.57. time = 0.020 s	Implemented only on static WL and NBI imaging, the model performance not performed on external clinical videos. Lack of detection evaluations (IoU, mAP).
Adamia (Adamian et al., 2021)	An open-source computer vision tool for automated vocal fold tracking from video endoscopy	Larynx	Segmentation	DeepLabCut (modified ResNet with deconvolutional layers for key point localization)	Normal VFs patients =20. unilateral VF patients=20. (private dataset)	VF anterior commissure key points localization to estimate the glottal opening angles. Pearson correlation coefficient (Estimated angles vs expert measured ones) = 0.97. AUC = 0.88 in estimating vocal	The dataset is limited. Techniques are less adaptable because of uneven illumination color of UADT during endoscopy. No external validation included.
Parker (Parker et al., 2021)	ML in laryngoscopy analysis: a proof of concept observational study for the	Larynx	Segmentation	UNET	Segmentation of ulcerations and granulomas on 127 laryngoscopes frames.	Sens. = 0.63-0.82. Spec. = 0.99. AUC = 0.89-0.99	The limited number of training images (100). The major segmentation metrics (DSC, IoU) not evaluated.

	identification of post-extubation ulcerations and granulomas.				(private dataset)		
Paderno (Paderno, Piazza, et al., 2021a)	DL for automatic segmentation of oral and oropharyngeal cancer using NBI imaging.	Oral cavity and oropharynx	Segmentation	UNET, UNET3, ResNet	Oral cavity = 100 and Oropharynx = 116 NBI Frames. (private dataset)	Oral cavity: DSC = 0.59-0.66. Accuracy = 0.86-0.89; recall = 0.67- 0.75; prec. = 0.59-0.67. Oropharynx: DSC = 0.69-0.76. accuracy = 0.78-0.84; Rec. = 0.78-0.86; Prec. = 0.62-0.80. Time = 0.059-0.115 s	This study has small datasets and does not include WL imaging. The external validation and transfer learning approach still open challenge in this study.
Lin (J. Lin et al., 2019)	Quantification and analysis of laryngeal closure from endoscopic videos	Larynx	Segmentation + Detection	Custom-made: Encoder (for detection of the RoI) +FCNN (for segmentation)	1206 frames. (private dataset)	Real-time segmentation of glottic opening. IoU = 0.65-0.85. Time = 0.055-0.060s	The work did not identify laryngeal cancers features. Only focus to identify VF and glottic areas.
Li (C. Li et al., 2018)	Development and validation of an endoscopic images-based DL model for detection with nasopharyngeal malignancies	Nasopharynx	Segmentation	Inception-based FCNN	Segmentation on 27,536 nasopharyngeal carcinoma images. (private dataset)	Performance comparison with human experts. DSC = 0.78; Accuracy = 0.89. sensitivity = 0.91; Spec. = 0.83. AUC = 0.94; Time = 0.028s	The large dataset only includes WL imaging there is no evident of using NBI modality.
Laves (Laves et al., 2019)	A dataset of laryngeal endoscopic images with comparative study on CNN semantic segmentation	Larynx	Segmentation	UNET, SegNet, Enet, ErfNet	Dataset of 536 micro laryngoscopy images. (private dataset). (public dataset)	Segmentation of anatomical structures and surgical tools. IoU = 0.54-0.84; inference Time = 0.048-0.008 s	This work considered only two patients that make less variation of dataset. The study focusses to detect larynx site not tumor delineation.
Pan (Pan et al., 2022)	RANT: cascade reverse attention segmentation framework with hybrid transformer for laryngeal endoscope images.	Larynx	Segmentation	Custom-made (RANT)	Segmentation of anatomical structures and surgical tools on the same dataset of (Laves et al., 2019)	DSC = 0.83-0.93; IoU = 0.77-0.89. inference time = 0.004 s	The study improved the segmentation performance, but limited variation and no external validation is still missing.

					(public dataset)		
Ji (Ji et al., 2020a)	A multi-scale recurrent fully convolution neural network for laryngeal leukoplakia segmentation	Larynx	Segmentation	Custom-made (Boldface Multi-scale Net)	Segmentation of leukoplakia on 649 laryngoscopy images. (private dataset)	Accuracy = 0.99; Rec. = 0.89. Prec.= 0.74. F1 score = 0.78. IoU = 0.83. Time = 0.205s	The study segmented only leukoplakia tumors. The major segmentation performance was not presented well. No external validation and more than one human annotation involved.
Hamad (Hamad et al., n.d.)	Automated segmentation of the VF in laryngeal endoscopy videos using convolutional regression networks.	Larynx	Segmentation	13-layer fully convolutional regression network	Patients=20. endoscopic frames = 7892.	Segmentation of the vocal fold and the glottal region. DSC = 0.86. Accuracy = 0.95. IoU = 0.88.	The study was not delineated any specific disease of larynx.
Ding (Ding et al., 2022)	Automatic glottis segmentation for laryngeal endoscopic images based on UNET.	Larynx	Segmentation	Custom-made (CNDA-UNET)	Patients=90. Total images = 928 (healthy =112, hypertrophic =218 polyp=532, Nodule=66)	Segmentation of the glottal space. DSC = 0.93. Sens. = 0.93. Spec. = 0.93.	The attention mechanism not presented for tumor prediction. No external validation used. No more than one human expert comparison found.

2.6 Summary

In the literature review it is observed that DL models applied to UADT endoscopy showed growing field that is approaching performances appropriate for clinical use, thus holding the promise to be a turning point in otolaryngological practice in the coming years. In this respect, investing in models suitable for real-time use might be the key to clinical implementation. Technically, to obtain a reliable AI system, it is of paramount importance to gather extensive image datasets and possibly establish publicly available open resources. Furthermore, every AI model should be validated on external cohorts to ensure a sufficient level of robustness for clinical use. The use of a including important evaluation metrics is recommended to allow a comprehensive assessment and benchmarking of the different AI models for facilitating their comparative analysis.

Chapter 3

SegMENT: DL model for UADT cancer Delineation

3.1 Overview

In this chapter UADT cancer delineation and margins evaluation explored using DL technique on NBI and WL imaging technology. A new CNN-based semantic segmentation model for video endoscopy of the UADT, named *SegMENT* specifically developed the identification and segmentation in videoendoscopic frames of UADT cancer, particularly LSCC, OCSCC and OPSCC. The *SegMENT* relies on *DeepLabV3+* CNN architecture, modified using Xception (Xception, 2017) as a backbone and incorporating ensemble features from other CNNs. The performance of *SegMENT* was compared to state-of-the-art CNNs (*UNet, ResUNet, and DeepLabv3*). *SegMENT* was then validated on two external datasets of NBI images of oropharyngeal (OPSCC) and oral cavity SCC (OSCC) obtained from a previously published study (Paderno, Piazza, et al., 2021a). The impact of in-domain transfer learning through an ensemble technique was evaluated on the external datasets also included in this chapter. The *SegMENT* achieved promising performances, showing that automatic tumor segmentation in endoscopic images is feasible even within the highly heterogeneous and complex UADT environment. *SegMENT* outperformed the previously published results on the external validation cohorts. The model demonstrated potential for improved detection of early tumors, more precise biopsies, and better selection of resection margins.

Related Publications

1. **Muhammad Adeel Azam**, Sampieri C, Ioppi A, Benzi P, Giordano GG, De Vecchi M, Campagnari V, Li S, Guastini L, Paderno A, Moccia S. "**Videomics of the upper aero-digestive tract cancer: Deep learning applied to white light and Narrow Band Imaging for automatic segmentation of endoscopic images**". Frontiers in Oncology. 2022 Jun 1; 12:900451. https://doi.org/10.3389/fonc.2022.900451.

3.2 Materials and methods

3.2.1 Data acquisition

Recorded video endoscopies of patients treated between 2014 and 2019 at the Unit of Otorhinolaryngology – Head and Neck Surgery of the IRCSS Ospedale Policlinico San Martino, Genoa, Italy, were retrospectively revised. Selection criteria comprised a pathology report positive for LSCC and the availability of at least one recorded video endoscopy before treatment. Local Institutional Review Board approval was obtained (CER Liguria: 230/2019). All patients were first examined through trans nasal video laryngoscopy (HD Video Rhino-laryngoscope Olympus ENF-VH, Olympus Medical System Corporation, Tokyo, Japan) in the office before treatment. For those submitted to transoral laryngeal microsurgery, an additional intraoperative endoscopic evaluation was conducted using 0°, 30° or 70° telescopes coupled to a HD camera head connected to a Visera Elite CLV-S190 light source (Olympus Medical System Corporation). In both settings, a thorough examination was conducted under WL video endoscopy, then switching to NBI. From each of the collected videos, four expert physicians extracted, when available, one WL and one NBI frame. Frames were selected to be the most representative of the lesion and possibly offer a clear view of its boundaries. Priority in image selection was given to steady frames with few artifacts and blur. The extracted videoframes were then labeled by the same physicians using the VGG Image Annotator (VIA) 2.0 (https://www.robots.ox.ac.uk/vgg/software/via/), an open-source web-based annotation software. The annotation process consisted of the manual segmentation of the neoplastic lesion borders: this was done by manually drawing its contour following the visible tumor margins either in WL or NBI in the same fashion.

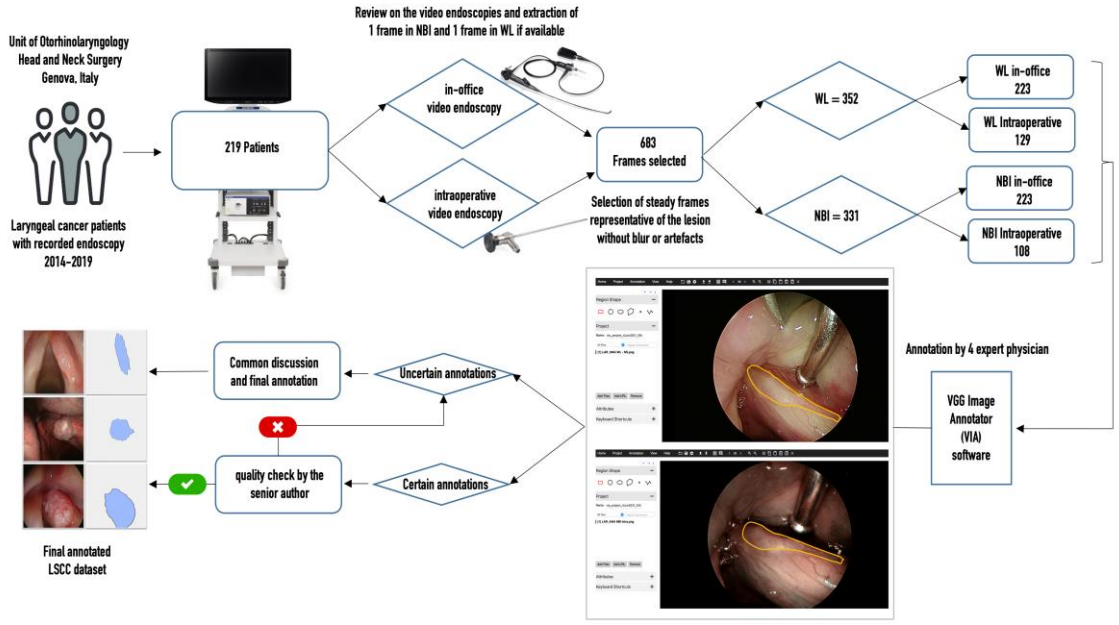


Figure 3. 1 Flow chart of data acquisition for the laryngeal squamous cell carcinoma (LSCC) dataset. WL, white light; NBI, narrow band imaging.

The pixel comprised in the traced mask were then classified as "LSCC" and no special labeling was assigned for NBI compared to WL images. If multiple lesions were visible, multiple segmentations were carried out to select all the LSCC pixels in the image. If one physician was not sure about the correctness of some annotations those were revised collectively by all four otolaryngologists. Finally, the senior author (G.P) checked the appropriateness of the annotations and referred the inexact ones to the collective revision. Figure 3.1 summarizes the data acquisition process. Finally, two datasets of NBI endoscopic images were obtained from a previous study on automatic segmentation of OCSCC and OPSCC (Paderno, Piazza, et al., 2021a). The datasets included the corresponding ground-truth annotations, as previously described.

3.2.2 *SegMENT* architecture

Figure 3.2 describes the architecture of *SegMENT*. The model is based on concepts introduced by the *DeepLabV3+* model (L. C. Chen et al., 2018a) which were expanded and customized here for precise cancer segmentation in UADT video endoscopies. *The segMENT* backbone was built on the *Xception* architecture (Xception, 2017), which was chosen for its high benchmark results on ImageNet (Jia Deng et al., 2009), the largest dataset of natural images publicly available. *Xception* has a smaller number of parameters compared to the most popular CNN architectures like *VGG16* and *VGG19*, but an almost equal number of parameters than *Resnet50* and *DensNet121* (Bianco et al., 2018). The *Xception* backbone

architecture is composed of two primary components: convolutional layers with pooling for feature extraction, and fully connected dense layers at the top of the network for classification. To customize this backbone network for segmentation tasks, we removed the network fully connected dense layers and maintained only the feature extraction layers. The functionalities to resize the input frames into 256x256 pixels were maintained. In addition, we used three-skip connections to get feature map outputs (of size 16×16, 32×32, and 64×64 pixels) from *Xception* backbone convolution layers. These were then merged using Atrous Spatial Pyramid Pooling (ASPP) blocks in the encoder part of our model.

The heterogeneous nature of UADT lesions in terms of dimension, form, and contour, we designed the encoder part of *SegMENT* to use two ASPP blocks. These can potentially contribute to increased segmentation accuracy. The ASPP block-1 is designed for high-level image features (shapes, tumor composition, etc.). It is fed with 16×16 pixels input images directly from the *Xception* backbone and contains 256 filters. ASPP block-1 comprises five rates of dilation convolution layers (1, 3, 6, 12, and 18), which were chosen given the small-scale input images. The ASPP block-2 is designed for low-level image features (edges, contours, texture) and accepts two scales of input images (32×32 and 64×64 pixels). It is also comprised of five rates of dilation convolution layers (1, 6, 12, 18, and 24), which are higher here because of the larger scale of the input images. The ASPP block-2 employs 48 filters for 64×64 pixels input images and 64 filters for 32×32 pixels input images. Finally, the convolutional layers of *SegMENT* use the Mish activation function (except for the last layers that use a sigmoid activation function). This activation function was selected to replace the traditional Relu activation function as it was shown to provide better performance (Misra, 2019). The decoder section accepts the encoder outputs with the three resized (16×16, 32×32, and 64×64 pixels) images. The output images of ASPP block-1 are four times up-sampled using the UpSampling2D layer and bilinear interpolation technique to yield 64×64 pixels images. The images from ASPP block-2 with 32×32 pixels are up-sampled two times. Next, 64×64-pixel images from ASPP block-2 are concatenated with the up-sampled 64×64-pixel images generated by ASPP block-1. This concatenation produces 64×64 pixels images. Afterward, a 2D convolutional layer with a kernel size of 3×3 and 256 filters are applied to these images. Again, images are further up-sampled four times to get 256×256-pixel images. Finally, the segmented tumor area is retrieved through a bitwise-AND operation, which generates an output image of 256x256 pixels.

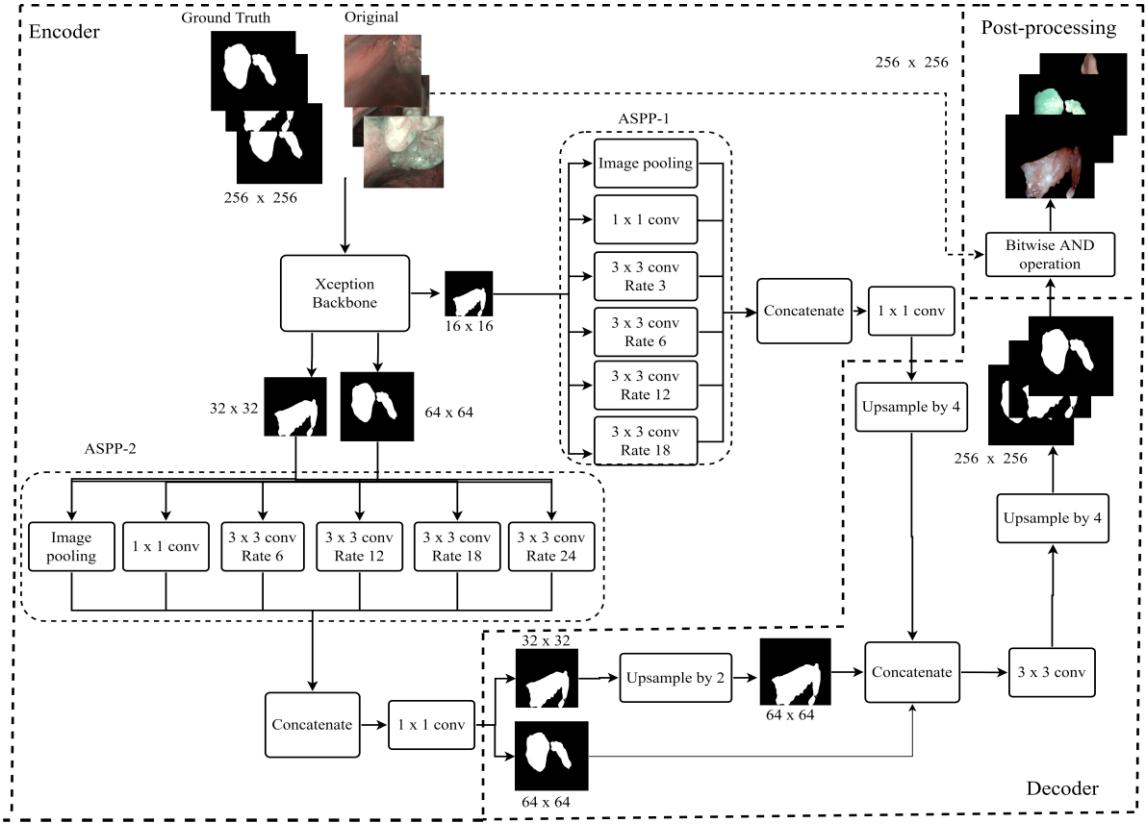


Figure 3. 2 Workflow diagram of the proposed *SegMENT* semantic segmentation architecture. The Xception backbone is customized for segmentation tasks by eliminating the dense layers for classification while maintaining the feature extraction layers. The model encoder uses two *ASPP* blocks designed, respectively, for high- and low-level image features. The decoder section concatenates the information coming from the two *ASPP* blocks, producing segmentation masks.

3.2.3 Baseline models

Three state-of-the-art baseline CNNs for semantic segmentation (i.e., *UNet, ResUNet, and DeepLabv3+*) were investigated and tested as part of a comparative study. Each segmentation model incorporates a backbone architecture for feature extraction. The backbones considered were *VGG16, VGG19, MobileNetV2, ResNet50, ResNet101V2, DenseNet121*, and *Xception*. Pre-trained backbone weights from the ImageNet dataset were used (Jia Deng et al., 2009). The *UNet* network (Ronneberger O, Fischer P, 2015) is equipped with an encoding path that learns to encode texture descriptors and a decoding path that achieves the segmentation task. *ResUNet* (Z. Zhang et al., 2018) is a segmentation model based on the *UNet* architecture that implements residual units instead of plain neural units, to obtain good performance with fewer parameters. Finally, *DeepLabV3+* (L. C. Chen et al., 2018b) uses dilated separable convolutions and spatial pyramid pooling in a U-shaped architecture to produce accelerated inference times and reduced loss values. The training

and testing of these baseline models were performed in the same environment and using the same data as *SegMENT*.

3.2.4 Ensemble technique

Multiple ensemble techniques are described in the literature for decreasing segmentation errors and optimizing efficiency (Mahendran et al., 2019; Nourani et al., 2018; Pham et al., 2021). Their utility becomes evident especially when the available training dataset for a new application area is small or highly heterogeneous, such as the case of the OCSCC and OPSCC datasets. In this work, we evaluated the value of using the weighted average ensemble approach (Laves et al., 2019) during the testing phase of the segmentation network. This technique implements a weighted ensemble of predictions from different models. Its integration into the proposed segmentation architecture is shown in Figure 3.3. In LSCC segmentation, the predictions of the two best-performing models (*SegMENT* and *UNet-DenseNet121*) were combined to improve the accuracy of laryngeal cancer segmentation. Initially, the two models were independently trained on the LSCC dataset. Their predictions were then combined following the weighted average ensemble approach. In OCSCC segmentation tasks, three different predictions were ensembled during testing. One was taken from a *SegMENT* model trained on the OCSCC dataset. The other two predictions were taken from *SegMENT* and *UNet-DenseNet121* trained on the LSCC dataset. To produce the single final output prediction, the three output predictions were multiplied by the assigned weight values, which were obtained through a grid search technique. The same strategy was used for OPSCC segmentation, with the difference that the first *SegMENT* model was trained on the OPSCC dataset. The other two predictions were taken from *SegMENT* and *UNet-DenseNet121* models trained on the LSCC dataset, as before.

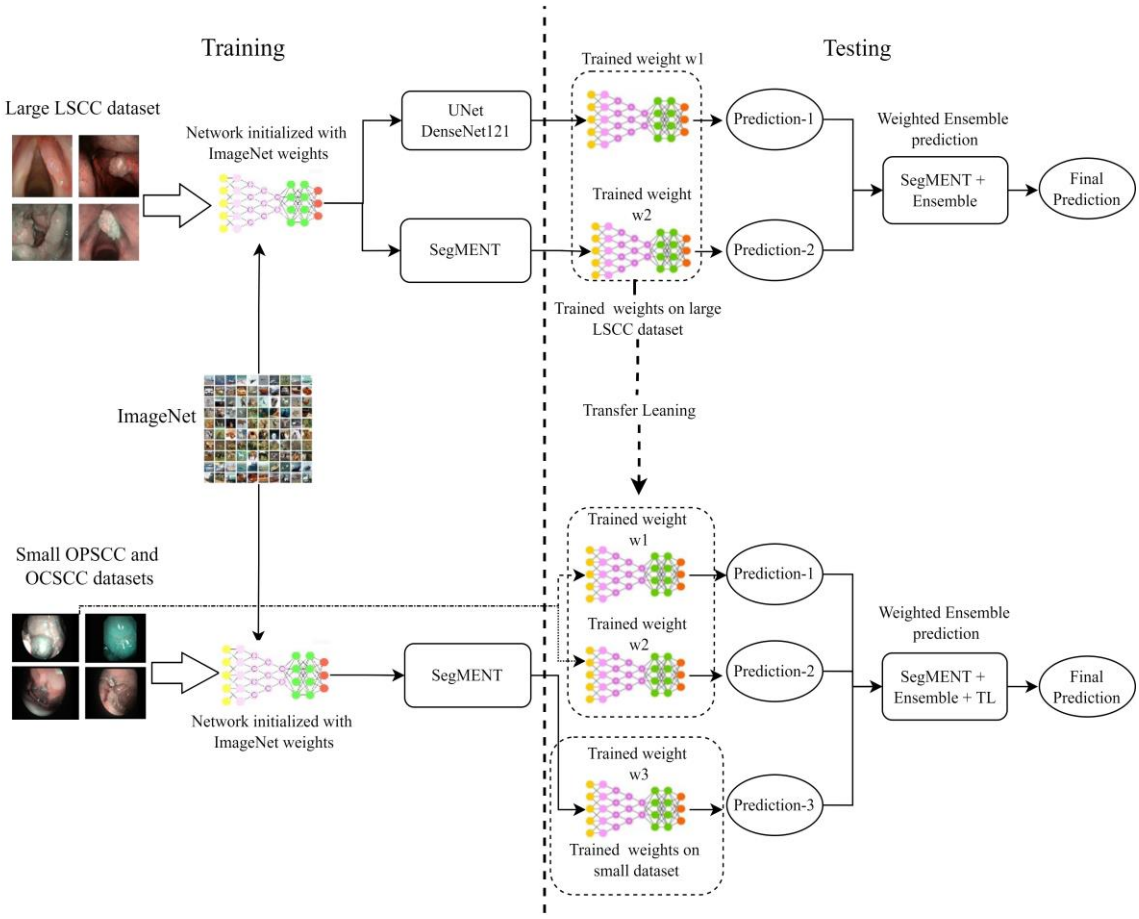


Figure 3. 3 Diagram describing the training and testing of the models assessed in this work. Initially, the models are initialized with weights obtained from training on the ImageNet dataset. Next, the models are trained on the specific datasets of interest. During testing, the trained models provide segmentation predictions, and ensemble used to generate the final segmentations. In addition, in-domain transfer learning (TL) can be used to enhance the segmentation performance on small datasets using trained weights from other anatomical subsites. LSCC: laryngeal squamous cell carcinoma. OCSCC: oral cavity squamous cell carcinoma. OPSCC: oropharyngeal squamous cell carcinoma.

3.3 Experimental Setup

3.3.1 Training parameters

For training of *SegMENT*, the Tverky loss function (Abraham & Khan, 2018) was used. This loss function is used for highly unbalanced datasets. In our model, we used a learning rate of 0.001 with a batch size of 8 images per epoch during training. The learning rate decay was set to a factor of 0.1. If the training loss did not improve after four consecutive epochs of learning, the decay was slowed down. Data augmentation was used during training to

increase the variability of the training dataset: flip, crop, translation, rotation, and scaling were applied, as well as hue, brightness, and contrast augmentation. A Tesla K80 GPU with 12 GB of memory and an Intel(R) Xeon(R) CPU running at 2.20 GHz with 13 GB of memory using Keras and a Tensorflow (Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, 2016) back-end were used for all experiments.

3.3.2 Validation on LSCC dataset

The first experiments were performed on the LSCC dataset, which was split into a training and a validation/test sets using a respective split ratio of 80% and 20% with randomly selected images. Initially, the baseline models and *SegMENT* were trained and tested on the LSCC dataset starting from pre-trained weights obtained from ImageNet. Afterward, the weights of *SegMENT* and *UNet-DenseNet121* were combined using the weighted ensemble method in the testing phase.

3.3.3 Validation on OCSCC and OPSCC datasets

The OCSCC and OPSCC datasets were separately used to validate *SegMENT* on these different UADT sites. Each dataset was split into a training and a validation/test groups with a 70/30 percent split ratio based on a random image selection process. *SegMENT* was first trained and tested separately on the OCSCC and OPSCC datasets, starting from ImageNet pre-trained weights without applying the ensemble technique. Following this, the described in-domain transfer learning (TL) method based on a weighted ensemble technique was used to assess potential performance improvements. We hypothesized that, compared to the standard TL provided by ImageNet, which is based on natural images (such as daily objects, and animals), a specific in-domain TL based on LSCC trained weights might enable better performance on images from other UADT regions. Therefore, the *SegMENT* ensemble model, incorporating features initially learned from the LSCC dataset, was tested on the OCSCC and OPSCC datasets during the testing phase of the segmentation framework.

3.3.4 Outcome analysis

The outcomes of each DL model were evaluated by comparing the predicted segmentations with the manual annotations performed by expert physicians (i.e., the ground-truth segmentations). Standard evaluation metrics for semantic segmentation were used as previously reported (Qiu et al., 2021). A classification of each pixel in the images as 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.

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

- *Precision* (positive predictive value): the fraction of pixels that are true positives (correctly predicted pixels of the targeting class) among the total predicted pixels:

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

- *Recall* (sensitivity): the fraction of pixels that are true positives among the total ground truth segmented pixels:

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

- *Dice similarity coefficient* (DSC): represent the harmonic weight of Precision and Recall values (also called F1 score):

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

- *Intersection over Union* (IoU): the fraction of pixels that are true positives among the union of pixels that are positive predictions and belong to the target class (Figure 3.4).

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

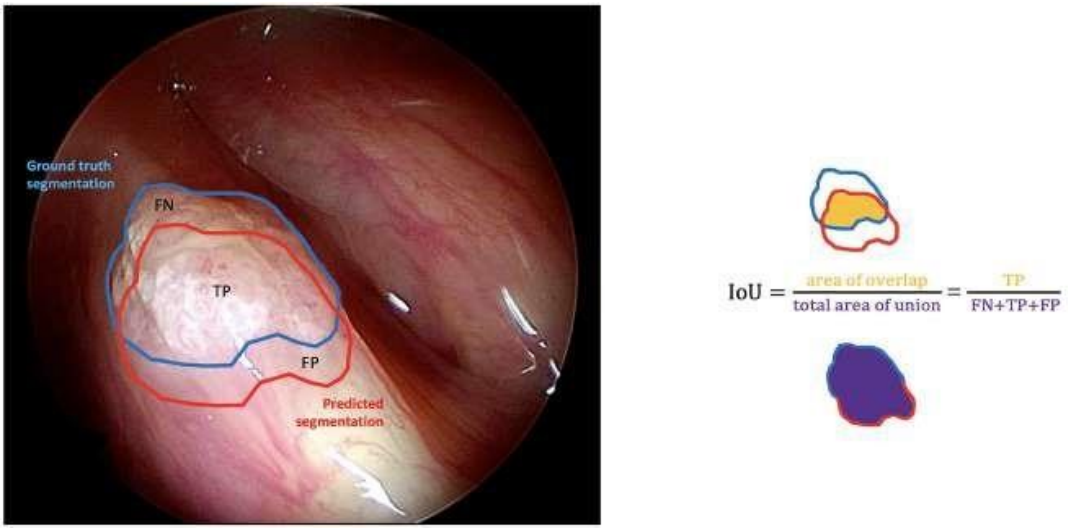


Figure 3. 4 Graphical representation of the intersection over union (IoU) calculation on a white light right glottic cancer intraoperative videoframe. The boundary traced in light blue represents the ground truth segmentation provided by an expert clinician, while the red one represents the model prediction. The IoU is calculated by dividing the overlapping area (containing the true positive pixels) by the total area of union (encompassing the false negative, true positive, and false positive pixels).

Table 3. 1 Performance evaluation of different semantic segmentation models during testing on the laryngeal squamous cell carcinoma (LSCC) dataset. The results represent the median scores from all the tests for each metric. Values in bold denote the best results. IoU: intersection over union. DSC: dice similarity coefficient.

Dataset	Model	Backbone	IoU	DSC	Recall	Precision	Accuracy
Laryngeal cancer (LSCC)	UNet	–	0.264	0.418	0.641	0.351	0.895
	ResUNet	–	0.309	0.473	0.486	0.490	0.928
	UNet	VGG16	0.595	0.746	0.817	0.823	0.968
	UNet	VGG19	0.618	0.763	0.900	0.827	0.967
	UNet	MobileNetV2	0.469	0.639	0.579	**0.855**	0.961
	DeepLabV3+	ResNet50	0.587	0.740	0.714	0.830	0.968
	UNet	ResNet50	0.476	0.645	0.834	0.745	0.952
	UNet	ResNet101V2	0.586	0.739	0.841	0.788	0.963
	UNet	DenseNet121	0.677	0.807	0.847	0.840	0.971
	SegMENT	Xception	**0.686**	**0.814**	0.916	0.830	0.969
	SegMENT ensemble	Xception	0.685	**0.814**	**0.951**	0.785	**0.973**

3.3.5 Statistical analysis

Differences in distributions of continuous variables among more than two independent groups were assessed with the analysis of variance (ANOVA) test. Post-hoc analysis was

performed using Tukey's multiple comparisons test to control for the inflated Type I error. A $p<0.005$ was considered significant. Data analysis was carried out using statistical functions (scipy.stat) and statistical models (statmodels v0.13.2) libraries in python (v3.9).

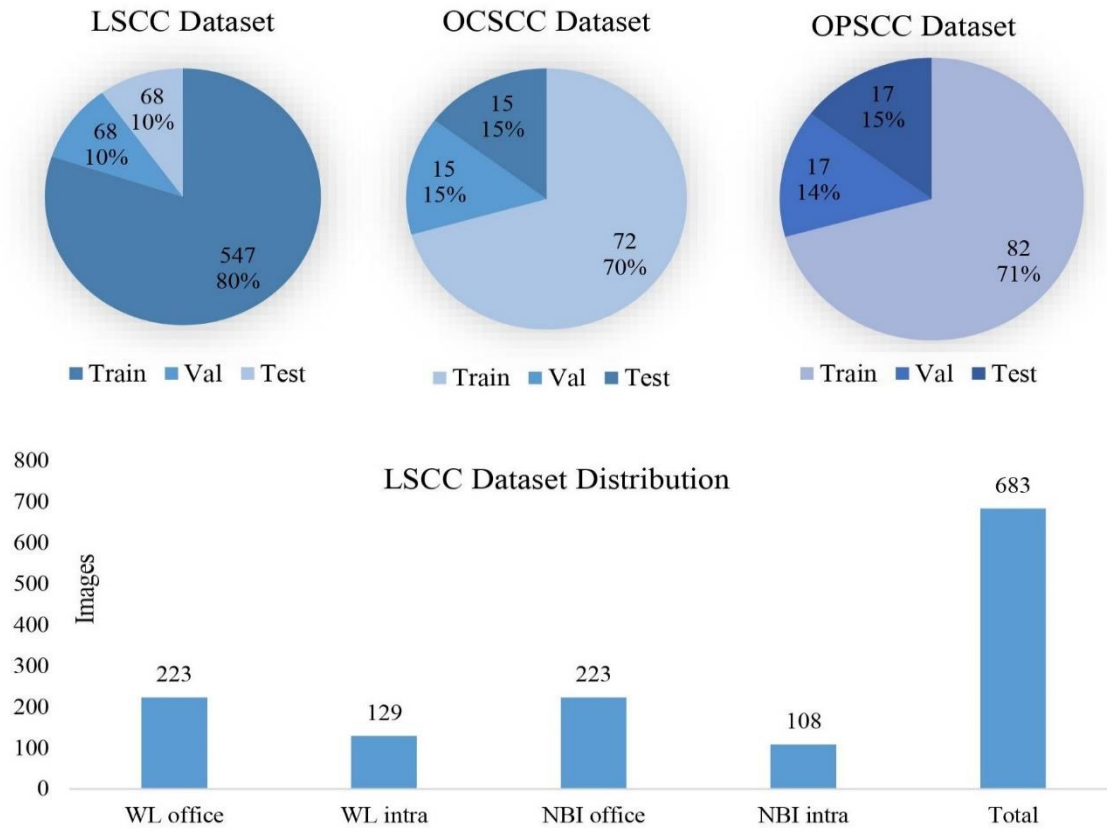


Figure 3. 5 Overview of final configuration of datasets for experiments DL models. Left: Pie chart represents the LSCC dataset split before training *SegMENT* and other DL models, Middle: Pie chart indicates the split ratio of OCSCC dataset, while Right: Pie chart is the overview of OPSCC dataset division. The below bar plot figure indicates the detail distribution of large LSCC dataset.

3.4 Results

3.4.1 Datasets

Two hundred and nineteen patients with a mean age of 67.9 years (SD ± 11.8 years) were enrolled. Among these, 196 (89.4%) were males and 23 (10.6%) females. A total of 683 frames representing LSCC were extracted from video laryngoscopies. 223 were in-office WL, 129 intraoperative WL, 223 in-office NBI, and 108 intraoperative NBI images. The semantic segmentation models were trained on the LSCC dataset after random distribution of the images into a training set (547 images), a validation set (68 images), and a test set (68 images). The external validation cohorts comprised 102 images for OCSCC (72 for training,

15 for validation, and 15 for testing) and 116 images for OPSCC (82 for training, 17 for validation, and 17 for testing). However, while the images used were the same, it must be highlighted that it was not possible to perform the same exact image allocation in the training/testing cohorts as in the previous study**.** Figure 3.5 represents an overview of the final configuration of the LSCC, OSCC and OPSCC datasets.

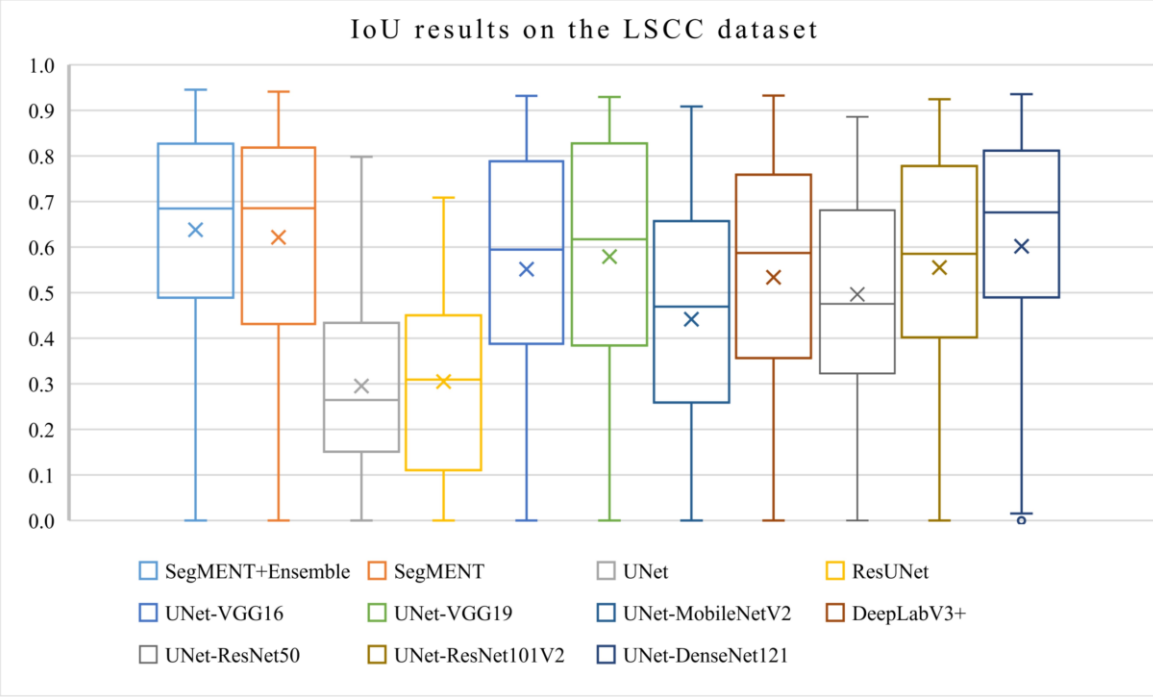


Figure 3. 6 Boxplots of Intersection over Union (IoU) results from *SegMENT* ensemble and the other state-of-the-art segmentation models on the laryngeal squamous cell carcinoma (LSCC) dataset. The cross sign represents the mean value while the horizontal line inside the boxplot shows the median value.

3.4.2 Experimental Results

During experiments, it was observed that the proposed *SegMENT* model outperformed the other state-of-the-art segmentation models. Among the baseline models, the *UNet-DenseNet121* performed better than the other baseline models. Thus, the ensemble technique was used to integrate the training weights from this model with those from *SegMENT*, leading to better segmentation performance during testing. The performances of the models on the test set of LSCC are shown in Table 3.1. The comparison with previously published work by (Paderno, Piazza, et al., 2021a) on external test set (OCSCC, OPSCC) are shown in Table 3.2. The median values achieved by *SegMENT* with the ensemble technique on the LSCC dataset were: IoU=0.685, DSC=0.814, recall=0.951, precision=0.785, and accuracy=0.973. The box plots showing the IoU and DSC score performances of *SegMENT*

and the other state-of-the-art segmentation models during testing on the LSCC dataset are shown in Figure 3.6 and Figure 3.7. The processing rates of all base-line models ranged from 6.3 to 8.7 frames per second (fps), while the proposed ensemble model processed an average of 2.1 fps (taking 0.48 seconds to process a single frame). Examples of LSCC segmentation including ground-truth labels and the resulting automatic segmentations are shown in Figure 3.8. The *SegMENT* model pre-trained on ImageNet already performed better than the previous study CNNs on all metrics. The in-domain TL also helped to improve the results, especially for the OPSCC dataset. Indeed, the median metrics on the OCSCC and OPSCC datasets improved compared to the previously published by, respectively, 10.3% and 11.9% for DSC, 15.0% and 5.1% for recall, 17.0% and 14.7% for precision, and 4.1% and 10.3% for accuracy. The processing rate of our model was 3.9 fps on both the OCSCC and OPSCC datasets. Examples of segmentation of OCSCC and OPSCC frames displaying both the ground-truth labels and the resulting automatic segmentations are shown in Figure 3.9.

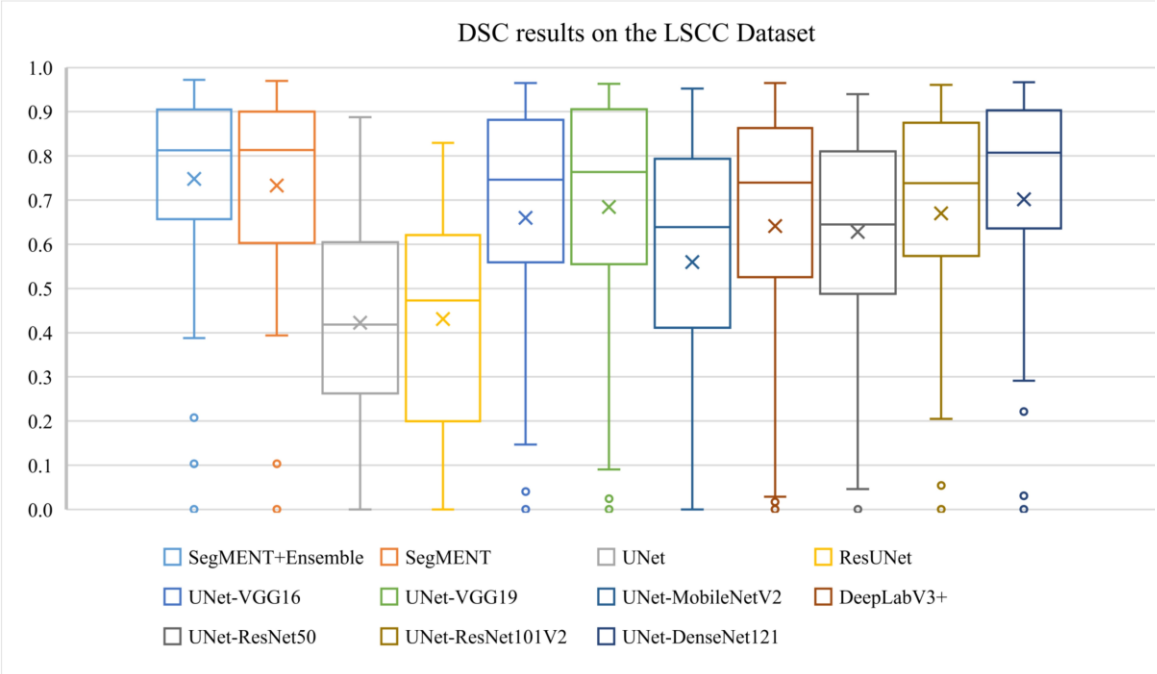


Figure 3. 7 Boxplots of Dice similarity coefficient (DSC) result from *SegMENT* ensemble and the other state-of-the-art segmentation models on the laryngeal squamous cell carcinoma (LSCC) dataset. The cross sign represents the mean value, while the horizontal line inside the boxplot shows the median value.

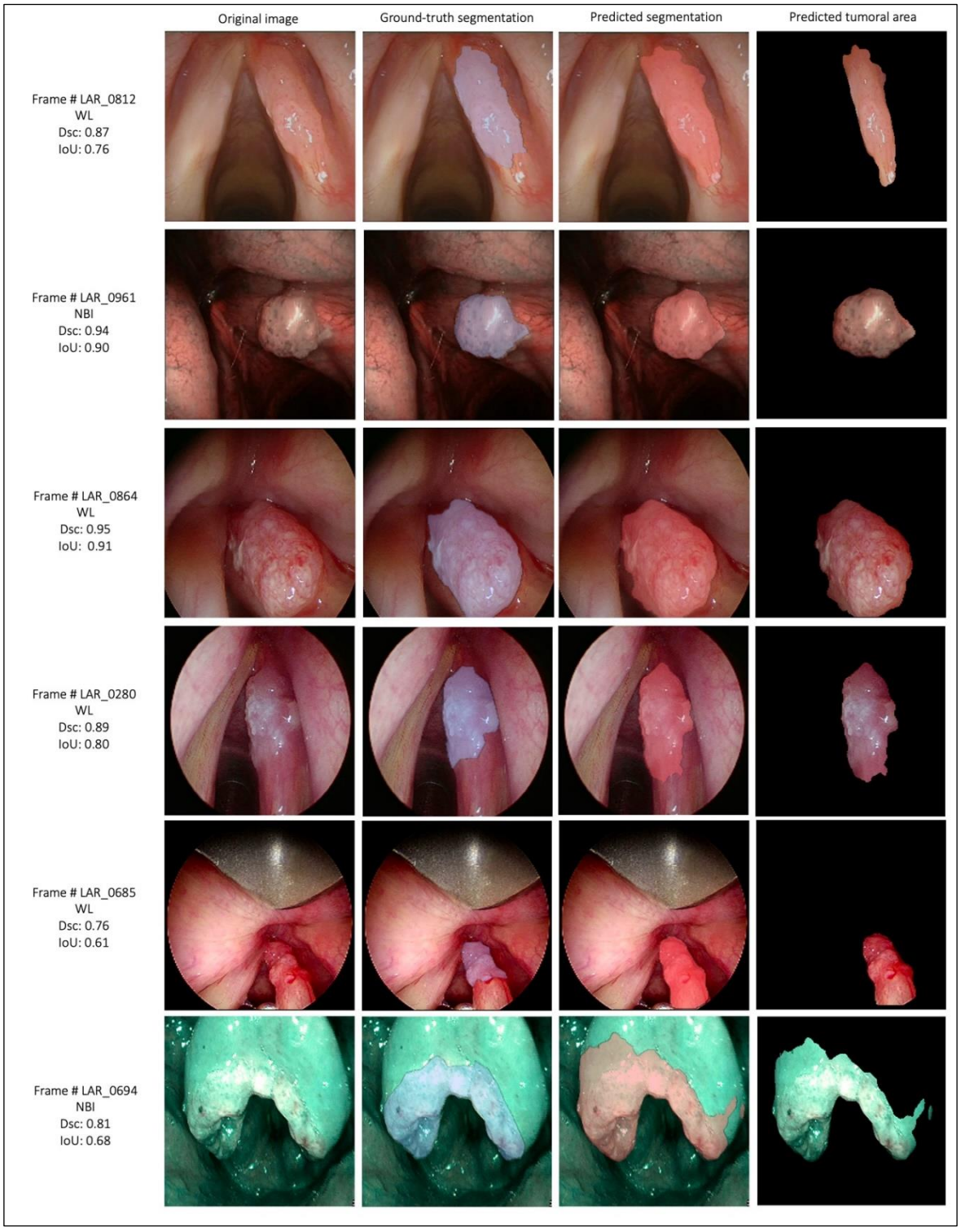

Original image
Ground-truth segmentation
Predicted segmentation
Predicted tumoral area
Frame # LAR_0812
WL
Dsc: 0.87
IoU: 0.76
Frame # LAR_0961
NBI
Dsc: 0.94
IoU: 0.90
Frame # LAR_0864
WL
Dsc: 0.95
IoU: 0.91
Frame # LAR_0280
WL
Dsc: 0.89
IoU: 0.80
Frame # LAR_0685
WL
Dsc: 0.76
IoU: 0.61
Frame # LAR_0694
NBI
Dsc: 0.81
IoU: 0.68

Figure 3. 8 Examples of automatic segmentation results for the laryngeal squamous cell carcinoma dataset using *SegMENT* ensemble. DSC: dice similarity coefficient; IoU: intersection over union; WL: white light; NBI: Narrow Band Imaging.

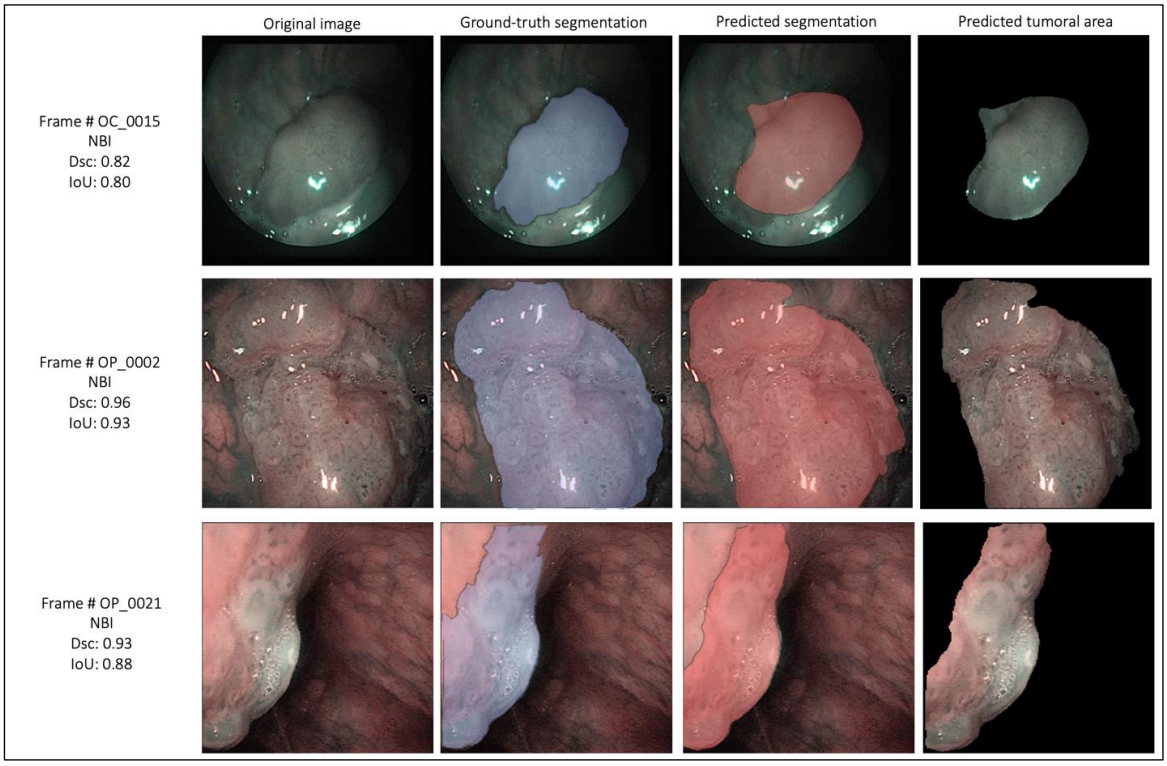


Figure 3. 9 Examples of automatic segmentation results for the oral cavity and oropharyngeal squamous cell carcinoma datasets using *SegMENT* + ensemble in-domain transfer learning. DSC: Dice similarity coefficient; IoU: intersection over union; NBI: Narrow Band Imaging.

3.4.3 Statistical Results

When comparing results among all the models using ANOVA test, the differences were significant for each metric ($p<0.001$). When performing multiple comparisons on IoU, the *SegMENT* ensemble model achieved significantly better results than *UNet*, *ResUNet*, *UNet-MobileNetV2*, and *UNet-ResNet*50 CNNs ($p=0.001$, $p=0.001$, $p=0.001$, and $p=0.03$, respectively). Concerning DSC, the *SegMENT* ensemble model achieved a significantly better result compared to *UNet*, *ResUNet*, and *UNet-MobileNetV2* CNNs ($p=0.001$, $p=0.001$, and $p=0.001$, respectively). Considering recall, the *SegMENT* ensemble model significantly outperformed *UNet*, *ResUNet*, *UNet-VGG16*, *UNet-MobileNetV2*, and *DeepLabv3+*CNNs ($p=0.001$, $p=0.001$, $p=0.02$, $p=0.001$, and $p=0.001$, respectively). For precision and accuracy values, the *SegMENT* ensemble model performed significantly better compared to *UNet* and *ResUNet* ($p=0.001$ and $p=0.001$, respectively).

Table 3. 2 Performance evaluation of models during testing on the oral cavity (OCSCC) and oropharynx squamous cell carcinoma (OPSCC) datasets. The results achieved by *SegMENT* trained only on the specific datasets and by the *SegMENT* ensemble model (i.e., incorporating features learned from the LSCC dataset) are compared to those reported employing *UNet* and *ResNet* in the previous study (Paderno, Piazza, et al., 2021a). The results represent the median scores from all the tests for each metric. Values in bold denote the best results. IoU: intersection over Union. DSC: dice similarity coefficient. TL: transfer learning.

Dataset	Model	Backbone	IoU	DSC	Recall	Precision	Accuracy
Oral Cavity Cancer (OCSCC)	UNet	–	–	0.654	0.755	0.632	0.890
	ResNet with 5×2 blocks	–	–	0.656	0.670	0.708	0.879
	SegMENT	Xception	0.612	**0.759**	0.757	**0.878**	**0.931**
	SegMENT + ensemble TL	Xception	**0.749**	0.598	**0.905**	0.602	0.917
Oropharynx Cancer (OPSCC)	UNet	–	–	0.712	0.815	0.704	0.819
	ResNet with 4×2 blocks	–	–	0.760	0.856	0.772	0.830
	SegMENT	Xception	0.685	0.786	0.767	0.874	0.932
	SegMENT + ensemble TL	Xception	**0.784**	**0.879**	**0.907**	**0.919**	**0.933**

3.5 Summary

This work represents the first multicentric validation of a DL-based semantic segmentation model applied on UADT videoendoscopic images of SCC. The model maintained reliable diagnostic performance analyzing both WL and NBI images from three distinct anatomical subsites. Ensemble strategies and in-domain transfer learning techniques demonstrated the potential to increase segmentation performance. Exploration of new CNNs should be carried out to pursue real-time clinical implementation, while further studies powered by a larger training dataset and larger external validation cohorts are needed before setting up clinical trials.

Chapter 4

Laryngeal cancer detection with deep learning

4.1 Overview

In this chapter real-time detection of LSCC in both WL and NBI video laryngoscopies based on the You-Only-Look-Once (YOLO) DL model presented. The real time performance tested on six video laryngoscopies. The ensemble algorithm presented (YOLOv5s with YOLOv5m—TTA) to achieve the best LSCC detection results on static LSCC frames.

Related Publications

1. **Muhammad Adeel Azam**, Sampieri C, Ioppi A, Benzi P, Giordano GG, De Vecchi M, Campagnari V, Li S, Guastini L, Paderno A, Moccia S. "**Videomics of the upper aero-digestive tract cancer: Deep learning applied to white light and Narrow Band Imaging for automatic segmentation of endoscopic images**". Frontiers in Oncology. 2022 Jun 1; 12:900451. https://doi.org/10.3389/fonc.2022.900451.
2. **Muhammad Adeel Azam**, Baldini C, Li S, Sampieri C, Ioppi A, Moccia S, Peretti G and Mattos LS, "**Real-Time Detection of Laryngeal Squamous Cell Carcinoma via Video Laryngoscopy with WL and NBI using YOLO**," Proceedings of the 11th Conference on New Technologies for Computer/Robot Assisted Surgery (CRAS 2022), pp. 92-93, Napoli, Italy, April 25 – 27, 2022.

4.2 Materials and methods

4.2.1 Data acquisition

A retrospective study was conducted under the approval of the IRCCS Ospedale Policlinico San Martino institutional ethics committee (CER Liguria: 230/2019) following the principles of the Declaration of Helsinki. It included patients treated between 2014 and 2019 at the Unit of Otorhinolaryngology–Head and Neck Surgery at the IRCSS Ospedale

Policlinico San Martino, Genoa, Italy. Selection criteria included histologically proven diagnosis of LSCC and the presence of at least one recorded video of the original laryngoscopy at the time of diagnosis. All patients were first examined through trans nasal video laryngoscopy (HD Video Rhino-laryngoscope Olympus ENFVH—Olympus Medical System Corporation, Tokyo, Japan) in the office before treatment. For those submitted to transoral laryngeal microsurgery, an adjunctive intraoperative evaluation by rigid endoscopy was performed using 0°, 30°, and 70° telescopes coupled to a CCD Camera Head connected to a Visera Elite CLV-S190 light source (Olympus Medical System Corporation). In both settings, a thorough examination was conducted under HD WL video endoscopy and then switching to NBI. For each patient enrolled, the available video laryngoscopes were reviewed, and several frames were extracted. If accessible, one WL frame and one NBI frame were selected from the videos of each patient. These frames were carefully chosen to ensure the selection of good-quality and steady images that provided a clear visualization of the tumor. Afterward, expert physicians segmented and labeled each frame using the VGG Image Annotator (VIA) 2.0 (https://www.robots.ox.ac.uk/vgg/software/via/), an open-source web-based software. The bounding box (BB) marked on the entire tumor with visible surface extension and labeled as LSCC according to the histology: if multiple lesions were present, multiple BB were added. Lastly, six unedited preoperative video laryngoscopes were selected for testing the real-time LSCC detection performance of the trained model.

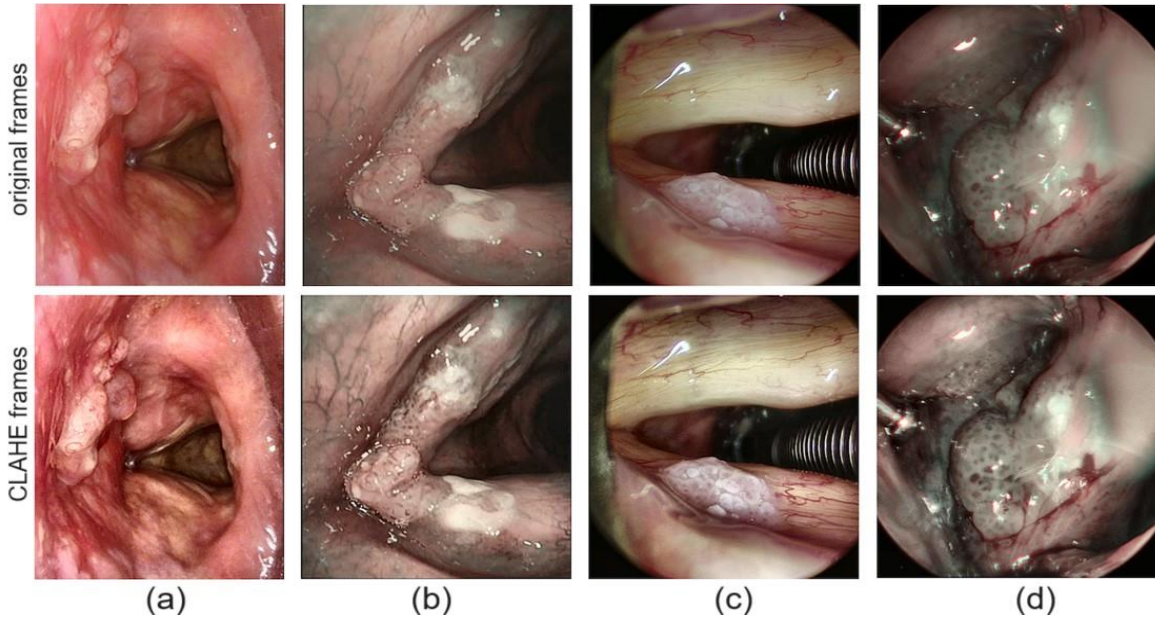


Figure 4. 1 Laryngeal cancer dataset sample images. The upper row contains the original Narrow Band Imaging (NBI) and white light (WL) laryngoscopy images, while in the lower row are reported the same frames after applying (CLAHE). Figure (a) represents an in-office WL view of infrahyoid epiglottic cancer. Figure (b) represents an in-office NBI videoframe of a left vocal fold cancer extending to the anterior commissure. Figure (c) represents an intraoperative WL videoframe of a right vocal fold cancer. While figure (d) represents an intraoperative NBI view of a left vocal fold cancer extending to the bottom of the ventricle.

4.2.2 Preprocessing and Augmentation

Since the contrast under WL and NBI is sometimes suboptimal, to improve the texture detail of input images (and consequently the information given to the DL algorithm during the training phase) contrast limited adaptive histogram equalization (CLAHE)8 methodology (Yadav et al., 2014) was applied to the original dataset of images. Figure 4.1 shows the output WL and NBI images obtained after CLAHE enhancement. Data augmentation is a standard methodology used in DL. It consists in modifying or mixing already existing images to artificially create new training data, which can help increase the reliability of the model. The data augmentation techniques used include image contrast, hue, brightness, saturation, and noise adjustments, geometric transformation such as rotation, randomized scaling, cropping, and flipping of images. Finally, mosaic data augmentation, which superimposes different images to create a new one allowing the model to recognize objects in a variety of settings and sizes, was implemented. Data augmentation was directly performed with the selected DL architecture.

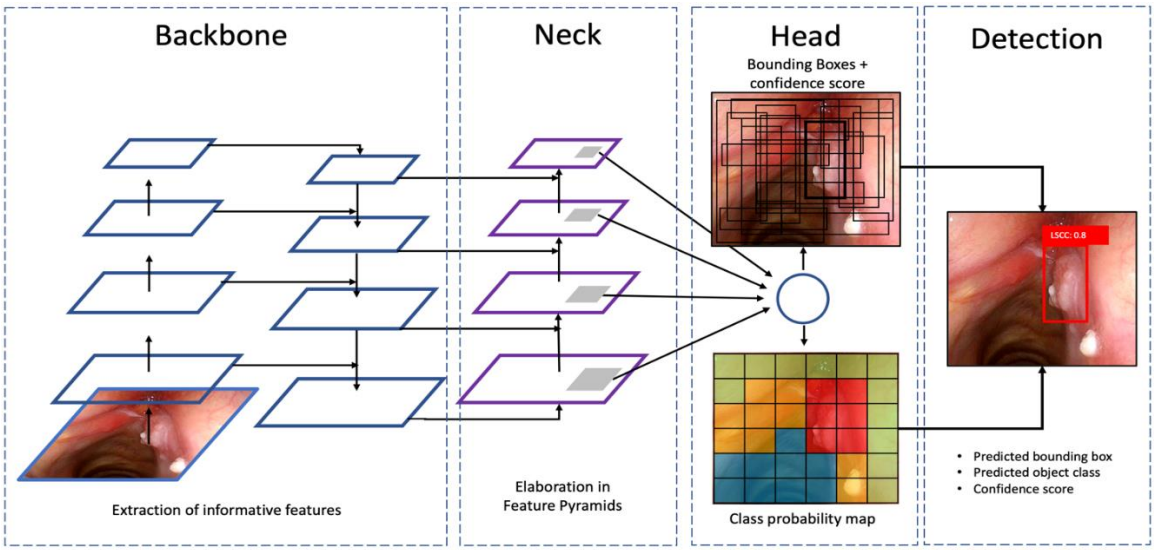


Figure 4. 2 YOLOv5 architecture included backbone for features extraction.

4.2.3 DL model development

The state-of-the-art You Only Look Once (YOLO) DL detection model (Redmon et al., 2016), an open-source software based on CNNs, was chosen as it offers optimal detection accuracy coupled with acceptable computational complexity. YOLO is a single-stage DL object detector, which can identify objects by framing them in a BB and, at the same time, classifying them according to the probability of belonging to a given class. The DL detection architecture is summarized in Figure 4.2. The most recent release of YOLO, at the time of our analysis, was version 5.0 (YOLOv5). The feature extractor that characterizes the backbone of YOLOv5 is CSP Darknet (C. Y. Wang et al., 2020) which is a reduced-size CNN that features high accuracy and superior inference speed. These characteristics led us to choose YOLOv5 over other DL approaches.

Table 4. 1 Performance Evaluation of Various YOLO Models During Training-Validation Phase. mAP@.5 = mean Average Precision with an Intersection over Union threshold of 0.5; TP% = rate of true positive predicted bounding boxes among the total number of bounding boxes predicted; FP% = rate of false-positive predicted bounding boxes among the total number of bounding boxes predicted; TTA = test time augmentation.

Model	Batch Size	Resolution (Pixels)	Model parameter (millions)	Precision	Recall	mAP@.5
YOLOv5s	64	640 x 640	7.06	**0.712**	0.538	**0.576**
YOLOv5m	32	640 x 640	21.0	0.561	0.590	**0.576**
YOLOv5l	16	640 x 640	46.6	0.585	0.615	0.545
YOLOv5x	16	640 x 640	87.3	0.550	**0.628**	0.571
YOLOv5s6	16	1280 x 1280	12.4	0.697	0.474	0.492
YOLOv5m6	8	1280 x 1280	35.5	0.661	0.474	0.506

YOLOv5 comprises four different models, differing in number of parameters, trainable weights size, and computation times. The models range from small to extra-large versions. (YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x), and two released variants are available (YOLOv5s6 and YOLOv5m6). After the training phase with the image dataset, the models were tested to identify the most efficient one in terms of LSCC detection performance and computational time. Finally, the best performing model in this test phase was employed in six unseen and unedited video laryngoscopes to evaluate its automatic LSCC detection performance and recording computation times. The training, validation, and testing hardware environment consisted of a single Tesla T4 GPU with 16 GB of RAM, as well as an Intel(R) Xeon(R) CPU running at 2.20 GHz for DL. YOLOv5 was implemented using a torch 1.8.1 + cu101 CUDA, executed via 13 GB of memory. The hyperparameters used to train all the YOLO models are the following: number of epochs 100; batch size ranging from

8 to 64; input image 640×640 pixels and 1280×1280 pixels; initial and final learning rate l_0=0.01 and l_R=0.2, respectively; momentum 0.337; weight decay 0.0005; threshold level to show BB on video validation analysis 0.4.

Table 4. 2 Performance Evaluation of Various YOLO Models on Testing Phase. mAP@.5 = mean Average Precision with an Intersection over Union threshold of 0.5; TP% = rate of true positive predicted bounding boxes among the total number of bounding boxes predicted; FP% = rate of false-positive predicted bounding boxes among the total number of bounding boxes predicted; TTA = test time augmentation

Model	Trained Weight (Mb)	TP%	FP%	Precision	Recall	mAP@.5
YOLOv5s	14.4	56	44	0.582	0.609	0.592
YOLOv5m	42.5	59	41	0.580	**0.621**	0.554
YOLOv5l	93.7	56	44	0.555	**0.621**	0.564
YOLOv5x	175	54	46	0.542	0.586	0.502
YOLOv5s6	25.2	63	33	0.608	0.609	0.581
YOLOv5m6	71.5	57	43	0.452	0.598	0.527
YOLOv5s—TTA	14.4	78	22	0.662	0.586	0.630
YOLOv5m—TTA	42.5	76	24	**0.677**	0.563	**0.610**
Ensemble YOLOv5s with YOLOv5m—TTA	14.4, 42.5	**82**	**18**	0.664	**0.621**	0.627

4.2.4 Outcome analysis

Statistics and metrics were calculated with the same software environment mentioned above. Precision-Recall curves were used to assess detection performances as described in the literature (Davis & Goadrich, 2006; Marie Schrynemackers, 2013). The true positive (TP) was defined when the Intersection over Union (IoU) of the BBs (Figure 4.3) was greater or equal to 0.5 (Everingham et al., 2009).

$$TP = \frac{\text{BBdetected} \cap \text{BBgroundTruth}}{\text{BBdetected} \cup \text{BBgroundTruth}} \geq 0.5$$

where BB ground Truth is the BB surrounding the segmented area provided by the expert physicians, and BB detected is the BB detected by the algorithm. Results with IoU<0.5 or

duplicated BB were therefore classified as false positive (FP). False negatives (FN) were defined when no detection occurred in presence of a ground truth classified object. A true negative (TN) can be defined as every part of the image where there is no ground truth object and no prediction happens, but this metrics is not useful for the specific case of object detection. Therefore, TN values were ignored. The Recall of a given class, which corresponds to the system's sensitivity, is expressed as:

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

The system's Precision in detecting a given feature, which corresponds to its Positive Predicted Value, is calculated as:

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

For object detection applications, mean average precision (mAP) is a standard metric for the evaluation of performance. The mAP is the area under the Precision-Recall curve and is defined by the equation:

$$mAP = \frac{\Sigma_{q=1}^{Q} AveP(q)}{\mathrm{Q}}$$

where Q is the number of queries in the set, and AveP(q) is the average precision for a given query, q. In our case, as we set the model threshold as 0.5 (at IoU=0.5), mAP@.5 denotes that this value was achieved under the condition of IoU≥0.5.

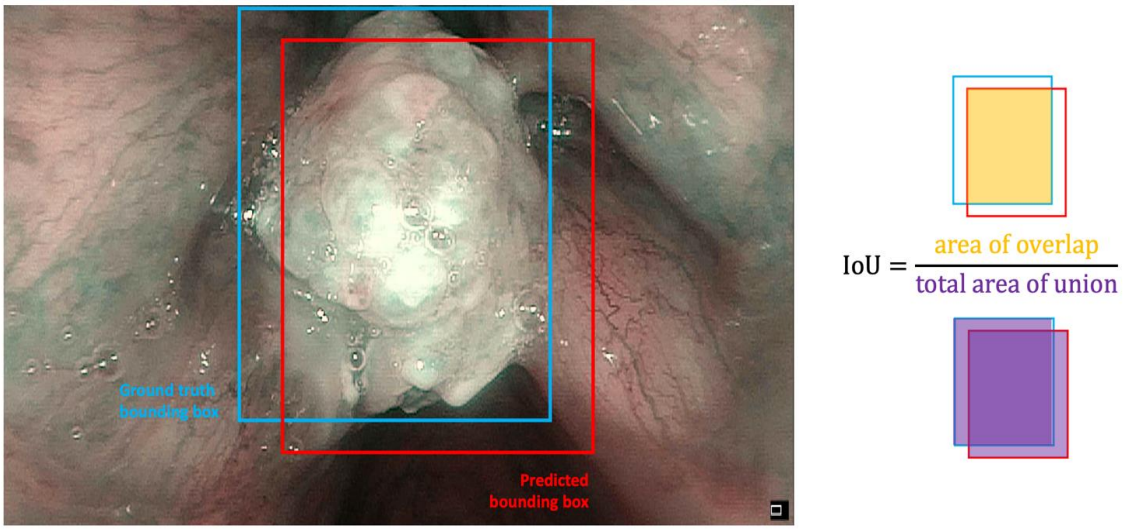


Figure 4. 3 Graphical representation of the Intersection over Union (IoU) calculation on a Narrow Band Imaging (NBI) videoframe. The light blue rectangle represents the ground truth bounding box, while the red rectangle represents the model prediction. The IoU is calculated by dividing the overlap area for the total area of union.

4.3 Results

4.3.1 Dataset

Two hundred and nineteen patients with a mean age of 67.9 years (SD ± 11.8 years) were enrolled. Among these, 196 (89.4%) were males and 23 (10.6%) females. A total of 657 frames representing LSCC were extracted from video laryngoscopes. Of those, 172 were WL in-office images, 146 WL intraoperative, 178 NBI in-office, 128 NBI intraoperative, and 33 images not containing LSCC. The YOLO CNNs models were trained after random distribution of the images into a training set, a validation set, and a test set. The training set consisted of 543 (82.6%) images, of which 256 were WL, 254 NBI, and 33 were from healthy tissues to help reduce the false detection rate of the model. The validation set consisted of 54 (8.2%) images composed of 32 images in WL and 22 in NBI. Finally, the test set consisted of 60 randomly selected images (9.2% of total images), 30 in WL and 30 in NBI. In addition, 6 unedited and unsubmitted videos were used to simulate real-time LSCC detection.

4.3.2 Experimental Results

The comparison of YOLO results during training-validation and testing are shown in Table 4.1, while the evolution of their performance metrics during training are represented in

Figure 4.4. The lightest CNNs models performed very well on our dataset. On the other hand, the more complex models, which typically require larger datasets for training, showed signs of overfitting the data, resulting in a higher number of FP.

Table 4. 3 Characteristics and Computation Times of the Testing Videos After Applying the Ensemble Model (YOLOv5s with YOLOv5m—TTA) for LSCC Detection. WHERE fps = frame per second; LSCC = laryngeal squamous cell carcinoma; Mb = megabytes; TTA = test time augmentation.

Video ID	Size (Mb)	Video Format	Video Resolution	fps	Total Frames	LSCC	Average Computation Time Per Frame (s)
1	23.1	avi	768 x 576	30	1321	Yes	0.027
2	25.3	mp4	778 x 480	25	1448	Yes	0.034
3	34.9	avi	768 x 576	30	1529	Yes	0.025
4	20.6	mp4	778 x 480	25	1421	Yes	0.023
5	27.3	mp4	860 x 480	25	1519	Yes	0.024
6	39.1	mp4	1280 x 720	30	946	No	0.028
Average computation time							0.026

This study also included an assessment of the test time augmentation (TTA) technique. This is a popular strategy used with DL models to increase detection performance.(Yap et al., 2020) TTA consists in performing inference on multiple altered versions of the same image, so that the predictions are subsequently aggregated to obtain higher detection performances. Here, this method was applied only on YOLOv5s and YOLOv5m as these models outperformed the others during training and validation. Finally, to further increase detection rates during testing, an ensemble model was implemented combining the two best-performing models along with the TTA technique (YOLOv5s with YOLOv5m – TTA). As reported in Table 4.2, this ensemble model increased the overall performances in testing compared to the other models, as it localized LSCC very close to the ground truth images: in particular, 82.0% of the predicted bounding box resulted to be TP, while only 18.0% were FN. Examples of BB indicating the location of LSCC according to the ground-truth labeled images and the ensemble model predictions are shown in Figure 4.5.

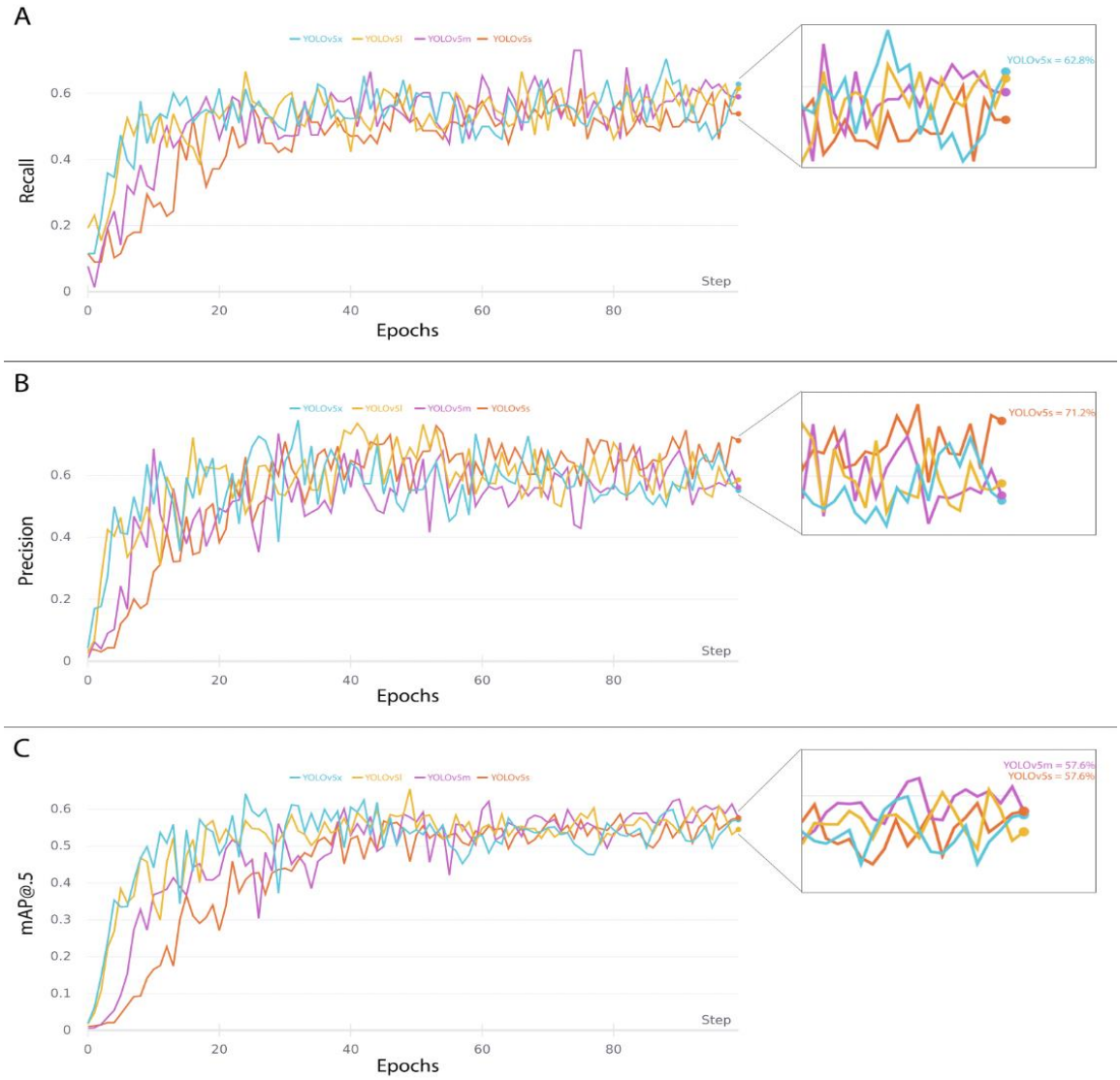


Figure 4. 4 YOLO models performance metrics graphs on training and validation data. Figure A, B, and C respectively represent Recall, Precision, and mAP@0.5 curves while trained up to 100 epochs.

For the testing on video streams, the study focused only on computation times to understand if the DL model could be implemented for real-time detection. The only CNN model used here was the TTA-ensemble model, as it was the best performing model based on the results presented above. The model processed with an average time per frame of 0.026 seconds. The characteristics of each video and the respective DL model computation times are reported in Table 4.3. For illustration purposes, selected frames from the original videos and the corresponding frames processed by the DL model are shown in Figure 4.6.

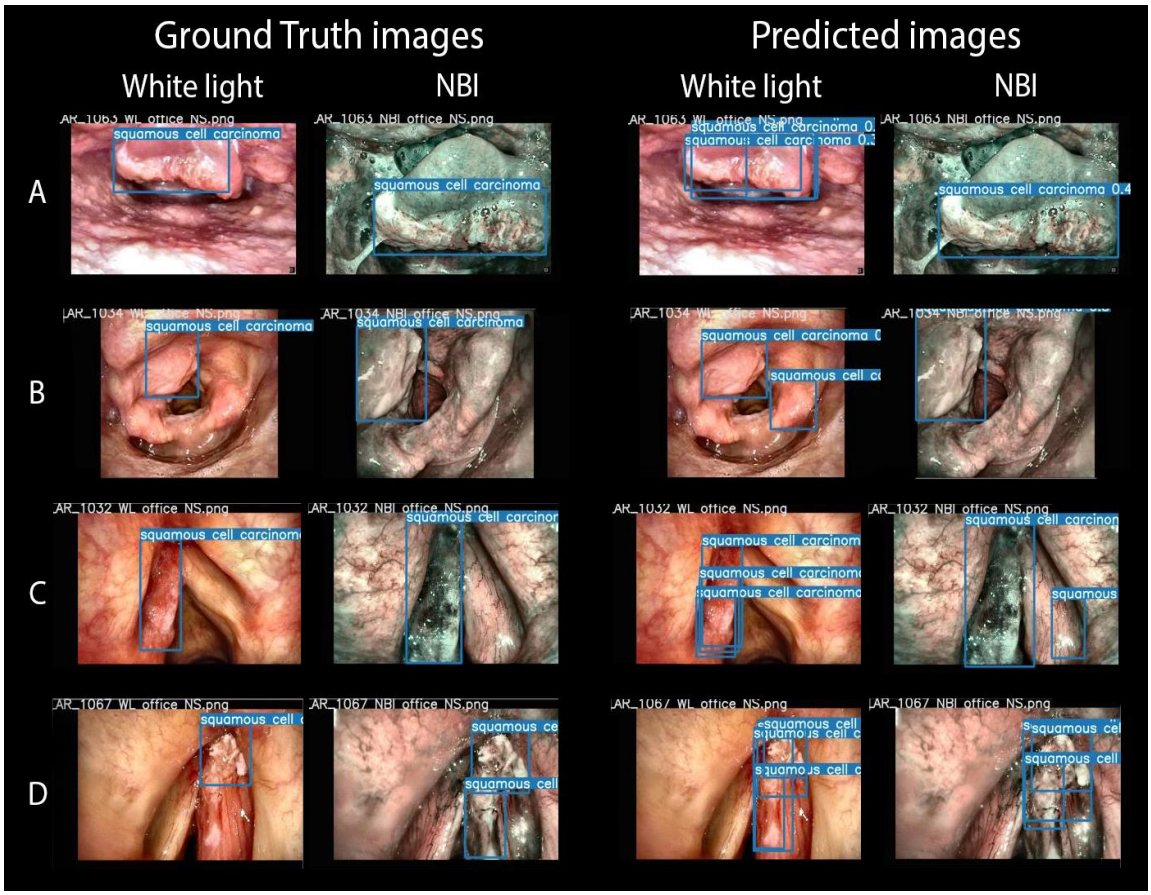


Figure 4. 5 Examples of automatic laryngeal cancer prediction provided by the ensemble model (YOLOv5s with YOLOv5m – TTA). The first two columns on the left contain images with ground truth bounding boxes, while the two columns on the right contain the same images with YOLO-predicted bounding boxes. Case A represents a carcinoma of the infrahyoid epiglottis; case B represents a carcinoma of the infrahyoid epiglottis; case C represents a carcinoma of the left vocal fold; case D represents a carcinoma of the right vocal fold involving the anterior commissure.

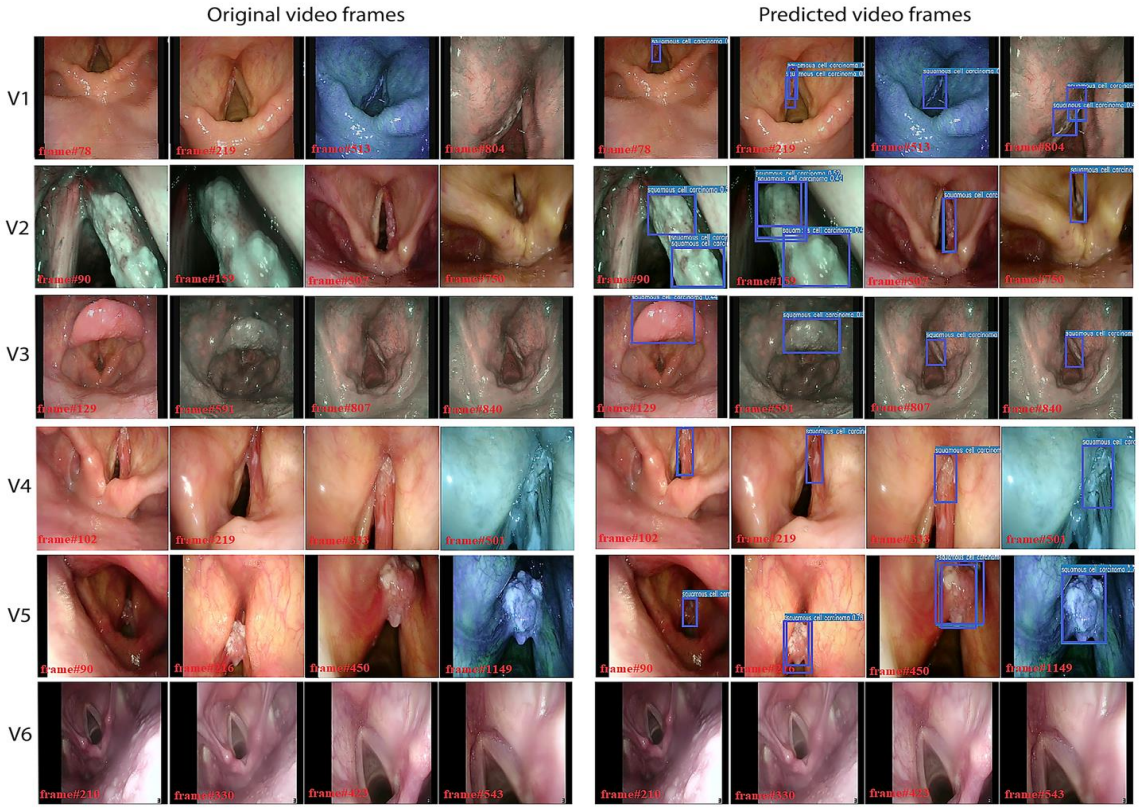


Figure 4. 6 Testing videoframes extracted from 6 video laryngoscopies. Each row represents a different video: the first 4 images of each row are extracted from the original video, while the last 4 images are the same frames extracted after the prediction of the ensemble model (YOLOv5s with YOLOv5m – TTA). The first video (V1) reports a carcinoma of the left vocal fold; the second video (V2) reports a cancer of the right vocal fold; third video (V3) depict a carcinoma affecting the laryngeal surface of the suprahyoid epiglottis and a severe dysplasia of the right vocal fold. Lastly, V4 reports a carcinoma of the right vocal fold extending to the anterior commissure, V5 shows a tumor of the left vocal fold, while V6 reports frames extracted from a healthy larynx.

4.4 Summary

The YOLO ensemble model proved to be efficient in detecting LSCC in video laryngoscopes. The remarkable computational times could represent the keystone in employing the YOLO ensemble model for real-time LSCC detection soon. The availability of NBI images to feed the algorithm represented a pivotal point to reach the detection performances observed, even considering the small training set used in this study. Our model represents a promising algorithm that is expected to reach even higher detection performances with the same short computational time if trained on an expanded image dataset.

Chapter 5

SegMENT-Plus: DL model for LSCC cancer Delineation

5.1 Overview

The WL and NBI endoscopy are widely used to assess the superficial spreading of laryngeal squamous cell carcinoma (LSCC). However, the analysis of images requires a high level of attention and extensive clinical expertise, leading to inter-clinician variability on the assessment of tumor margins. Computer-aided segmentation can automate the identification of LSCC margins, supporting clinicians in this challenging task. In this paper, we present *SegMENT-Plus*, a Deep Learning segmentation convolutional network specifically developed and optimized for accurate delineation of LSCC. *SegMENT-Plus* uses *EfficienstNetB5* as encoder with a new modified atrous spatial pyramid pooling (*m-ASPP*) block that integrates CBAM and squeeze excitation (SE). In this new architecture, CBAM extracts local and global LSCC features from the encoder, while the SE block refines cancer segmentation on each dilated convolution output. *SegMENT-Plus* was trained and evaluated on a multi-center dataset including clinical data from three different hospitals. A total of 4289 annotated laryngeal images from 766 patients were included in this study. The experiments showed that *SegMENT-Plus* achieved a Dice Similarity Coefficient (DSC) between 81.4% and 84.9% and an Intersection over Union (IOU) between 81.8% and 85.7% on the data from the different hospitals, attesting its high performance and generalization capability. The proposed segmentation architecture also demonstrated statistically significant improvement in DSC and IoU compared to other state of the art architectures, showing that this work is a concrete foundation towards a clinical system for the automatic delineation of LSCC margins in endoscopic images.

Related Publication

1. **Muhammad Adeel Azam**, Sampieri C, Ioppi A, Azam MA, Baldini C, Li S, Moccia S, Peretti G, Mattos LS. "**Automatic delineation of laryngeal squamous cell carcinoma during endoscopy**". Biomedical Signal Processing and Control. 2024 Feb 1; 88:105666. https://doi.org/10.1016/j.bspc.2023.105666.

2. **Muhammad Adeel Azam**, Baldini C, Li S, Sampieri C, Ioppi A, Moccia S, Peretti G and Mattos LS, "**Automatic Segmentation of Laryngeal Cancer Using Deep Learning: A Multicentric Dataset Study**," Proceedings of the 12th Conference on New Technologies for Computer/Robot Assisted Surgery (CRAS 2023), ISBN: 9789491857089, pp. 50-51, Paris, France, September 11 – 13, 2023.

5.2 Materials and methods

5.2.1 *SegMENT-Plus* Architecture

SegMENT-Plus is a deep neural network designed to segment LSCC images. Figure 5.1 illustrates its encoder-decoder architecture. The selected encoder (i.e., the network's backbone) is the *EfficientNetB5* (Tan & Le, 2019) which is pre-trained on the large-scale ImageNet database (Jia Deng et al., 2009) to speed-up model convergence and enhance its performance. Three skip connections are selected from the encoder layers to extract features maps: 1) *block6a expand* for high-level features, 2) *block3a_expand* for middle-level features, 3) *block4a_expand_activation* for low-level features. A new *m-ASPP* module is integrated in *SegMENT-Plus* and used as a bridge between the encoder and the decoder. It accepts high level features maps from the encoder layer (*block6a_expands_activation*), and outputs refined features maps. Following the *m-ASPP*, the decoder block of *SegMENT-Plus* accepts three inputs and uses bilinear interpolation up-sampling (Up sampling by 2D) to expand the spatial dimensions of the input feature maps before residual blocks.

In this decoder architecture, our *m-ASPP* module gives high-level features, while other features maps are taken from middle and low-level skip connections from the encoder. These skip connections are concatenated with the output of features maps in the decoder block. The Residual blocks propagate feature map information over many CNN layers and refine feature maps after skipping connection concatenation. The collected (128×128-pixels) feature maps are up-sampled twice to obtain 256×256-pixels feature maps. Finally, the 1×1 convolution with sigmoid activation at the output layer. The input laryngeal image is represented in the *SegMENT-Plus* model encoder $I \in F(C,H,W)$. The "I" is for input image of pixels (256x256), F for feature maps, and C, H, and W stand for the channel, length, and width of the feature map, respectively. The input images passed through pretrained *EfficientNetB5* encoder and extracted three level of features maps. The low-level images features extracted from *block4a_expand_activation* layers expressed in Equation (5.1).

$$I_1 \in F_{Low}((C = 240), \frac{H}{4}, \frac{W}{4}) \tag{5.1}$$

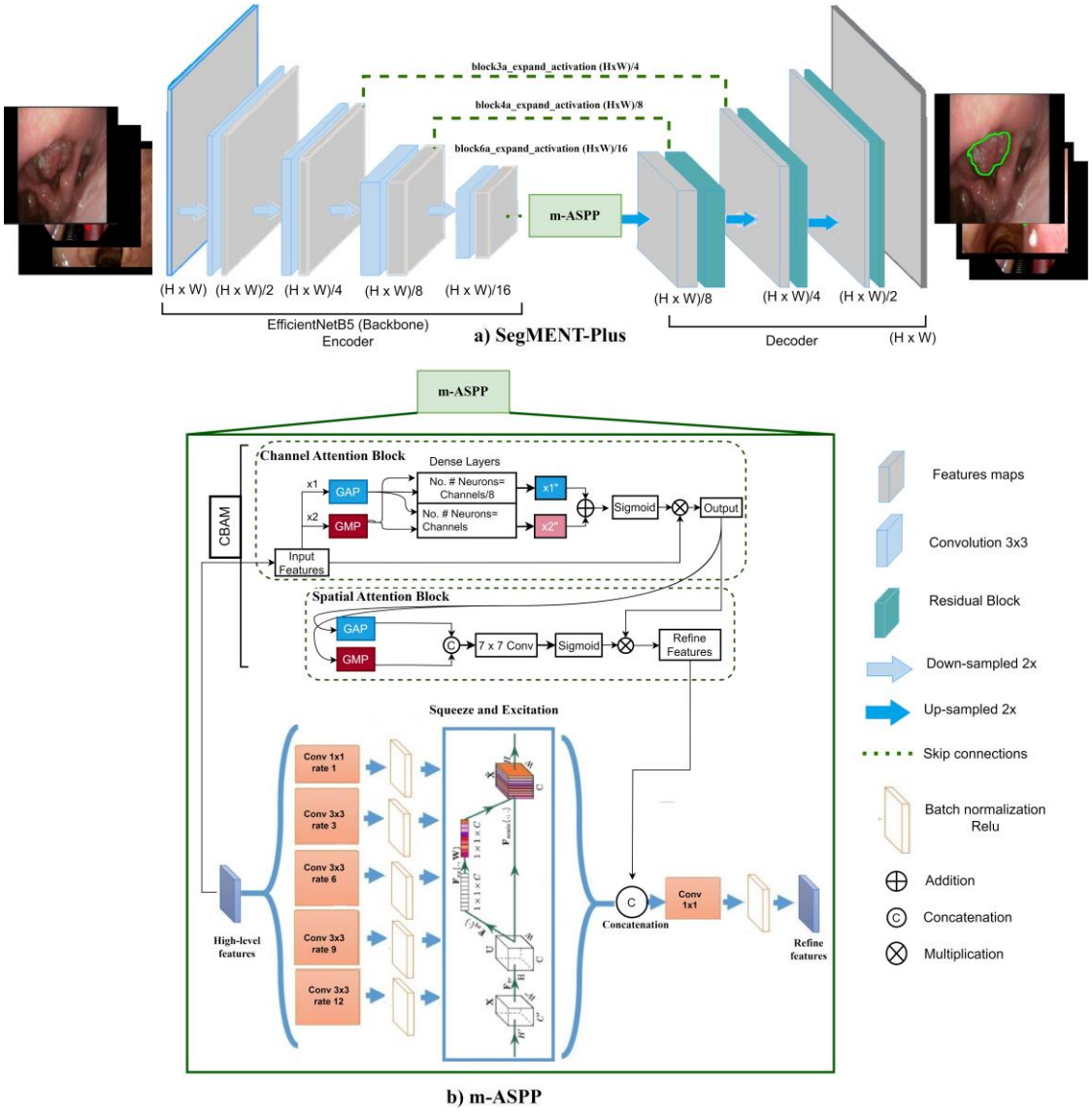


Figure 5. 1 *SegMENT-Plus* block diagram with integrated components. (a) *SegMENT-Plus* with skip connections (in dashed arrows) within the encoder-decoder architecture. (b) The *m-ASPP* module consists of five 3x3 convolutions with different dilated rates (1, 3, 6, 9 and 12), followed SE block and concatenated with the CBAM.

The Flow indicated the low-level features while the output features maps extracted from the *block4a_expand_activation* layers layer consist of 240 channels with pixels (64x64), The middle level features F_{Mid} from *block3a_expand_activation* and high-level features F_{High} from *block6a_expand_activation* mathematically expressed in Equation (5.2) and (5.3) respectively.

$$I_2 \in F_{Mid}((C = 384), \frac{H}{8}, \frac{W}{8}) \tag{5.2}$$

$$I_3 \in F_{High}((C = 1056), \frac{H}{16}, \frac{W}{16}) \quad (5.3)$$

The high-level features from last encoder layer feed to the proposed *m-ASPP* module to get the refined features maps. The Equation (5.3) transformed to refined features and expressed in Equation (5.4). The I* denotes the output images with refined features after *m-ASPP* module operation.

$$I^* = mASPP[I_3 \in F_{refine}((C = 256), \frac{H}{16}, \frac{W}{16})] \quad (5.4)$$

The detailed mathematical operation of *m-ASPP* module is described in section B "The *m-ASPP* module".

5.2.2 The *m-ASPP* module

The concept of the proposed *m-ASPP* module integrated in *SegMENT-Plus* is inherited from the original *ASPP* (L. Chen et al., 2018a). However, here it includes CBAM (Woo et al., 2018) and SE blocks (Hu et al., 2020) to increase the network's receptive field with different Fields of View (FoV), allowing it to take in more context feature information. Within *m-ASPP*, CBAM is used to assist in refining the significance of input feature maps. With its channel and spatial attention modules, CBAM prioritizes global and local information over irrelevant background information (Woo et al., 2018). CBAM replaces the previous global averaging pooling (GAP) process used in traditional *ASPP* architectures, in which the outputs of different dilated convolution blocks are directly concatenated with GAP without the refinement of global feature information. This traditional approach leads to limitations regarding reduced attention to local features and difficulties in retrieving relevant information (Zafar et al., 2022). CBAM solves these issues with its attention modules (L.-C. Chen et al., 2017). Figure 5.2 shows details of the CBAM attention modules. As introduced by Woo et al. (Woo et al., 2018), this module adds a Channel Attention Block on top of spatial attention to improve important channels and significant spatial regions. CBAM's Channel Attention Block includes GAP and Max pooling on feature map channels. Channels represent objects differently (i.e., tumor and background); attention is implemented as weights for each channel. Pooling operations and higher weights help GAP and Max pooling focus on tumor regions. Channel block only considers cross-channel dependencies and ignores spatial information, while spatial block considers local and global region information spatially. Together, channels and spatial information yield adaptive feature information. In parallel to information retrieval using CBAM, the *m-ASPP* extracts multiscale features from the high-level features maps using parallel dilated convolutions

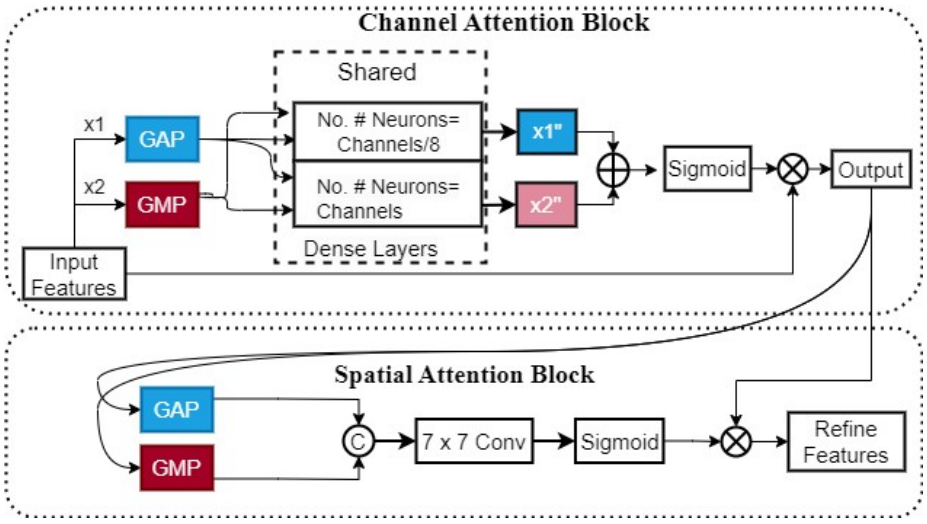


Figure 5. 2 CBAM architecture block diagram. The 'GAP' represents global averaging pooling while GMP denotes global max pooling.

followed by SE blocks. The goal here is to reweight each feature of the map channel based on its relevance to LSCC characteristics. High-level input features (F_{High}) are first processed by the CBAM module in *m-ASPP*. The features pass through one Dimensional (1D) channel attention mechanism [$A_C \in (C, 1, 1)$] to improve important channels. The result is then improved by a 2D spatial attention step [$A_S \in (1, H, W)$] that looks at spatial relationships of features maps. The A_c and A_s are channel and spatial attention modules, respectively. This CBAM-based strategy improves the characteristics through both channel and spatial attention, as shown in Equations (5.5) and (5.6). Where I_3 is the output of F_{High} feature maps from last encoder layer, I' stands for attention channel output features maps, and I" stands for attention module output features maps, respectively. The symbol $\otimes$ denotes the elementwise multiplication operation.

$$I' = [A_C \in (C,1,1)](I_3) \otimes I_3) \tag{5.5}$$

$$I^{"} = [A_S \in (1,H,W)](I') \otimes I') \tag{5.6}$$

Then I_3 passed through various dilated convolutions to capture various ranges of receptive filed of F_{High} features maps shown in Equation (5.7). The dilated convolutions contribute to this by allowing multiple FoV on the feature maps and the extraction of more refined information than a standard *ASPP* module. Here, five dilated 3×3 convolution blocks with different dilation rates (1, 3, 6, 9 and 12) generate the desired feature maps receptive fields (Chollet, 2017). The output of each dilated convolution is then elementwise multiplied with SE blocks to further enhance the relevant features represented in Equation (5.8). Finally, the

output of CBAM (I") and $[y1^{"}, y2^{"}, y3^{"}, y4^{"}, y5^{"}]$ are concatenated to get combined informative features maps (y) represented in Equation (5.9).

$$\begin{matrix} y1 \\ y2 \\ y3 \\ y4 \\ y4 \\ y5 \end{matrix} = \begin{cases} [(I_3)(conv(1,1), d = 1) \in (C, H, W)] \\ [(I_3)\ (conv(3,3), d = 3) \in (C, H, W)] \\ [(I_3)\ (conv(3,3), d = 6) \in (C, H, W)] \\ [(I_3)\ (conv(3,3), d = 9) \in (C\ , H, W)] \\ [(I_3)\ (conv(3,3), d = 12) \in (C\ , H, W)] \end{cases} \tag{5.7}$$

$$\begin{matrix} y1^{"} \\ y2^{"} \\ y3^{"} \\ y4^{"} \\ y5^{"} \end{matrix} = \begin{cases} [y1 \otimes Se] \\ [y2 \otimes Se] \\ [y3 \otimes Se] \\ [y4 \otimes Se] \\ [y5 \otimes Se] \end{cases} \tag{5.8}$$

$$y = Concat[I", y1^{"}, y2^{"}, y3^{"}, y4^{"}, y5^{"}] \tag{5.9}$$

The *m-ASPP* module finally completes its operations with the merging of the information provided by the CBAM and SE blocks. For this, the output of each SE block is concatenated with the CBAM output. Figure 5.3 compares *m-ASPP* with standard *ASPP*, Dense *ASPP*, and *WASPP*. The *m-ASPP* replaces averaging pooling in *ASPP* with a CBAM block integrating channel and spatial attention mechanisms. Then, the five dilated convolution blocks are followed by SE blocks to generate feature maps receptive fields. Overall, this modified *ASPP* architecture includes new functionalities designed to overcome the limitations of standard *ASPP* for LSCC segmentation.

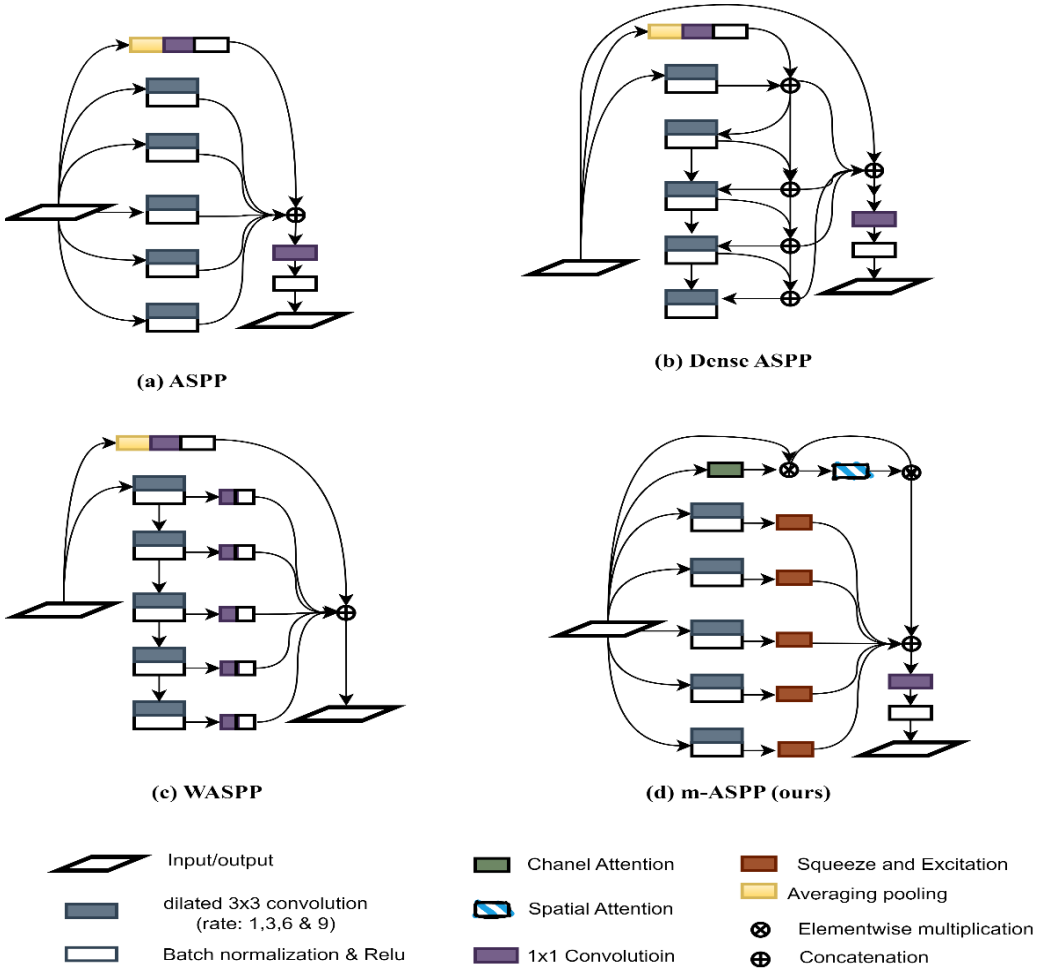


Figure 5. 3 *Comparison of various Atrous Spatial pyramid pooling modules.* (a) *ASPP* (b) Dense *ASPP,* (c) *WASPP* and (d) Proposed *m-ASPP.*

5.2.3 Datasets

This study was based on video laryngoscopy (VL) data acquired from three different hospitals, which were used to build three independent datasets as described below. The dataset contains images in both WL and NBI modalities. The WL images include the entire color spectrum of tissue interaction. Most colors are absorbed, except for reflections from distinct tumor colors. WL imaging is cost-effective and produces high-quality images of tissue. The NBI, on the other hand, increases vascular and tissue visibility, assisting in early lesion detection and diagnostic accuracy. The wavelengths used in NBI are blue and green, which are absorbed by hemoglobin in blood vessels. The submucosal region reflects green light while mucosal capillaries absorb blue light, giving NBI more contrast between blood vessels and surrounding tissue than WL. We combined the two imaging modalities in the framework of our experimental endeavors to increase the generalizability of DL models for tumor delineation across diverse imaging modalities.

Genoa LSCC dataset: Endoscopic videos were selected from the database of the Otorhinolaryngology Unit of the San Martino Hospital (Genoa, Italy) considering endoscopies performed between 2014 and 2020. This dataset contains videos obtained in both the clinic and operating room. All videos included endoscopic images of the patient's larynx under WL and NBI. A total of 557 patients were included in this dataset. Once the VLs were selected, four expert physicians retrieved five WL and NBI frames with squamous cell carcinomas from each video. Frame selection prioritized informative images with low artifacts and blur. Subsequently, three expert laryngologists annotated the selected frames using the CVAT annotation software (*CVAT*, n.d.). This produced pixel-by-pixel masks of the tumor regions, labeled as LSCC, and considered as ground-truth segmentations for the purposes of this study. In total, 3933 annotated images were included in the Genoa dataset. *SegMENT-Plus* was developed based on the Genoa dataset using a patient-level split ratio of 80:10:10 for the training, testing, and validation sets. As a result, the training set included 1678 WL images (1082 in-clinic and 596 intraoperative) and 1468 NBI images (1027 in-clinic and 441 intraoperative). The validation set included 206 WL images (126 in-clinic and 80 intraoperative) and 187 NBI images (126 in-clinic and 61 intraoperative). Finally, the testing set included 208 WL images (129 in-clinic and 79 intraoperative) and 186 NBI images (122 in-clinic and 64 intraoperative). In total, the dataset included 3933 images, 3146 used for training, 393 for validation, and 393 for testing. Table 5.1 summarizes the data distribution of the Genoa LSCC dataset.

Table 5. 1 Genoa LSCC dataset distribution of images in subsets with type of examination and imaging technology used for training, validation, and testing.

Dataset Distribution	Type of examination	No.	Imaging technology	No.	No.
Training	In-clinic	1082	WL	1678	3146
	Intraoperative	596			
	In-clinic	1027	NBI	1468	
	Intraoperative	441			
Validation	In-clinic	126	WL	206	393
	Intraoperative	80			
	In-clinic	126	NBI	187	
	Intraoperative	61			
Testing	In-clinic	129	WL	208	394
	Intraoperative	79			
	In-clinic	122	NBI	186	
	Intraoperative	64			
Total	In-clinic	1337	WL	2092	3933
	Intraoperative	755			
	In-clinic	1275	NBI	1841	
	Intraoperative	566			

Brescia LSCC dataset: The Unit of Otolaryngology of the Spedali Civili at the University of Brescia (Italy) provided this first external cohort dataset. It was comprised of 200 WL and NBI images annotated by the same clinicians using the same methodology as done for building the Genoa dataset. This dataset was used only for evaluating the performance of *SegMENT-Plus*.

Seoul LSCC dataset: This second external cohort dataset was used exclusively for testing *SegMENT-Plus*. It consisted of 156 WL and NBI images provided by the Unit of Otolaryngology of Severance Hospital, Yonsei University (Seoul, South Korea), which were segmented by the same clinicians as described previously.

5.2.4 Experimental Analysis

Training Setting: *SegMENT-Plus* and other baseline models were implemented in Keras and Tensorflow. All models were trained for 30 epochs with 8 batches in all experiments. The initial learning rate was set to 0.001, but it decreased with multiplicative factor of 0.1 if the loss did not decrease after three epochs. All experiments were performed on a Dual Intel Xeon Gold 5222 3.8GHz CPU, 128 GB RAM, Nvidia Quadro RTX 8000 GPU with 48GB of memory. Before training the models, the dataset images were resized to 256 by 256 pixels to reduce computational load. In addition, three Test Time Augmentations (TTAs) with 90˚, 180˚, and 270˚ rotations were implemented. The Dice loss (R. Zhao et al., 2020) was used to deal class imbalance in our dataset (since cancer pixels occupy fewer segments than background). Dice loss was trained using the Adam optimizer.

Performance metrics: Segmentation performance was evaluated by comparing the predicted segmentations with the manual annotations (ground-truth). Standard evaluation metrics for semantic segmentation (Azam et al., 2022a) were used, including Accuracy, Precision, Recall, Dice Coefficient (DSC) and Intersection over Union (IoU). True Positive (TP) represents a tumor pixel correctly identified by the model. True Negative (TN) is a correctly identified non-tumor pixel. When the model detects a tumor in a non-tumor region, this is considered a false positive (FP). False Negative (FN) occurs when the model misses a tumor. The mathematical expressions of these metrics were already written in Chapter 3. To find significant differences in continuous variable distributions, the analysis of variance (ANOVA) test was used. Tukey's multiple comparisons test was used for pairwise difference comparison. A p-value below 0.05 was considered significant. This study used Python 3.9 with scipy.stat and statmodels 0.13.2.

5.2.5 Ablation Study

An ablation study was conducted to examine the effectiveness of the *SegMENT-Plus* architecture and validate the individual contributions of its network components. The study builds upon the research described in (Artacho & Savakis, 2021), which evaluated the efficacy of these components in a different segmentation model. In the first experiment, *SegMENT-Plus* was evaluated using various backbone models as encoders while keeping the rest of the segmentation architecture, hyper-parameters, and modules the same. This experiment considered only the Genoa LSCC dataset. The backbones tested include: Resnet-50 (K. He et al., 2016), Xception (Chollet, 2017), EfficientNetB5, B6, B7 (Tan & Le, 2019) and V2B3 (Tan & Le, 2021). In the first ablation study experiments different backbones of segmentation network changed while same standard *ASPP* module used. The proposed *m-ASPP* did not integrated into all encoder backbones, it was only used for second ablation study with *EfficientNetB5* encoder only. The standard *ASPP* module used an equal number of filters from the 16x16 pixel maps of the final encoder layer for all encoders. As various backbones are used in the first ablation study, the number of filters is matched with each last encoder layer resulting in diverse feature map outputs. In the second ablation study experiment we changed the *ASPP* modules while keeping the encoder (*EfficientNetB5*) constant. In these experiments, we used a single encoder, therefore all *ASPP* modules had a fixed set 256 number of filters. A fixed filter size was chosen since the final encoder layer consistently outputs a predetermined number of feature maps. The *m-ASPP* module performance compared with other state of art *ASPP* modules in later ablation study experiments. The *SegMENT-Plus* considers five different *ASPP* module conditions: without *ASPP*, standard *ASPP* (L. Chen et al., 2018a), Dense *ASPP* (M. Yang, n.d.), *WASPP* (Artacho & Savakis, 2019, 2021) and *m-ASPP.* In this case, the backbone (EfficientNetB5) and the rest of segmentation architecture were kept constant. This experiment used data only from the Genoa LSCC dataset.

5.2.6 Comparison with baseline architectures

A study was performed to compare *SegMENT-Plus* and *Segment-Plus* + TTA with five CNN models: *UNET* (Ronneberger O, Fischer P, 2015), *Res-UNET* (Khanna et al., 2020), *Double-UNET* (Jha et al., 2020a), *DeepLabv3Plus* (L. Chen et al., 2018b), and *SegMENT* (Azam et al., 2022a). *UNET*'s decoders segment images using encoder-learned deep features. The decoder can propagate dense feature maps to relevant layers using *UNET*'s encoder skip connections. *Res-UNET* is a *UNET*-based segmentation model that uses residual units instead of neural units for higher performance with fewer parameters. *Double-UNET* uses an architecture based on two *UNET*'s, a *VGG-19* preliminary encoder, *ASPP* and SE blocks. *DeepLabV3Plus* uses dilated separable convolutions and pyramid pooling in a U-shaped

architecture to increase processing speed with lower loss. *SegMENT* uses *Xception* backbone and three *ASPP* blocks for the multiscale feature maps. All these models were trained on Genoa LSCC dataset, and their performances are compared herein. The comparative experiment in our study evolves in two aspects. First, we used segmentation models (*UNET, ResUNET, DeepLabV3Plus*, and *SegMENT*) in our investigation. These segmentation techniques previously demonstrated exceptional performance in 2021(Paderno, Piazza, et al., 2021a) and 2022 (Azam et al., 2022b) on laryngeal WL and NBI image datasets. Furthermore, the *DoubleUNET* (Jha et al., 2020b)model, developed in 2020 for medical polyp segmentation, is incorporated, broadening the area of our comparison research. By using these well-known models as a baseline, we can highlight the distinct contributions of our recommended adjustments in identifying laryngeal tumor delineation. Secondly, our study extensively evaluates *ASPP* modules in the segmentation network for laryngeal tumor delineation. Notably, *DeepLabV3Plus* uses one *ASPP* module while in the *SegMENT* architecture three *ASPP* modules integrated for better laryngeal segmentation. The proposed *SegMENT-Plus* model enhances *ASPP* module with a *m-ASPP*. This enhancement is explained by a careful comparison with other Atrous modules already in use, such as *ASPP*, *WASPP*, and Dense *ASPP*. The inclusion of these baseline models is crucial since it offers a baseline for our comprehensive comparative study of laryngeal segmentation. Kernel density estimation (KDE) curves were used to visualize the distribution of the DSC and IoU metrics across the three testing datasets. For this purpose, the scikit-learn library for Python (Heer, 2021) was used. The estimated probability density functions were plotted as curves, with the x-axis representing DSC and IoU values and the y-axis representing probability density. A Gaussian Kernel was used to estimate the kernel density curves for IoU and DSC. The KDE curves allow an intuitive visual comparison of the performance of the different segmentation models. For example, curves with small spread and peak towards higher values of the metric considered indicated models with superior segmentation performance.

5.3 Results

The *SegMENT-Plus* model was trained on a significantly larger dataset compared to our previous study (Azam et al., 2022a). Expert clinicians annotated 4289 WL and NBI endoscopic LSCC images. External validation cohorts included 356 images from two clinical centers, overcoming the monocentric dataset limitation of prior similar studies. All segmentation models in our training studies converged at different rates during the duration of 30 epochs. The proposed *SegMENT-Plus* model performs better in terms of rate of convergence, stability, and minimal loss within 20 epochs than other state-of-the-art models. *SegMENT* and *DeepLabV3Plus* achieve slightly slow convergence at the 14th epoch and no loss reduction until the 22nd epoch. However, *UNET* and *ResUNET* are unable to converge after the 20th epoch, which results in no convergence further. *Double-UNET* exhibits the slowest convergence, taking 25 epochs to stabilize with constant loss curve. In validation

loss the convergence trend of each segmentation model follows the similar slops as of training plots. The detailed convergence curves of train and valid loss graphs are shown in Figure 5.4.

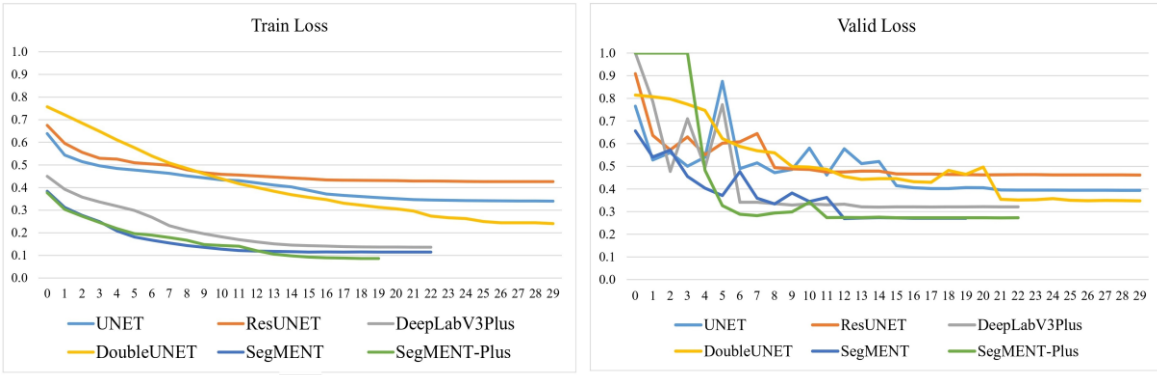


Figure 5. 4 The loss convergence plot of each segmentation model, figure (a) shows the training convergence plot while figure (b) shows the validation convergence plots.

5.3.1 Ablation study results

Results from the ablation study regarding *SegMENT-Plus* with different backbone models are shown in Table 5.2. The results show that *ResNet-50* and *EfficientNetV2B3* perform worst on IOU, DSC, and Precision. *EfficinetNetB5* and EfficientNetB6 score highest in all quantitative criteria. EfficientNetB5 outperforms *EfficientNetB6* in DSC, Precision, and Accuracy. Table 5.3 summarizes the results of the ablation study regarding the *ASPP* module. *SegMENT-Plus* with an *ASPP* module showed a performance improvement of 0.7% in DSC and 0.4% in IoU when compared to the architecture without *ASPP*. Using Dense *ASPP* instead, the performance increase was 1.6% in DSC and 0.8% in IoU with respect to the architecture without *ASPP*. The use of *WASPP* practically did not change the performance without *ASPP*: only a slight increase of 0.1% in DSC was recorded. Finally, the proposed *m-ASPP* module showed the largest improvement in performance, providing a 3.0% increase in DSC and a 1.7% increase in IoU. The CBAM module significantly increases network sensitivity of laryngeal tumor regions by utilizing both channel and spatial attention approaches. The combination of the SE and CBAM components inside the *m-ASPP* module significantly reduces the prevalence of inattention in non-tumor regions. Figure 5.5 clearly shows the Grad Cam visualization of tumor region with each component of *m-ASPP* module. In the initial deployment of our *m-ASPP* architecture without SE and CBAM integration, the segmentation model recognizes tumor locations with restricted tumor area. The addition of SE blocks, on the other hand, broadens the range of focus to include tumor regions as well as non-tumor attention areas. To avoid this problem, the GAP operation is

replaced by the integration of CBAM to effectively reduce unneeded attention. As a result, the network's ability to precisely identify tumor areas only improves. Overall, the *m-ASPP* module's design successfully addresses the shortcomings of conventional and other *ASPP* modules, resulting in a significant improvement in laryngeal tumor segmentation.

Table 5. 2 *SegMENT-Plus* performance on genoa LSCC test dataset with different backbone models*

Backbone	Par. (m)	DSC	IoU	Rec	Prc	Acc	FPS
Resnet-50	17.8	0.690 (0.87−0.59)	0.750 (0.86−0.66)	0.746 (0.95−0.64)	0.730 (0.93−0.61)	0.940 (0.98−0.92)	24.5
Xception	24.5	0.743 (0.89−0.67)	0.789 (0.89−0.71)	0.767 (0.95−0.70)	0.783 (0.95−0.68)	0.951 (0.99−0.94)	20.1
EfficientNetB5	15.1	**0.745 (0.90−0.67)**	0.788 (0.90−0.72)	0.733 (0.95−0.68)	**0.784 (0.95−0.70)**	**0.955 (0.99−0.94)**	26.2
EfficientNetB6	18.7	0.744 (0.89−0.68)	**0.790 (0.89−0.71)**	**0.789 (0.96−0.74)**	0.764 (0.94−0.68)	0.952 (0.99−0.94)	22.7
EfficientNetB7	24.3	0.712 (0.88−0.62)	0.769 (0.88−0.70)	0.728 (0.94−0.61)	0.774 (0.95−0.68)	0.948 (0.99−0.94)	18.8
EfficientNetV2B3	10.5	0.688 (0.87−0.58)	0.754 (0.87−0.67)	0.711 (0.94−0.59)	0.739 (0.93−0.63)	0.943 (0.98−0.93)	**28.8**

*The values are presented as median with IQR (3rd quartile−1st quartile). Values in bold indicate the highest values. 'Par': number of parameters ('m': million). 'DSC': Dice Coefficient. 'IoU': Intersection over Union. 'Rec': Recall. 'Prc': Precision. 'Acc': Accuracy. 'FPS: frames per second.

5.3.2 Results from comparison with baseline architectures

1) Performance on the Genoa LSCC dataset

The *SegMENT-Plus* model achieved exceptional results on the Genoa LSCC dataset, with DSC, IoU, and Rec scores above 80%. The addition of TTA to the model resulted in a modest improvement in Prc. and Acc. Furthermore, *SegMENT-Plus* outperformed our previous *SegMENT* model (which integrated standard *ASPP* module shown in Table 5.3). The new model showed an improvement of 2.3% in DSC and 1.3% in IoU with respect to the previous one, which had already demonstrated superiority with respect to the baseline models (Azam et al., 2022a). This can also be seen in the box plots in Figure 5. 6, which shows the DSC and IoU results of different segmentation models obtained on the Genoa LSCC testing dataset. Among the compared models, *UNET* and *Res-UNET* demonstrated the poorest segmentation performance. Table III shows that the *SegMENT-Plus* model without TTA, with EfficientNetB5 backbone achieved a frame rate of 26.2 FPS in terms of processing speed. This model is 2.9 times faster than the *SegMENT-Plus*+TTA model, which has a 9.1 FPS, and 12.5 times faster than our previous *SegMENT* model (Azam et al., 2022a).

Table 5. 3 *SegMENT-Plus* performance with ASPP modules on genoa LSCC testing dataset*

Element	Par. (m)	DSC	I* w/o ASPP	I* ASPP	IoU	I* w/o ASPP	I* ASPP	Rec	Prc	Acc
w/o ASPP	10.1	0.737 (0.89–0.68)			0.784 (0.89–0.72)			0.782 (0.96–0.69)	0.769 (0.94–0.67)	0.950 (0.99–0.94)
ASPP	17.4	0.742 (0.89–0.66)	0.7%		0.787 (0.89–0.72)	0.4%		0.792 (0.95–0.72)	0.766 (0.93–0.68)	0.952 (0.98–0.95)
Dense ASPP	58.1	0.749 (0.90–0.67)	1.6%		0.790 (0.89–0.72)	0.8%		0.788 (0.95–0.72)	0.780 (0.94–0.67)	0.953 (0.99–0.94)
Waterfall ASPP	13.7	0.738 (0.89–0.66)	0.1%		0.784 (0.89–0.71)	0.0%		0.767 (0.95–0.66)	**0.782** **(0.95–0.68)**	0.951 (0.99–0.94)
m-ASPP (proposed)	17.6	**0.759** **(0.83–0.76)**	**3.0%**	**2.3%**	**0.797** **(0.84–0.80)**	**1.7%**	**1.3%**	**0.804** **(0.88–0.80)**	0.779 (0.86–0.78)	**0.954** **(0.97–0.95)**

The values are presented as median with IQR (3rd quartile−1st quartile). Values in bold indicate the highest values. 'Par': number of parameters ('m': million). 'DSC': Dice Coefficient. 'IoU': Intersection over Union. 'Rec': Recall. 'Prc': Precision, and 'Acc': Accuracy. I w/o ASPP shows each model's improvement over without ASPP while I* ASPP measures the percentage improvement of the *m-ASPP* module over the *ASPP* module.

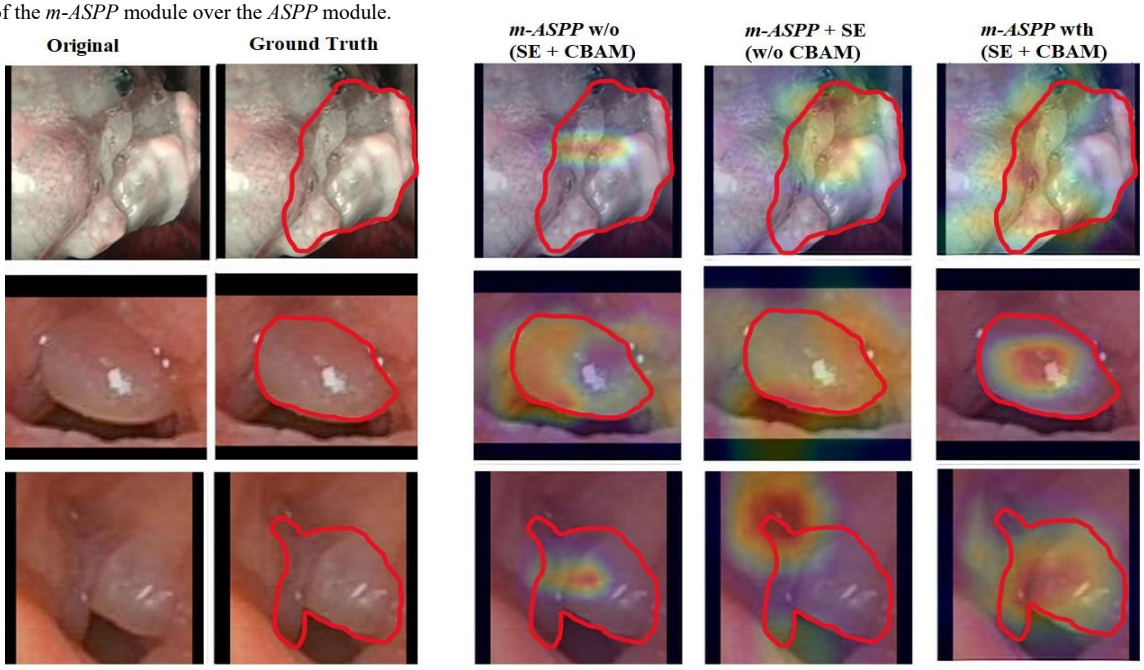


Figure 5. 5 The m-ASPP grad-Cam visualization sample images from Genova LSCC dataset. The grad-cam in third column shows m-ASPP module without SE and CBAM and only used GAP instead of CBAM. The fourth column indicates the grad-cam of m-*ASPP* with five parallel SE attention blocks. The fifth column indicates the grad-cam of *m-ASPP* module with SE and CBAM.

2) Performance on external cohorts

When tested on the Seoul LSCC dataset, *SegMENT-Plus*+TTA demonstrated superior performance compared to all other models. Res-UNET, on the other hand, gave the poorest performance. Notably, *SegMENT-Plus*+TTA exhibited substantial improvements in DSC and IoU scores of 8.9% and 6.8%, respectively, when compared to our previous *SegMENT*

model. In this dataset, *SegMENT-Plus* and *SegMENT-Plus*+TTA achieved processing speeds of 24.5 FPS and 8.4 FPS, respectively. Similarly, *SegMENT-Plus*+TTA also outperformed *SegMENT* on the Brescia LSCC dataset, achieving improvements of 2.3% and 1.9% in DSC and IoU scores, respectively. In this dataset, *SegMENT-Plus* achieved 26.2 FPS, while *SegMENT-Plus*+TTA reached 8.8 FPS. Figures 5.7 and 5.8 show the DSC and IoU boxplots results of different segmentation models obtained on the external cohort's dataset.

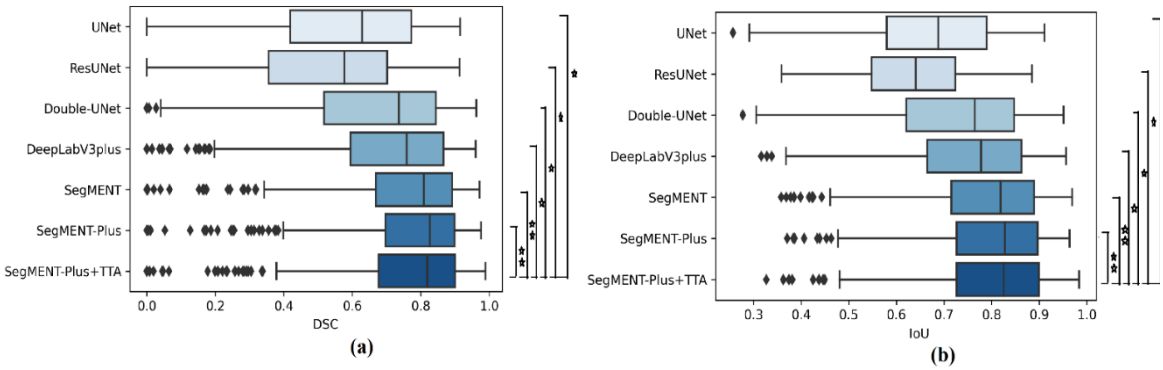


Figure 5. 6 Boxplots showing the DSC and IoU results of different segmentation models obtained on the Genoa LSCC dataset.

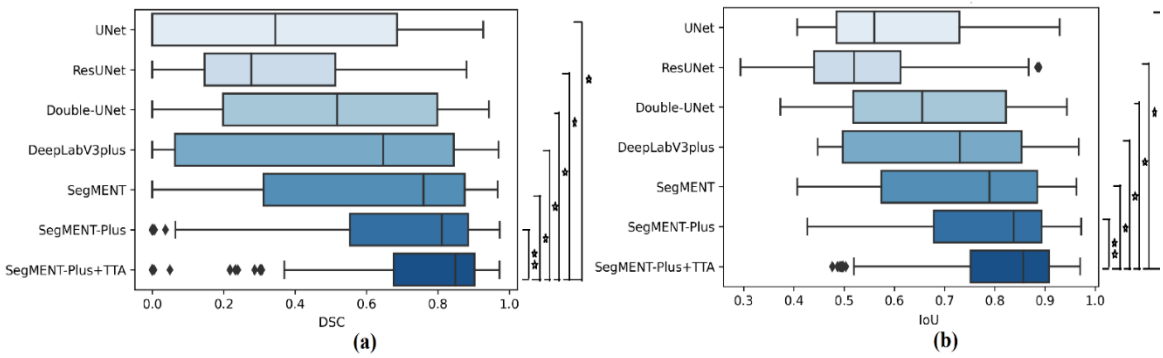


Figure 5. 7 Boxplots showing the DSC and IoU results of different segmentation models obtained on the Seoul, South LSCC dataset.

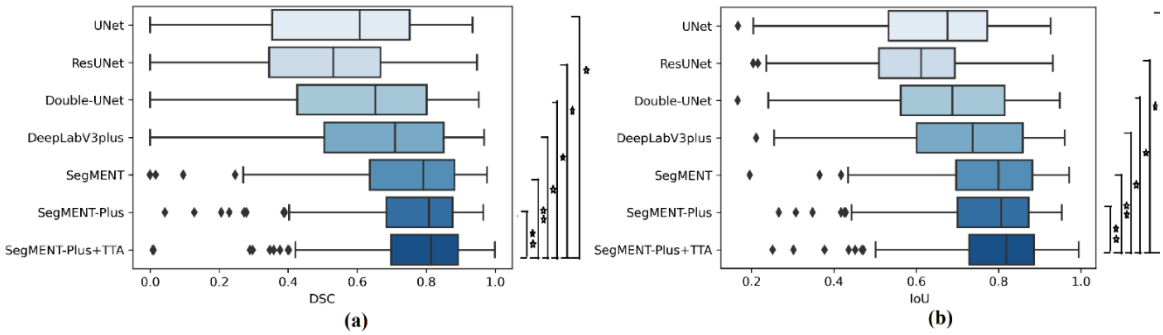


Figure 5. 8 Boxplots showing the DSC and IoU results of different segmentation models obtained on the Brescia LSCC dataset.

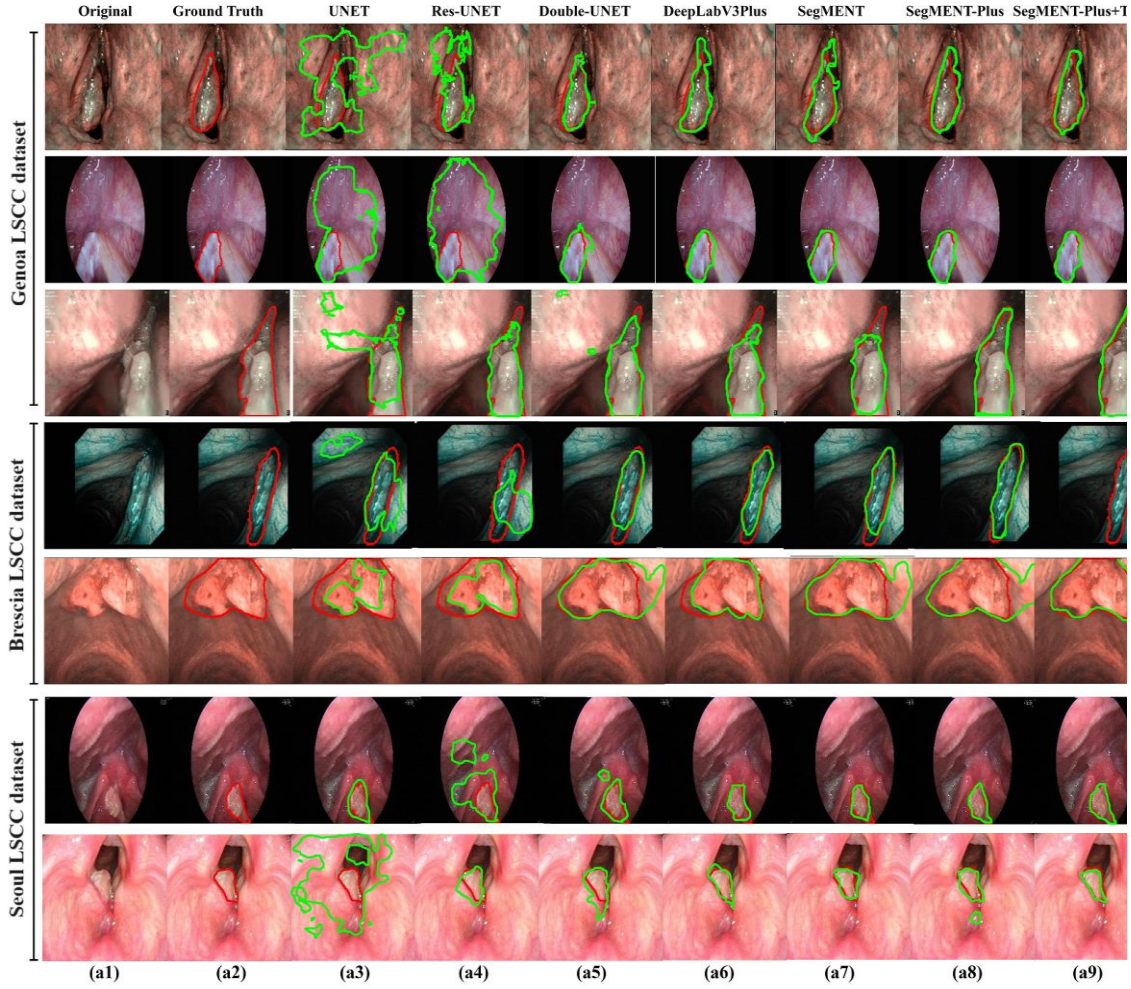


Figure 5. 9 Sample LSCC segmentation results obtained with different segmentation models. The red contour in the images represent the ground truth, while the green contours are the segmentation results. From top to down rows: First three rows are from the Genoa LSCC test dataset; the following two rows are from the Brescia LSCC testing set; the last two rows are from the Seoul LSCC testing dataset. Images in column (a1) are the original images, while column (a2) shows the ground truth images. The images in columns (a3) – (a7) show the segmentations produced by the baseline segmentation models. The images in column (a8) are from *SegMENT-Plus* and in (a9) from *SegMENT-Plus* + TTA.

Figure 5.10 presents the KDE curves for the Genoa, Seoul, and Brescia datasets. Examining the plots, Segment-Plus and *SegMENT-Plus* with TTA consistently surpassed other state-of-the-art segmentation models in terms of performance.

5.3.3 Statistical Analysis

Based on the ANOVA T-test, DSC and IoU metrics showed a significant difference ($p<0.001$) between *SegMENT-Plus* and all the other segmentation models for all three datasets. Fig. 6-8 show boxplots of quantitative results comparing the performance of the different segmentation models tested. The Pairwise difference Tukey's Honestly Significant Differences (HSD) test was also conducted for additional analysis. On IoU and DSC, SegMENT-Plus outperformed *UNet, ResUNet, Double-UNet*, and *DeepLabV3Plus* on the Genoa LSCC testing dataset ($p=0.001$). On the Seoul LSCC dataset, *SegMENT-Plus*+TTA outperformed *UNet, ResUNet*, *Double-UNet*, *DeepLabv3Plus*, and *SegMENT* on DSC metric ($p=0.001$, $p=0.001$, $p=0.001$, $p=0.001$, $p=0.001$, $p=0.002$, respectively). In addition, *SegMENT-Plus*+TTA outperformed *UNet, ResUNet, Double-UNet,* and *DeepLabv3Plus* for IoU values ($p=0.001$, $p=0.001$, $p=0.02$, $p=0.001$). On the Brescia LSCC dataset, *SegMENT-Plus* outperformed *UNet, ResUNet, Double-UNet,* and *DeepLabV3Plus* in terms of IoU and DSC ($p=0.001$). Sample LSCC segmentation results obtained with different segmentation models on the three test sets are shown in Figure 5.9. The quantitative results from three test sets are shown in Tables 5.4–5.6.

Table 5. 4 Genova LSCC median with (IQR) values as representative of test image findings.

Method	Backbone	DSC	IoU	Rec	Prc	Acc	FPS
UNET	-	0.628 (0.77-0.42)	0.688 (0.79-0.58)	0.709 (0.91-0.42)	0.720 (0.91-0.45)	0.944 (0.97-0.89)	25.3
Res-UNET	-	0.576 (0.70-0.35)	0.641 (0.72-0.55)	0.735 (0.93-0.49)	0.561 (0.78-0.31)	0.917 (0.95-0.87)	26.6
Double-UNET	VGG19	0.736 (0.84-0.52)	0.765 (0.85-0.62)	0.827 (0.95-0.62)	0.772 (0.93-0.54)	0.957 (0.98-0.91)	20.2
DeepLabv3plus	Resnet-50	0.759 (0.87-0.59)	0.778 (0.86-0.66)	0.847 (0.95-0.64)	0.814 (0.93-0.61)	0.961 (0.98-0.92)	**30.8**
SegMENT	Xception	0.809 (0.89-0.67)	0.819 (0.89-0.71)	0.847 (0.94-0.67)	0.871 (0.95-0.69)	0.971 (0.99-0.94)	21.6
SegMENT-Plus	EfficientNetB5	**0.827 (0.90-0.70)**	**0.828 (0.90-0.73)**	**0.881 (0.96-0.74)**	0.855 (0.94-0.70)	0.972 (0.99-0.95)	25.6
SegMENT-Plus +TTA	EfficientNetB5	0.819 (0.90-0.68)	0.825 (0.90-0.72)	0.880 (0.96-0.70)	**0.882 (0.95-0.71)**	**0.974 (0.99-0.95)**	9.1

Table 5. 5 Seoul LSCC median with (IQR) values statistical results on external test images.

Method	Backbone	DSC	IoU	Rec	Prc	Acc	FPS
UNET	-	0.345 (9.69-0.0)	0.560 (0.73-0.48)	0.697 (0.97-0.0)	0.259 (0.70-0.0)	0.957 (0.98-0.92)	27.0
Res-UNET	-	0.278 (0.52-0.14)	0.520 (0.62-0.44)	0.914 (1.0-0.86)	0.165 (0.40-0.08)	0.885 (0.94-0.80)	28.1
Double-UNET	VGG19	0.519 (0.80-0.19)	0.656 (0.82-0.52)	0.894 (1.0-0.49)	0.434 (0.75-0.14)	0.966 (0.98-0.93)	19.5
DeepLabv3plus	Resnet-50	0.647 (0.84-0.06)	0.731 (0.86-0.50)	0.817 (0.97-0.05)	0.625 (0.85-0.07)	0.982 (0.99-0.96)	**29.4**
SegMENT	Xception	0.760 (0.88-0.31	0.789 (0.89-0.57)	0.893 (0.98-0.39)	0.744 (0.88-0.34)	0.987 (0.99-0.97)	22.4
SegMENT-Plus	EfficientNetB5	0.811 (0.89-0.55)	0.838 (0.89-0.68)	0.918 (0.98-0.72)	0.758 (0.86-0.49)	0.987 (0.99-0.97)	24.5
SegMENT-Plus +TTA	EfficientNetB5	**0.849 (0.90-0.67)**	**0.857 (0.91-0.74)**	**0.925 (0.99-0.76)**	**0.820 (0.90-0.67)**	**0.991 (1.0-0.98)**	8.4

Table 5. 6 Brescia LSCC median with (IQR) statistical results on external test images.

Method	Backbone	DSC	IoU	Rec	Prc	Acc	FPS
UNET	-	0.608 (0.75-0.35)	0.676 (0.77-0.53)	0.630 (0.86-0.28)	0.741 (0.92-0.47)	0.925 (0.96-0.85)	25.3
Res-UNET	-	0.531 (0.67-0.34)	0.612 (0.69-0.51)	0.693 (0.87-0.39)	0.652 (0.87-0.35)	0.886 (0.94-0.82)	19.6
Double-UNET	VGG19	0.653 (0.80-0.42)	0.6988 (0.82-0.56)	0.762 (0.92-0.44)	0.736 (0.92-0.45)	0.923 (0.96-0.86)	18.4
DeepLabv3plus	Resnet-50	0.710 (0.85-0.50)	0.737 (0.86-0.60)	0.760 (0.93-0.41)	**0.867 0.95-0.66)**	0.944 (0.98-0.88)	23.1
SegMENT	Xception	0.791 (0.88-0.64)	0.799 (0.88-0.69)	0.869 (0.96-0.66)	0.859 (0.95-0.64)	0.954 (0.98-0.91)	23.0
SegMENT-Plus	EfficientNetB5	0.808 (0.88-0.68)	0.807 (0.87-0.70)	0.897 (0.98-0.70)	0.822 (0.91-0.66)	0.956 (0.98-0.92)	**26.2**
SegMENT-Plus +TTA	EfficientNetB5	**0.814 (0.89-0.70)**	**0.818 (0.89-0.73)**	**0.902 (0.98-0.67)**	0.858 (0.93-0.73)	**0.961 (0.98-0.93)**	8.3

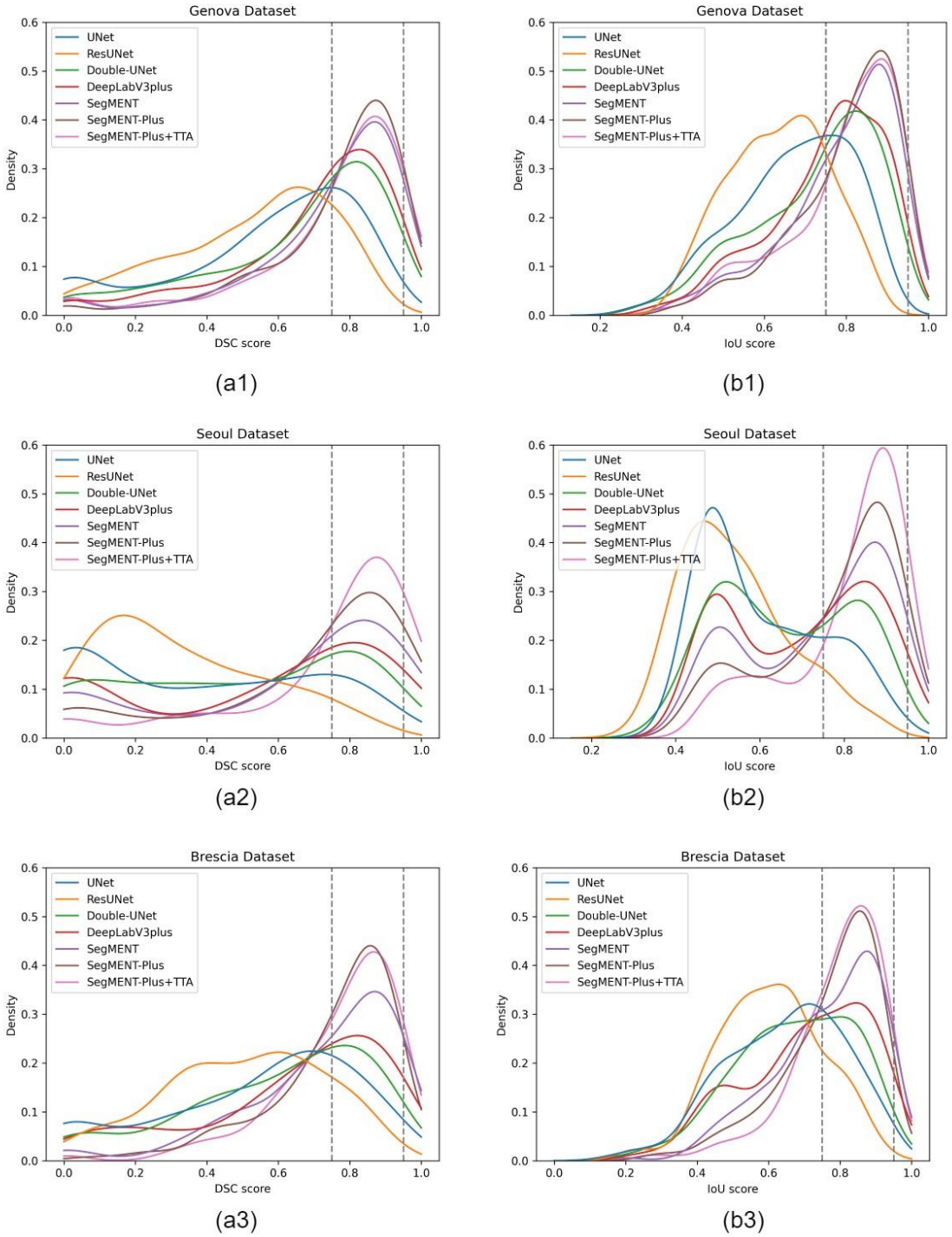


Figure 5. 10 Kernel density estimation (KDE) curves on the three test datasets (Genoa, Seoul, and Brescia LSCC) for the baselined model with *SegMENT-Plus* and *SegMENT_Plus* + TTA. Figures (a1-a3) the KDE curves of the Dice Similarity Coefficient (DSC) metric, while (b1-b2) presents the KDE curves of the Intersection over Union (IoU) metric.

5.4 Summary

This work presented *SegMENT-Plus*; a Deep Learning CNN specifically designed for automatic LSCC delineation in endoscopy images. Notably, *SegMENT-Plus* is an encoder-decoder structure with a new *m-ASPP* module including CBAM and SE blocks. The segmentation performance of the network was evaluated using a multi-center annotated dataset of WL and NBI laryngeal endoscopies. Experimental results showed that the proposed network outperformed other state of the art segmentation models. They also demonstrated the benefit of the new *m-ASPP* module, which leads to significant improvements to segmentation performance. Future research will focus both on expanding the training dataset to further improve LSCC segmentation performance, and on translating *SegMENT-Plus* to a clinical platform to start prospective trials with the technology.

Chapter 6

Towards Realtime LSCC Delineation

6.1 Overview

This chapter investigates the potential of DL for automatically delineating (segmenting) laryngeal cancer superficial extent on endoscopic images and videos. A retrospective study was conducted extracting and annotating WL and NBI frames to train a custom-made segmentation model called *SegMENT-Plus* (proposed in Chapter 5) without TTA. 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. *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 optimal excellent generalization capabilities. The computation speed of the model allowed its application on video laryngoscopies simulating real-time use. Clinical trials are needed to evaluate the role of this technology in surgical practice.

Related Publication

Sampieri C, **Muhammad Adeel Azam,** Ioppi, A., Baldini, C., Moccia, S., Kim, D., Tirrito, A., Paderno, A., Piazza, C., Mattos, L.S. and Peretti, G. (2024), **"Real-Time Laryngeal Cancer Boundaries Delineation on White Light and Narrow-Band Imaging Laryngoscopy with Deep Learning".** The Laryngoscope.
https://doi.org/10.1002/lary.31255.

6.2 Materials and methods

6.2.1 Data Acquisitions

The recordings of the video laryngoscopes of laryngeal cancer patients who underwent an endoscopic examination 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 video laryngoscopy. 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 in two different settings: in the office and in the operating room. In the office, the video laryngoscopes were performed using a flexible naso-pharyngo-laryngoscope (HD Video Rhino-laryngoscope Olympus ENF-VH, Olympus Medical System Corporation, Tokyo, Japan) through a trans nasal route. Patients undergoing transoral laser microsurgery were also examined pre-operatively in the operating room under suspension micro laryngoscopy with 0°, 30° or 70° rigid telescopes (high-definition camera head connected to a Visera Elite CLV-S190 light source, Olympus Medical System Corporation, Tokyo, Japan).

Both types of video laryngoscopes were performed first using WL and then switching to NBI. From each video laryngoscopy, several representative laryngeal cancer frames were extracted. Particularly, when available, 5 WL and 5 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. Therefore, priority in frame selection was given to informative images with few artifacts and blur. Overall, the frames obtained from micro laryngoscopy generally appeared to be sharper, brighter and of higher definition (due to the rigid telescopes used instead of flexible ones). The collected images were varied in terms of spatial resolution, with widths and heights ranging from 768 to 1920 pixels and 576 to 1072 pixels, respectively. A supervised training strategy, where a dataset is labeled to train the DL algorithm to recognize objects, was chosen for achieving the segmentation task. The extracted frames were annotated using CVAT annotation software (Intel Corporation, Santa Clara, United States)(*Computer Vision Annotation Tool,* n.d.) by 3 expert physicians who had a specific background of at least 5 years 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. NBI and WL images were used together for training so that the DL model could learn to recognize a tumor regardless of the imaging modality.

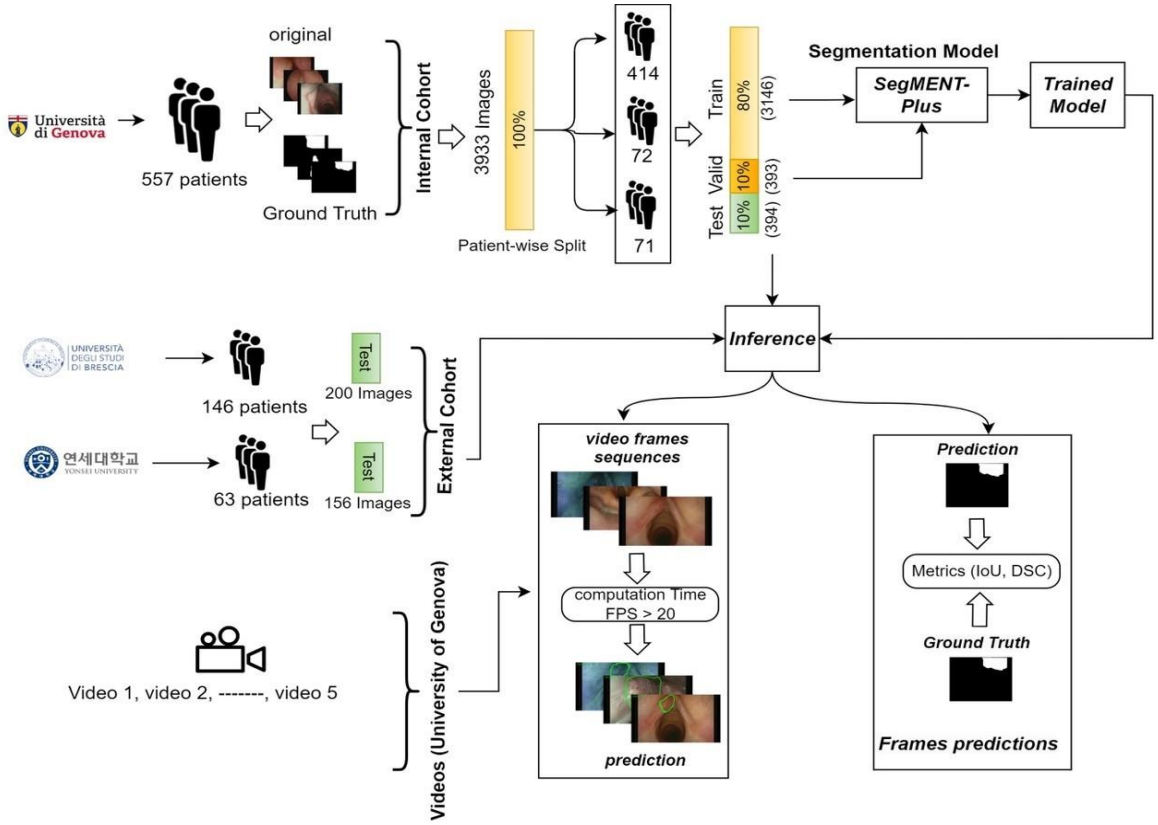


Figure 6. 1 Workflow diagram of the study. IoU = Intersection over Union; DSC = dice similarity coefficient; FPS = frames per second.

If multiple lesions were visible, multiple segmentations were carried out to select all the laryngeal cancer pixels in the image. If one physician was not sure about the margins' annotation, for example in presence of an uncertain NBI vascular pattern or mucous, the images were reviewed collectively by the three of them. At the end of the review, the frame was annotated based on the consensus of most of the experts. This image dataset (from now on referred 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. Two external datasets containing laryngeal cancer images both in WL and NBI were used to verify the generalizability of the model. 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 five years. The two external testing datasets were entirely dedicated to the testing phase of the DL model and were not used for the training. Lastly, 5 unedited preoperative video laryngoscopies of 5 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 6.1 synthetizes the workflow of the study.

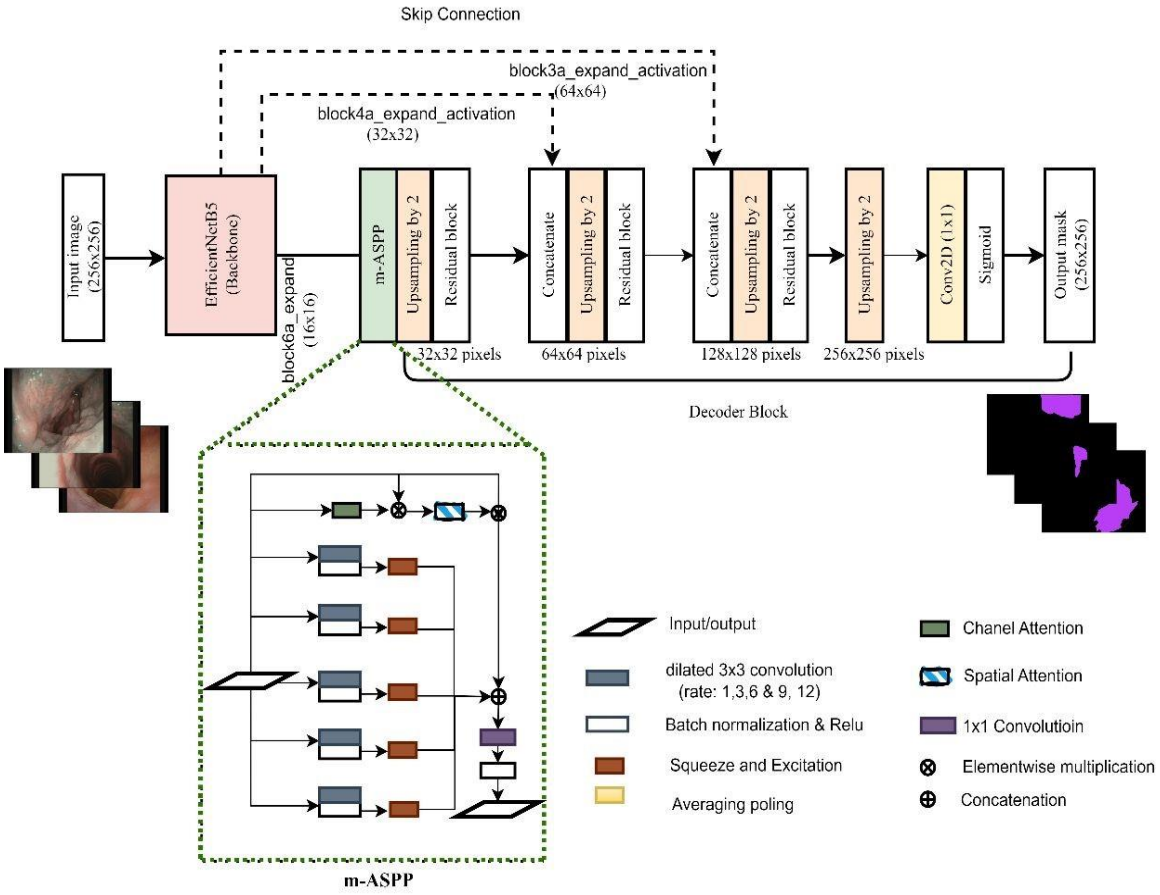


Figure 6. 2 The figure illustrates the encoder-decoder with m-ASPP module in the middle architecture used in *SegMENT-Plus.*

6.2.2 Deep learning model development and validation

The detailed model architecture already presented in (chapter 5), however overall architecture shown in Figure 6.2 and short description inserted here: The *SegMENT-Plus* is a lighter and more efficient version with 17.7 million fewer parameters (38.54%) fewer than the original *SegMENT* model (38.5% reduction in model size). While the architecture of the previous model used three standard *ASPP* modules, making it more computationally expensive, the newly proposed *SegMENT-Plus* architecture uses only one *m-ASPP* module. In fact, the previous *SegMENT* model was focusing entirely focused mostly on the number of multiscale features and struggled to refine obtain fine tumor local features of the tumors. The new m-ASPP module removes this limitation by introducing in this new version, the introduction of a Squeeze and Excitation block and a Convolution Block Attention Module, which allow, allows *SegMENT-Plus* the algorithm to focus more attention on local features. modify the image features at the end of the encoder block. Furthermore, a new residual link in the decoder of the *SegMENT-Plus* model overcomes the enables the restriction of retaining of the earlier fine encoder features that occurred in the previous *SegMENT* model.

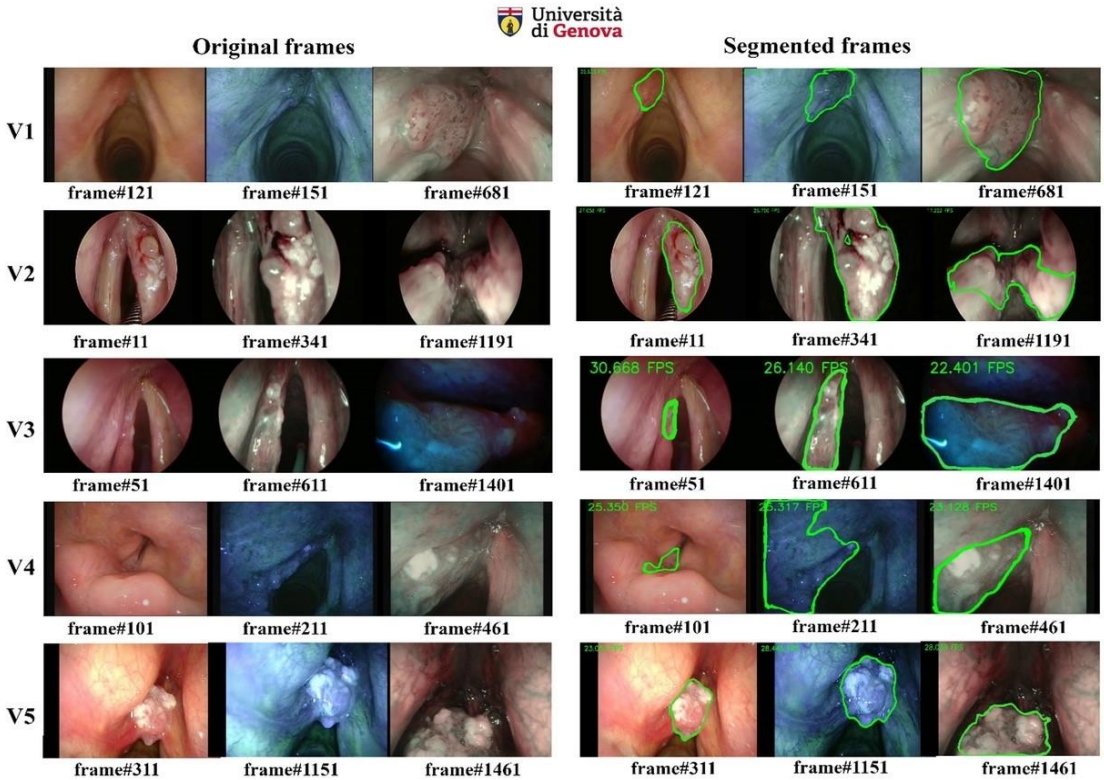


Figure 6. 3 Testing video frames extracted from five video laryngoscopes. Each row represents a different video: the pictures in the panel on the left are white light or narrow band imaging frames belonging to the original video, while the pictures in the panel on the right are the same frames after the elaboration with the deep learning model.

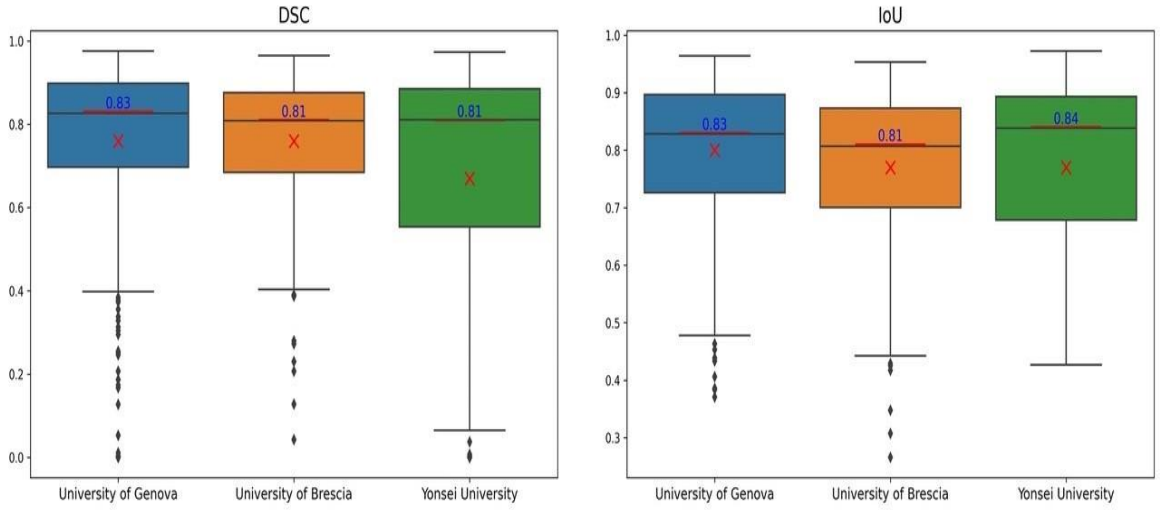


Figure 6. 4 Boxplots representing the segmentation performance of *SegMENT-Plus* on the three testing datasets. The horizontal bar inside the box represents the median value (also reported in numbers), the "x" represents the mean, and the box represents 50% of the distribution within the 1st and the 3rd quartiles. DSC, dice similarity coefficient; IoU, Intersection over union.

6.3 Experimental Setup

6.3.1 Training and validation settings

The model was programmed using Python (version 3.9) in Keras (version 2.11.0) and Tensorflow (version 2.11.0). *SegMENT-Plus* was trained for 30 epochs with a batch size of 8, starting with a learning rate of 0.001 and decreasing it with a 0.1 factor after three consecutive epochs if loss did not reduce. The Adam optimizer was used to train the model minimizing the dice similarity coefficient loss. The experiments were conducted on a workstation with a Dual Intel Xeon Gold 5222 (3.8GHz) CPU, 128 GB RAM, and a NVIDIA RTX A6000 GPU with 48 GB. The input images are of various spatial resolutions, with widths and heights ranging from 768 to 1920 pixels and 576 to 1072 pixels, respectively. The entire images are resized into 256 by 256 pixels using OpenCV library in Python (version 3.9). For standardization and reducing the computational load, all the images were resized using bilinear interpolation to 256 by 256 pixels before being

submitted into the SegMENT-Plus network. In bilinear interpolation, the weighted mean of the four closest neighboring pixels is used to calculate the resultant value of a pixel.

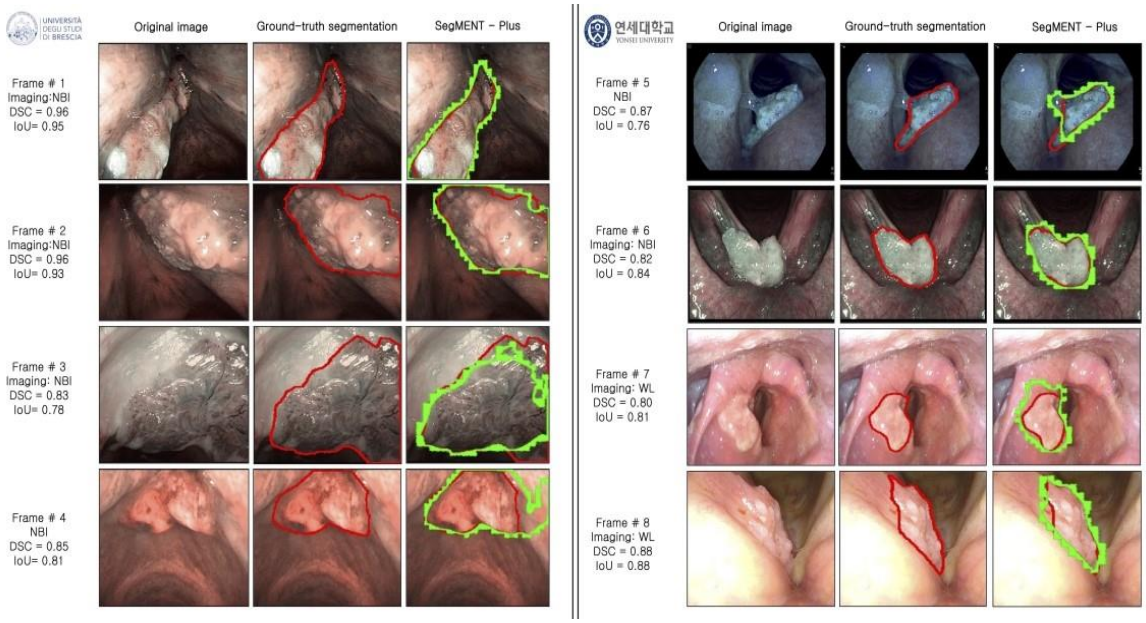


Figure 6. 5 Examples of *SegMENT-Plus* automatic segmentation of laryngeal carcinoma. The panel on the left reports four examples from the University of Brescia dataset. Frames #1 and #2 show a left and a right vocal cord carcinoma, respectively, while frames #3 and #4 represent two different anterior 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.

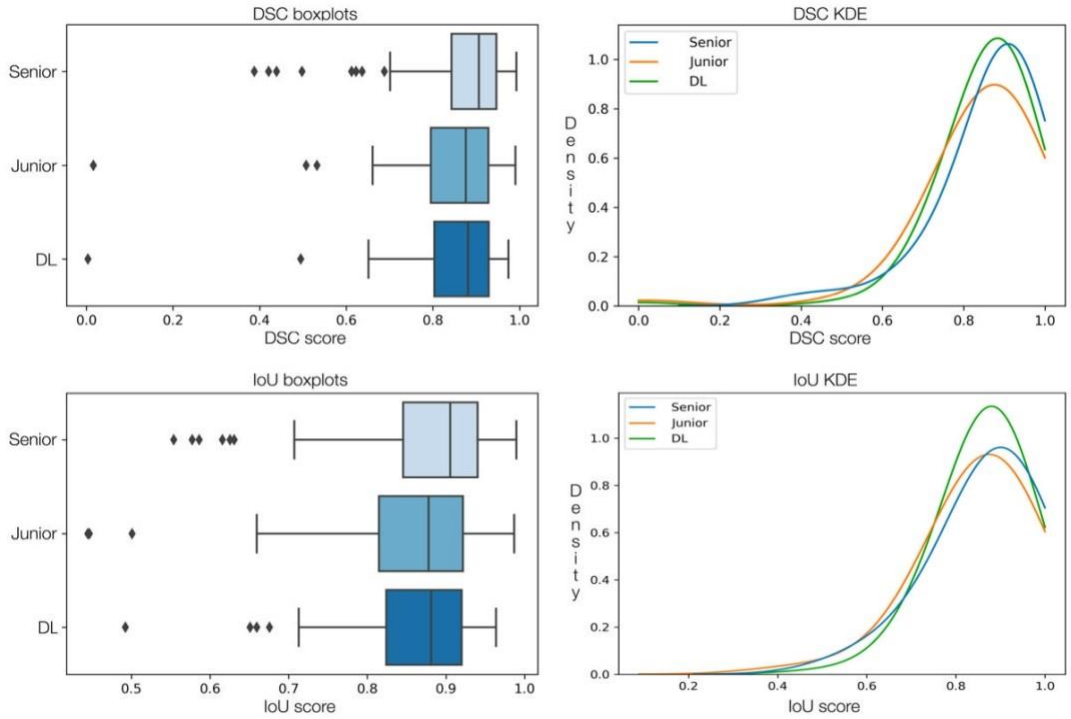


Figure 6. 6 Resident physicians' and *SegMENT-Plus* segmentation performances are represented with boxplots and kernel density estimates (KDE) curves. In the left figures, the boxplots are presented: the vertical bar inside the boxes represents the median value, while the box represents 50% of the distribution 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*.

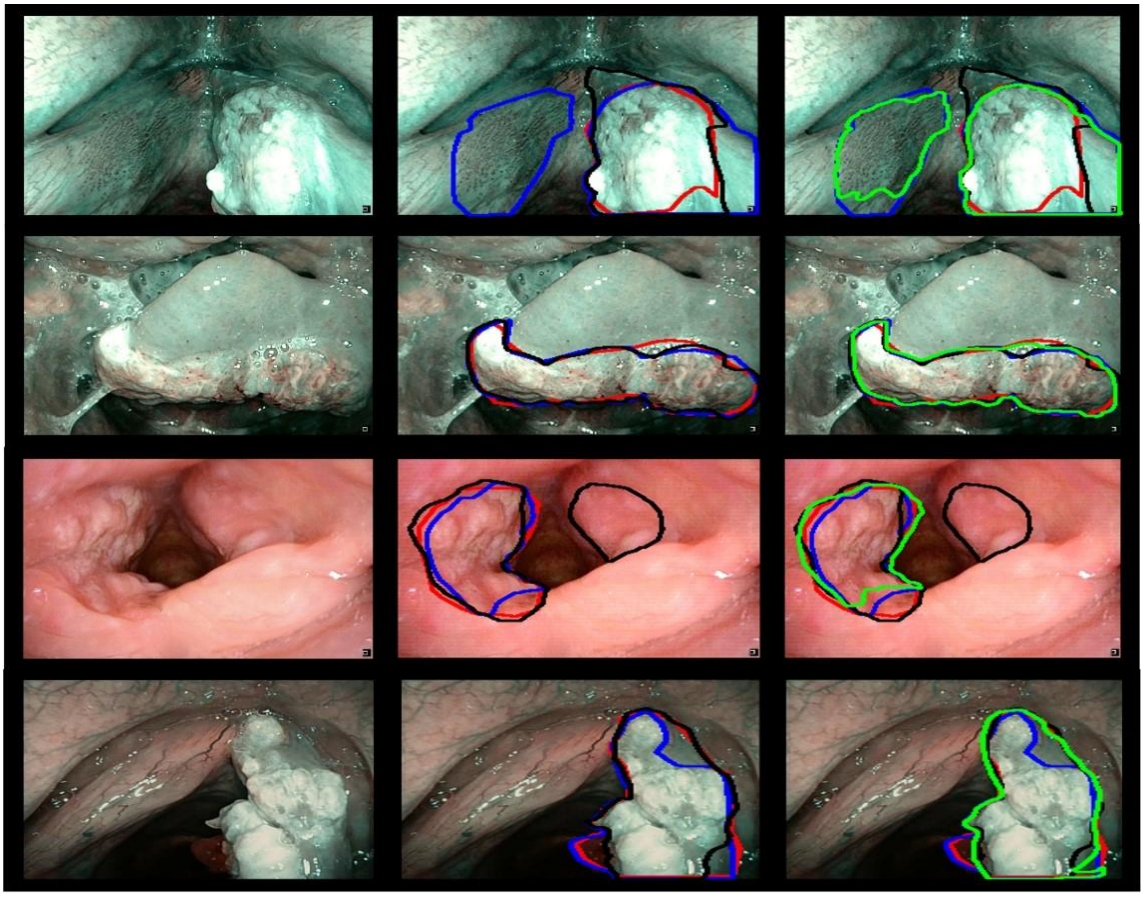


Figure 6. 7 Examples of the segmentations performed by the human physicians and *SegMENT-Plus* on the University of Genova test dataset. The column on the left shows the original frames, the column in the middle reports the annotation performed by the human physicians, and in the column on the right the DL model prediction is added. Blue = ground truth; Red = senior resident; black = junior resident; green = *SegMENT-Plus*.

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 (Ali et al., 2020; Sampieri et al., 2023; Yue et al., 2023). 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. As the most comprehensively representative 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 video laryngoscopes are generally acquired with an image rate of 25 fps, an algorithm with a median processing time of around 25 fps is compatible with real-time implementation (P. Wang et al., n.d.; X. X. Yang et al., 2021a).

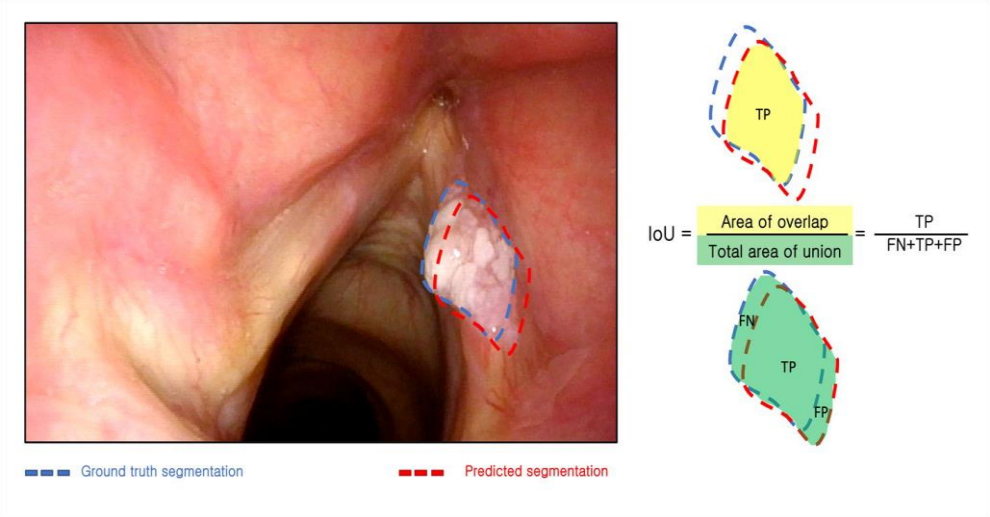


Figure 6. 8 Graphical representation of the intersection over union (IoU) The boundary traced in light blue represents the ground truth segmentation, while the red one represents the model prediction. The IoU is calculated by dividing the overlapping area (containing the true positive pixels) by the total area of union (encompassing the false negative, true positive, and false positive pixels).

6.3.2 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 for the task. The annotation of these images was 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 were previously trained in the use of the annotation software. They were also familiar with endoscopies of the upper aerodigestive tract and with the NBI analysis of laryngeal neoplasms. 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*.

6.3.3 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 and interquartile ranges (IQR: 25th and 75th). 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-Walli's test. In case of significant differences at 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).

Table 6. 1 Composition of the University of Genova LC image dataset. The subdivision of frames in training-validation and test datasets is reported. The acquisition setting (in-office vs. intra-operative) and imaging modality (white light = WL vs. Narrow Band Imaging = NBI) are described.

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			

6.4 Results

6.4.1 Internal Cohort

The video laryngoscopies of 557 patients examined at the unit of otolaryngology of IRCSS San Martino hospital, University of Genova were retrospectively retrieved. They were 485 (87.1%) males and 72 (12.9%) females with a median age of 67.0 ±7. From these video laryngoscopies, 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 6.1**.** 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 versus 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 5 previously unseen video laryngoscopies. The characteristics of the 5 testing videos and the computational time of the algorithm are reported in Table 6.3. Examples of the original video frames and the corresponding ones processed by the *SegMENT-Plus* model are shown in Figure 6.3.

Table 6. 2 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). IoU = Intersection over Union; DSC = dice similarity coefficient; FPS = frames per second.

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)

6.4.2 External Cohort

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 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 6.2 summarizes the segmentation outcomes on the three test datasets. Figure 6.4 shows the box plots of *SegMENT-Plus* performances on the three test sets. Figure 6.5 shows some automatic segmentation examples for the external validation datasets.

6.4.3 Comparison with human physicians

The result of the junior resident, senior resident, and *SegMENT-Plus* on a sub cohort 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-Walli's 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 curves and box plots are reported in Figure 6.6**.** Some graphical examples of the model's segmentation outputs compared to residents' and the ground truth ones are reported in Figure 6.7.

Table 6. 3 Testing videos characteristics and computation time expressed as median (Q1-Q3) after applying *SegMENT-Plus*. MB = megabyte; s = seconds.

Video ID	Frames number	Size (MB)	Resolution	Frame Rate	Time per frame (s)
Video 1	875	44.5	1280 x 720	25	0.044 (0.041-0.046)
Video 2	1350	69.5	1280 x 720	25	0.040 (0.036-0.041)
Video 3	1025	5.69	480 x 360	25	0.042 (0.039-0.044)
Video 4	1680	2.69	392 x 240	30	0.039 (0.036-0.041)
Video 5	1500	27.0	860 x 480	25	0.039 (0.036-0.041)

6.5 Summary

The proposed DL segmentation model was able to accurately delineate laryngeal cancer boundaries in endoscopic images and videos. The high diagnostic performance was also maintained when tested on two different external datasets demonstrating robust generalization capabilities. The fast computation speed of the model allowed to successfully apply it on video laryngoscopies 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.

Chapter 7

Conclusions and Future Directions

7.1 Overview

The discussion and overview of each study mentioned in previous chapters are presented in this chapter to conclude the thesis. The discussed problems based on nature of work distinguish into four distinct categories in this chapter: UADT cancer delineation, automatic LSCC detection, precise LSCC delineation and automatic real-time LSCC delineation. Each section describes the proposed solution(s) to the highlighted challenges, discusses the limitations of such solutions, and provides future research areas. Finally, a summary of the findings brings the thesis to a conclusion.

7.2 UADT cancer delineation

The development of semantic segmentation AI models for medical image analysis is a field of study that is becoming increasingly widespread, especially for radiologic imaging (Jha et al., 2021; Qiu et al., 2021; Wallner et al., 2019). Conversely, the exploitation of these algorithms to investigate videoendoscopic images represents a sphere of research less explored in the literature, as demonstrated by the lack of proper terminology to indicate such a field of interest before our first proposal to identify it as “Videomics” (Paderno, Holsinger, et al., 2021). Notably, most studies analyzing the applicability of DL-based semantic segmentation models in endoscopy come from the gastrointestinal field (Wu et al., 2021; Zhong et al., 2021)(Mendel et al., 2017; Wu et al., 2021). On the other hand, reports regarding the use of such algorithms in videoendoscopic examination of the UADT are scarce. This can be attributed to many factors: first, the insidious and diversified mucosal anatomy and a wide variety of lesions arising from this region are a challenge for AI-based image recognition models; second, while HD colonoscopy and esophago-gastroduodenoscopy have been routinely used in worldwide screening protocols for decades,(Saito et al., 2021; Winawer et al., 2006) implementation of UADT endoscopic examination coupled with HD-video endoscopy is more recent and less widespread, thus limiting the amount of data available for the application of AI technologies. In our study the promising results of a new CNN specifically designed for the automatic segmentation of UADT SCC. This new network was first tested on a dataset of LSCC images and

subsequently validated on OCSCC and OPSCC images obtained from a previous study (Paderno, Piazza, et al., 2021a). The proposed model showed similar diagnostic outcomes for all three investigated sites, demonstrating good generalization capacity and, thus, the potential for using it in real-life clinical scenarios. Considering the tests performed on the LSCC dataset, the *SegMENT* ensemble model performed better than the state-of-the-art CNNs, especially considering the IoU and Dsc metrics, which are the most reliable and widely used for the evaluation of semantic segmentation models. These performances were maintained when the model was validated on the OCSCC and OPSCC cohorts, where *SegMENT* outperformed the results of the state-of-the-art models previously investigated on the same datasets (Paderno, Piazza, et al., 2021a). Notably, the adjunction of specific-TL features borrowed from the LSCC dataset allowed reaching even higher results, especially for the OPSCC dataset, compared to the basal *SegMENT* pre-trained on ImageNet. While most methods for medical image analysis employ TL from general natural images (e.g., ImageNet) this strategy has been proven to be less effective compared to in-domain TL due to the mismatch in learned features between natural and medical images (Saito et al., 2021). Similarly, our results indicate that in-domain TL is a promising strategy for the processing of UADT video endoscopies. Nonetheless, our findings should be further validated on larger datasets, as the results on the OCSCC dataset were improved less compared to OPSCC. These contradictory outcomes may be explained by the heterogeneous image composition of the OCSCC dataset. Indeed, endoscopic examination of the oral cavity often includes videoframes of the lip cutaneous surfaces, alveolar ridges, or teeth crowns. Therefore, these non-mucosal areas that differ markedly from the laryngeal and oropharyngeal endoscopic appearance may have contributed to confuse the model and decrease its performance. Moreover, the limited number of images included in the validation datasets might limit the effect of in-domain TL which we believe could lead to even better results if tested on more images. The comparable results obtained by the proposed model for the LSCC (WL+NBI), the OCSCC (NBI only), and OPSCC (NBI only) datasets confirm the good generalization capacity of this model, which performs well regardless of the light source used to acquire images. Moreover, even without in-domain TL, the *SegMENT* ensemble model maintained its performance on the OCSCC and OPSCC datasets regardless of the small training datasets (13% and 15%, respectively, compared to the LSCC dataset). This finding possibly suggests that EEFs images, with their enhanced visual characteristics, may offer more information to the model during the training phase, thus helping to mitigate the shortage of images. Nevertheless, a prospective study comparing WL vs. EEFs cohorts is recommended to better investigate this finding. To date, the present study represents the first attempt to validate a DL-based semantic segmentation model capable of achieving good results in the endoscopic assessment of SCC arising from the oral cavity, oropharynx, and larynx. The segmentation task has been seldom applied to videoendoscopic images of the UADT, making this field of study innovative. (Laves et al., 2019) previously tested different CNNs to automatically

delineate different tissues and anatomical subsites on images obtained during intraoperative endoscopic evaluation of the glottis. The authors reported high mean values of IoU (84.7%) by segmenting every object in the image but did not focus on the annotation of cancer. Moreover, their dataset consisted of similar images obtained from only two patients, hence impairing a reasonable comparison with our results. A more specific paper on laryngeal lesion segmentation was published by (Ji et al., 2020a). In this work, several CNN models were implemented to delineate glottic leucoplakias on a dataset of 649 images with segmentation metrics in line with our results (DSC=0.78 and IoU=0.66). Interestingly, the best processing performance achieved by their models was 5 fps which, together with the previous results of (Paderno, Piazza, et al., 2021b), ranging (8.7-16.9) fps were faster than the 3.9 fps processed by our proposed model. Notwithstanding, all these processing times are still far from real-time inferences (20-30 fps), meaning that different strategies and different CNNs must be explored to maintain high diagnostic performance while reaching real-time efficiency. Investigating a different UADT subsite, Li and colleagues employed a CNN to automatically segment endoscopic images of nasopharyngeal carcinoma. In (C. Li et al., 2018) work a large dataset of WL in-office endoscopic frames (30,396 images) obtained in a single tertiary-level institution. Of note, the only segmentation metric reported is mean DSC (0.75±0.26), which is comparable to the results of our model. Interestingly, the authors underlined the value of the proposed semantic segmentation algorithm as an instrument to perform a target biopsy in case of suspicious nasopharyngeal lesions. Indeed, the common presence of adenoid/lymphatic hyperplasia in this area burdened the performance of endoscopic biopsies taken in an office-based setting, thus resulting in considerable false negatives rates. The same issue is frequently encountered when performing endoscopy-driven biopsies of suspicious neoplasms of other UADT districts. Regarding the oropharynx, lesions arising from the base of tongue and amygdala-glossal sulcus are frequently hidden by lymphatic tissue present in this area and can sometimes be misinterpreted and confused with mucosal/lymphatic hyperplasia. Similarly, tumors in the oral cavity can nest in inflammatory lesions such as leucoplakias or lichen, which may also hinder the real target to biopsy. Regarding the larynx, the performance of incisional biopsies under WL has been largely questioned due to its low sensitivity (Alzubaidi et al., 2020). The introduction of EEFs, allowing us to better select the most suspicious area to target, represented an important step forward leading to a significant improvement in the performances of endoscopic biopsies. Nevertheless, the difficulties encountered during human evaluation of such images burden the capillary application of EEFs, hence paving the way to the innovative application of AI for these tasks (Carobbio et al., 2021; Galli et al., 2021; Nair et al., 2021; Piazza, Cocco, De Benedetto, et al., 2010b). Indeed, the use of trustworthy semantic segmentation DL models during endoscopy may automatically delineate the superficial area where to conduct the biopsy, even when facing heterogeneous and mystifying lesions such as SCCs of the UADT. Additionally, automatic-segmentation

models may find a field of application even intraoperatively for surgical guidance (Piazza, Cocco, De Benedetto, et al., 2010a), or for driving tumor excision and provide an improved rate of negative surgical margins. Of note, the use of NBI during surgery for SCC of the oral cavity, oropharynx, and larynx has been already shown to be effective in these tasks, (Garofolo et al., 2015; Gong et al., 2021) but we believe that AI-based tools, once rigorously validated, will represent a more precise and objective method for surgical margins sampling. Finally, pursuing automatic segmentation in this field is expected to become increasingly relevant in the future not only for surgical practice but in other fields as well. In fact, semantic segmentation is paramount for establishing boundaries between objects to explain complex situations to computers. In the future, highly elaborated AI tools might be able to autonomously understand the relationships between different elements, even in a highly complex environment such as the UADT and suggest meaningful clinical decisions to physicians.

7.2.1 Limitation and future direction

The present work has limitations that must be acknowledged. First, the study conclusions are restricted by the small size of the datasets, especially the external validation cohorts, and by its retrospective design. To overcome such drawbacks, an enriched data collection will characterize our future projects to increase the dataset's size. Additionally, data acquisition protocols will be implemented by gathering videos from different video sources, with the purpose of enhancing the generalization capability of the algorithm. Furthermore, the previously collected NBI images from the validation cohorts did not allow head-to-head comparison of WL vs. NBI. Moreover, as *SegMENT* is trained and validated on both WL and NBI images, the "bioendoscopic" appearance of the lesion is of crucial importance to achieve optimal diagnostic performance. In this light, the finding of diffuse and narrow intrapapillary capillary loops typical of inflammatory diseases or radiotherapy, may contribute to confuse the algorithm in detecting the accurate contour of the lesion. Future projects should be aimed at training CNN models with endoscopic images obtained by heterogeneous cohorts of patients, such as previously irradiated or concomitantly affected by inflammatory diseases. Finally, the annotations performed by physicians were not cross validated, representing a bias that will be addressed in future studies.

7.3 Automatic LSCC detection

In recent years, the use of AI in medicine has been rising exponentially. In particular, the application of CADe and CADx systems to video endoscopy has been explored in the field of gastroenterology to detect areas of mucosal inflammation, polyps, precancerous, and cancerous lesion. Conversely, to our knowledge, there are only a few reports in the present literature regarding cancers of the upper aerodigestive tract (UADT), making it an emerging

field of research (Esmaeili et al., 2020a; Mascharak et al., 2018; Paderno, Piazza, et al., 2021a; Ren et al., 2020; Xiong et al., 2019). Among these, the largest reports regarding LSCC are those from Xiong et al. (Xiong et al., 2019) and Ren et al.,(Ren et al., 2020) who applied CADx to classify precancerous and cancerous laryngeal lesions, with both reporting very high values of sensitivity and specificity. Nevertheless, these studies investigate the use of DL models only with WL endoscopy. It is our belief, however, that modern laryngological workup should not be performed without a bio endoscopic approach such as that allowed using NBI. This technology, in fact, has demonstrated a tremendous impact on LSCC diagnosis, intraoperative definition of excisional margins, and post-treatment follow-up (C et al., 2010; Ren et al., 2020). In this context, our models were trained and tested with a mixed dataset of WL and NBI images to obtain a reliable algorithm that can provide a fluid experience while switching between WL and NBI during real-time videolaryngoscopy. In this regard, to our knowledge, only three studies have explored NBI imaging applied to AI in the UADT. A promising pilot study by Mascharak et al. aimed to investigate if NBI imaging evaluation was feasible for machine learning CADx in oropharyngeal carcinoma (Mascharak et al., 2018). Even with a very limited cohort (30 patients), they demonstrated the superiority of NBI compared to WL in detecting this type of neoplasm, reporting a specificity of 70.0% versus 52.4%, and a sensibility of 70.9% versus 47.0%, respectively. Paderno et al. more recently applied different CNNs to the automated segmentation of WL and NBI frames of the oral cavity and oropharyngeal neoplastic lesions, obtaining accuracy rates like those reported herein (Paderno, Piazza, et al., 2021a). Finally, Inaba and colleagues reported rates of sensitivity, specificity, and accuracy of 95.5%, 98.4%, and 97.3%, respectively, in the diagnosis of very small cancerous lesions arising from the mucosa of the UADT while performing esophagogastroduodenoscopy (Inaba et al., 2020). However, they focused mainly on hypopharyngeal carcinoma using only NBI images, while the only laryngeal subsite examined (epiglottis) was the worst-performing in terms of sensitivity (85.7%). Nevertheless, as stated, the implementation of NBI in the development of the DL algorithm led to an enhancement of tumor visibility, thus helping their model to better recognize neoplasms. This also explains the promising detection performance achieved by our pilot CADe models, which will likely improve as our training dataset grows. One of the main focuses of this work was to explore the feasibility of the CADe in real-time video laryngoscopy, a topic rarely investigated in the literature so far. The validation performed on video streams was possible, thanks to the low computational time required by the YOLO ensemble model, requiring only 26 milliseconds to analyze one videoframe. Considering that video laryngoscopies typically range from 25 to 30 frames-per-second, with its average computing time of 38.5 frames-per-second, the model could achieve real-time processing performances. Differing from UADT, real-time CADe technology has already been implemented in the detection of mucosal abnormalities in gastroenterology with similar computing times (P. Wang et al., 2018; X. X. Yang et al., 2021b). Of note, algorithms that

can perform in real-time necessarily require a reduction in terms of complexity. In fact, as shown by Cho et al., models characterized by a high number of parameters, even if performing remarkably (reported accuracy of 99.7%), failed to perform in real-time, therefore, needing to limit the inference process to one image every five frames (Cho & Choi, 2022). On the other hand, our model, even if it successfully identified LSCC in videos with good inference and reduced computation time, occasionally detected FP objects. This happened since LSCC comes in a variety of forms, colors, textures, and sizes, making the building of a solid DL detection model very demanding. This pilot study allowed us to identify a suitable detection model for real-time endoscopy implementation while exploring several pre-processing strategies that can enhance diagnostic performances. To our knowledge, this is the first report of AI-aided UADT cancer detection using YOLO. This DL model proved to be suitable to detect LSCC on both images and videos with adequate performance in real time applications. Our ensemble model demonstrated the best LSCC detection performance in terms of precision, recall, and mAP@.5. From a technical point of view, we underline that pre-processing techniques like CLAHE can be critical in fields like endoscopy where the quality of video images is often suboptimal. On the other hand, data augmentation processes should be used extensively to help the model to learn to identify heterogeneous lesions such as LSCCs. Indeed, implementing these techniques in the DL framework leads to higher accuracy performances. In addition, the TTA methodology assessed in this study also enhanced the inference power of the original YOLO model, leading to increased TP rate. Therefore, it represents another promising strategy for future algorithms.

7.3.1 Limitation and future direction

The shortcomings of the present study comprise the relatively small LSCC dataset and the exclusion of benign laryngeal lesions. The next phase of this research will be directed to build a solid and trustworthy algorithm by enriching the training dataset with thousands of LSCC frames and a comparable number of images from multiple and heterogeneous benign lesions (vocal polyps, cysts, nodules, papilloma's, etc.) both sourced by a multicenter collaboration. Finally, the comparison of this model's detection accuracy with expert physicians will be of paramount importance for definitive validation. A comprehensive model suitable for clinical practice must be validated by rigorous research. We believe these preliminary findings will help other groups progress research in this field.

7.4 Precise LSCC delineation

The *SegMENT-Plus* architecture has improved the *SegMENT* segmentation model, which was developed in 2022 for the purpose of segmenting laryngeal cancer. This architectural add included a comparison of the *m-ASPP* module to other cutting-edge *ASPP* modules. Later study focused on the synergistic integration of CBAM and SE modules to improve

network performance, as opposed to the standard *ASPP* module. The incorporation of the SE block after each dilated convolution refines the numerous receptive FoVs and extends the focus, bringing the laryngeal tumor locations into attention. Analyzing the results from the ablation study, the first thing noted is that the poor performance of *ResNet-50, EfficientNetV2B3*, and *EfficientNetB7*. This can be explained by the fact that EfficientNetV2B3 has the smallest number of parameters, while *ResNet-50* is likely too simple for LSCC images. *EfficientNetB7*, on the other hand, has a large backbone that over-fits our LSCC dataset. With Dense *ASPP*, the segmentation performance was increased because the denser connections enable this model to retrieve more global information. However, the dense module increases the size of our segmentation network parameters, making computation more expensive. Furthermore, the retrieval of local information is still limited. Segmentation performance with *EfficientNetB5* and *EfficientNetB6* were practically the same, scoring the top values in all metrics. Nonetheless, *EfficientNetB5* demonstrated to be faster, reaching higher FPS because it has fewer parameters. Therefore, the backbone model selected for *SegMENT-Plus* was the *EfficientNetB5.* Regarding the assessment of *SegMENT-Plus* with different *ASPP* modules, our new *m-ASPP* module demonstrated outperforming all the other state-of-the-art modules. On the other extreme, it is interesting to note that *WASPP* had practically no impact on the segmentation performance, which can be explained by the fact that *WASPP* has fewer parameters than the standard *ASPP*, which makes the model computationally cheaper, but compromise segmentation performance.

When comparing the segmentation performance of *SegMENT-Plus* with that of *UNET, Res-UNET, Double-UNET, DeepLabV3Plus* and *SegMENT,* two important characteristics can be observed. The first is that the models without *ASPP* (*UNET, Res-UNET* and *Double-UNET*) struggle to extract multiscale tumor features, likely because they miss local and global information, as argued in (L. Chen et al., 2018a). The second characteristic is that the models with *ASPP* (*DeepLabV3Plus* and *SegMENT*) can indeed overcome the limitation regarding multi-scale and global features of laryngeal images, but still struggle to extract local information of tumor regions. This is likely due to absence of attentions modules (CBAM and SE) in their *ASPP* modules, which agrees with the discussions in (Woo et al., 2018). Our proposed *m-ASPP* module solved the two limitations mentioned above, showing that CBAM effectively maintains low and high-level image features, while SE contributes to enhance the network's attention on tumor regions. The *m-ASPP* module has no densely connected layers and has a reduced number of parameters compared to Dense *ASPP*. As a result, the proposed *m-ASPP* without TTA is suitable for real-time implementation on high-end systems. The *SegMENT-Plus* model has 17.7 million parameters, which is 38.54% fewer than the original *SegMENT* model, with benchmark results on DSC and IoU that are 2.2% and 1.1% better than *SegMENT*, respectively. Furthermore, we tested the generalizability of the model on two external cohorts of laryngeal cancer endoscopic images with improved results. However, the *Double-UNET* model is challenging to apply in real-time clinical

settings due to its immense number of parameters (29.3 million), which is 65.54% more computationally expensive than the *SegMENT-Plus* model. The quantitative experiments showed *that SegMENT-Plus* improved 12.5% in DSC and 8.24% in IoU than *Double-UNET* model. It is also observed that during training the *Double-UNET* model converges slowly. However, during testing the model did not learn tumor features well and was not able to converge even though we increased the rate of convergence after three consecutive epochs with condition if loss did not improve during testing. It was noted that although *SegMENT-Plus* without TTA is fast and uses less memory, its performance exhibits slightly lower accuracy and robustness. This is because it may not capture all the variations in the test data when compared to model with TTA.

7.4.1 Limitation and future direction

SegMENT-Plus with TTA improved generalization to external test datasets by exploiting random modifications to the test images. However, this makes the model computationally expensive and not suitable for real-time segmentations. Nonetheless, if real-time performance is not required, *SegMENT-Plus* with TTA would be the preferred segmentation model. Finally, when testing *SegMENT-Plus* on complete endoscopic video sequences, we noticed consistency issues with the segmentation performance when the endoscope moves too much. This is because such movements cause motion blur and changes in illumination, leading to erroneous segmentations. Therefore, it may be indeed preferred to remove the real-time performance requirement in favor of better DSC and IoU performance. In any case, the 9.1 FPS achieved with *SegMENT-Plus*+TTA is already quite a decent performance for near real-time implementations. In the future, further improvements to segmentation performance may be pursued by pre-training the network on other larger medical image datasets before fine-tuning LSCC images. In addition, the inclusion of images from other medical centers can strengthen the generalization capabilities of the model.

7.5 Automatic real-time LSCC delineation

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 (Fiz et al., 2017b; Gorphe & Simon, 2019b; Sampieri et al., 2022b) . Several technologies have been employed to enhance the visibility of malignant tissue on the UADT mucosa (i.e. bio endoscopy, autofluorescence, contact endoscopy), yet each of these tools suffers from specific drawbacks that have limited their use (de Kleijn et al., 2023). Among those, NBI represents the most used and studied (Y. C. Lin et al., 2010; Rosenthal, 2014; Vilaseca et al., 2017b). During surgery, NBI can provide accurate real-time tumor margins information and has been proven to reduce the rate of positive margins in laryngeal transoral surgery (Zwakenberg et al., 2023).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 (Valls-Mateus et al., 2018). 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 (Azam et al., 2022). *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 (Azam et al., 2022), 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 experimental 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 (Paderno et al., 2023). 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., 2020). In this paper the authors developed a DL model able to delineate laryngeal leukoplakia from laryngoscope 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 paper, 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 margins identification, regardless 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 rate 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. *SegMENT-Plus* are not intended to replace the clinician but rather to represent a support in decision making, offering potential for standardization towards more equitable treatment.

7.5.1 Limitation and future direction

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 like 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 filters modality (other than NBI). Second, the ground truth annotations lack external cross validation while inter- and intra-annotators' reliability was not assessed in this thesis. 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.

7.6 Conclusion

The multicentric validation of a DL-based semantic segmentation model named "*SegMENT"* applied to UADT videoendoscopic images of SCC is covered in this thesis. When assessing WL and NBI images from three separate anatomical subsites (larynx, oropharynx, and oral cavity), the model maintained consistent diagnostic performance. Ensemble strategies and in-domain transfer learning techniques showed promise for improving segmentation performance. The YOLO ensemble model proved to be efficient in detecting LSCC in video laryngoscopes. The remarkable computational times could represent the keystone in employing the YOLO ensemble model for real-time LSCC detection soon. The availability of NBI images to feed the algorithm represented a pivotal point to reach the detection performances observed, even considering the small training set used in this study. The model represents a promising higher detection performance within pre-operative and intra-operative laryngeal examination. The *SegMENT-Plus* CNN was designed specifically for automatic LSCC delineation in endoscopic images. *SegMENT-Plus*

is comprised of encoder-decoder structure that includes a new m-ASPP module with SE and CBAM blocks. A multi-center annotated dataset of WL and NBI laryngeal endoscopies was utilized to evaluate the network's segmentation performance. According to testing results, the proposed network outperformed other cutting-edge segmentation methods. It also demonstrated the benefits of the recently added *m-ASPP* module, which results in significant improvements in segmentation efficiency. Furthermore, it precisely indicates the boundaries of laryngeal cancer in endoscopic images and videos. Testing on two different external datasets indicated high generalization skills while keeping good diagnostic performance. The model's fast computing speed allowed it to be effectively applied to video laryngoscopies, indicating that it may be utilized in real time. This data implies that a clinical implementation is a realistic technique to verify the model's utility in a real-world context.

Future research will focus on both expanding the training dataset and adapting *SegMENT-Plus* to a clinical platform to launch prospective trials leveraging the technology to improve LSCC segmentation performance. This project remains in progress, with several publications planned. Expect to see papers in engineering-related journals soon. The dataset used in this project has been declared as confidential due to its commercial value in the development of AI systems for the transformation of AI technology in healthcare settings. However, it is vital to emphasize that the dataset will be made available for future study reasons. As of today, the team has decided to disclose the AI model weights, as well as any related software and platform programs. This facilitates collaboration and allows other researchers to expand on the present approach by training and testing their models on comparable datasets. While currently not using a federated approach due to the dataset's confidentiality and strategic value, the project plans to reassess in the future. This aims to transition towards a federated model for broader sharing, addressing privacy concerns and encouraging collaborative progress in the field.

References

Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, K. M. (2016). TensorFlow: A System for Large-Scale Machine Learning. *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16) 2016*, 265–283. https://doi.org/10.1016/0076-6879(83)01039-3

Abraham, N., & Khan, N. M. (2018). *A Novel Focal Tversky loss function with improved Attention U-Net for lesion segmentation*. http://arxiv.org/abs/1810.07842

Adamian, N., Naunheim, M. R., & Jowett, N. (2021). An Open-Source Computer Vision Tool for Automated Vocal Fold Tracking From Videoendoscopy. *Laryngoscope*, *131*(1), E219–E225. https://doi.org/10.1002/lary.28669

Ali, S., Zhou, F., Braden, B., Bailey, A., Yang, S., Cheng, G., Zhang, P., Li, X., Kayser, M., Soberanis-Mukul, R. D., Albarqouni, S., Wang, X., Wang, C., Watanabe, S., Oksuz, I., Ning, Q., Yang, S., Khan, M. A., Gao, X. W., … Rittscher, J. (2020). An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. *Scientific Reports*, *10*(1). https://doi.org/10.1038/S41598-020-59413-5

Alzubaidi, L., Fadhel, M. A., Al-Shamma, O., Zhang, J., Santamaría, J., Duan, Y., & R. Oleiwi, S. (2020). Towards a Better Understanding of Transfer Learning for Medical Imaging: A Case Study. *Applied Sciences*, *10*(13). https://doi.org/10.3390/app10134523

Araújo, T., Santos, C. P., De Momi, E., & Moccia, S. (2019). Learned and handcrafted features for early-stage laryngeal SCC diagnosis. *Medical and Biological Engineering and Computing*, *57*(12), 2683–2692. https://doi.org/10.1007/s11517-019-02051-5

Arens, C., Piazza, C., Andrea, M., Dikkers, F. G., Tjon Pian Gi, R. E. A., Voigt-Zimmermann, S., & Peretti, G. (2016). Proposal for a descriptive guideline of vascular changes in lesions of the vocal folds by the committee on endoscopic laryngeal imaging of the European Laryngological Society. *European Archives of Oto-Rhino-Laryngology*, *273*(5), 1207–1214. https://doi.org/10.1007/s00405-015-3851-y

Artacho, B., & Savakis, A. (2019). *Waterfall Atrous Spatial Pooling Architecture for Efficient Semantic Segmentation*. 1–17. https://doi.org/10.3390/s19245361

Artacho, B., & Savakis, A. (2021). *GourmetNet : Food Segmentation Using Multi-Scale Waterfall*.

Azam, M. A., Sampieri, C., Ioppi, A., Benzi, P., Giordano, G. G., De Vecchi, M., Campagnari, V., Li, S., Guastini, L., Paderno, A., Moccia, S., Piazza, C., Mattos, L. S., & Peretti, G. (2022a). Videomics of the Upper Aero-Digestive Tract Cancer: Deep Learning Applied to White Light and Narrow Band Imaging for Automatic Segmentation of Endoscopic Images. *Frontiers in Oncology*, *12*(June), 1–14. https://doi.org/10.3389/fonc.2022.900451

Azam, M. A., Sampieri, C., Ioppi, A., Benzi, P., Giordano, G. G., De Vecchi, M., Campagnari, V., Li, S., Guastini, L., Paderno, A., Moccia, S., Piazza, C., Mattos, L. S., & Peretti, G. (2022b). Videomics of the Upper Aero-Digestive Tract Cancer: Deep Learning Applied to White Light and Narrow Band Imaging for Automatic Segmentation of Endoscopic Images. *Frontiers in Oncology*, *12*. https://doi.org/10.3389/fonc.2022.900451

Azam, M. A., Sampieri, C., Ioppi, A., Benzi, P., Giordano, G. G., De Vecchi, M., Campagnari, V., Li, S., Guastini, L., Paderno, A., Moccia, S., Piazza, C., Mattos, L. S., & Peretti, G. (2022c). Videomics of the Upper Aero-Digestive Tract Cancer: Deep Learning Applied to White Light and Narrow Band Imaging for Automatic Segmentation of Endoscopic Images. *Frontiers in Oncology*, *12*. https://doi.org/10.3389/fonc.2022.900451

Barbalata, C., & Mattos, L. S. (2016). Laryngeal tumor detection and classification in endoscopic video. *IEEE Journal of Biomedical and Health Informatics*, *20*(1), 322–332. https://doi.org/10.1109/JBHI.2014.2374975

Bianco, S., Cadene, R., Celona, L., & Napoletano, P. (2018). Benchmark analysis of representative deep neural network architectures. *IEEE Access*, *6*, 64270–64277. https://doi.org/10.1109/ACCESS.2018.2877890

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *68*(6), 394–424. https://doi.org/10.3322/caac.21492

C, P., D, C., L, D. B., F, D. B., P, N., & G, P. (2010). Narrow band imaging and high definition television in the assessment of laryngeal cancer: a prospective study on 279 patients. *European Archives of Oto-Rhino-Laryngology : Official Journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS) : Affiliated with the German*

Society for Oto-Rhino-Laryngology - Head and Neck Surgery, *267*(3), 409–414. https://doi.org/10.1007/S00405-009-1121-6

Carobbio, A. L. C., Vallin, A., Ioppi, A., Missale, F., Ascoli, A., Mocellin, D., Bagnasco, D., Mora, R., Peretti, G., & Canevari, F. R. M. (2021). Application of bioendoscopy filters in endoscopic assessment of sinonasal Schneiderian papillomas. *International Forum of Allergy and Rhinology*, *November 2020*, 1–4. https://doi.org/10.1002/alr.22760

Carta, F., Sionis, S., Cocco, D., Gerosa, C., Ferreli, C., & Puxeddu, R. (2016). Enhanced contact endoscopy for the assessment of the neoangiogenetic changes in precancerous and cancerous lesions of the oral cavity and oropharynx. *European Archives of Oto-Rhino-Laryngology*, *273*(7), 1895–1903. https://doi.org/10.1007/s00405-015-3698-2

Chai, J., Zeng, H., Li, A., & Ngai, E. W. T. (2021). Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, *6*, 100134. https://doi.org/10.24433/CO.0411648.v1

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018a). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018b). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184

Chen, L., Papandreou, G., Member, S., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018a). *DeepLab : Semantic Image Segmentation with Deep Convolutional Nets , Atrous Convolution , and Fully Connected CRFs*. *40*(4), 834–848.

Chen, L., Papandreou, G., Member, S., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018b). *DeepLab : Semantic Image Segmentation with Deep Convolutional Nets , Atrous Convolution , and Fully Connected CRFs*. *40*(4), 834–848.

Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). *Rethinking Atrous Convolution for Semantic Image Segmentation*. http://arxiv.org/abs/1706.05587

Cho, W. K., & Choi, S. H. (2022). Comparison of Convolutional Neural Network Models for Determination of Vocal Fold Normality in Laryngoscopic Images. *Journal of Voice*, *36*(5), 590–598. https://doi.org/10.1016/j.jvoice.2020.08.003

Cho, W. K., Lee, Y. J., Joo, H. A., Jeong, I. S., Choi, Y., Nam, S. Y., Kim, S. Y., & Choi, S. H. (2021). Diagnostic Accuracies of Laryngeal Diseases Using a Convolutional Neural Network-Based Image Classification System. *Laryngoscope*, *131*(11), 2558–2566. https://doi.org/10.1002/lary.29595

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, *2017-Janua*, 1800–1807. https://doi.org/10.1109/CVPR.2017.195

Computer Vision Annotation Tool. (n.d.). Retrieved November 24, 2022, from https://cvat.org/auth/login

CVAT. (n.d.). Retrieved February 28, 2023, from https://github.com/opencv/cvat

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. *ACM International Conference Proceeding Series*, *148*, 233–240. https://doi.org/10.1145/1143844.1143874

de Kleijn, B. J., Heldens, G. T. N., Herruer, J. M., Sier, C. F. M., Piazza, C., de Bree, R., Guntinas-Lichius, O., Kowalski, L. P., Vander Poorten, V., Rodrigo, J. P., Zidar, N., Nathan, C. A., Tsang, R. K., Golusinski, P., Shaha, A. R., Ferlito, A., & Takes, R. P. (2023). Intraoperative Imaging Techniques to Improve Surgical Resection Margins of Oropharyngeal Squamous Cell Cancer: A Comprehensive Review of Current Literature. *Cancers*, *15*(3). https://doi.org/10.3390/CANCERS15030896

Deganello, A., Paderno, A., Morello, R., Fior, M., Berretti, G., Del Bon, F., Alparone, M., Bardellini, E., Majorana, A., & Nicolai, P. (2021). Diagnostic Accuracy of Narrow Band Imaging in Patients with Oral Lichen Planus: A Prospective Study. *Laryngoscope*, *131*(4), E1156–E1161. https://doi.org/10.1002/lary.29035

Ding, H., Cen, Q., Si, X., Pan, Z., & Chen, X. (2022). Automatic glottis segmentation for laryngeal endoscopic images based on U-Net. *Biomedical Signal Processing and Control*, *71*. https://doi.org/10.1016/j.bspc.2021.103116

Esmaeili, N., Illanes, A., Boese, A., Davaris, N., Arens, C., Navab, N., & Friebe, M. (2020a). Laryngeal lesion classification based on vascular patterns in contact endoscopy and narrow band imaging: Manual versus automatic approach. *Sensors (Switzerland)*, *20*(14), 1–12. https://doi.org/10.3390/s20144018

Esmaeili, N., Illanes, A., Boese, A., Davaris, N., Arens, C., Navab, N., & Friebe, M. (2020b). Manual versus Automatic Classification of Laryngeal Lesions based on Vascular

Patterns in CE+NBI Images. *Current Directions in Biomedical Engineering*, *6*(3). https://doi.org/10.1515/cdbme-2020-3018

Everingham, M., Gool, L. Van, Williams, C. K. I., Winn, J., & Zisserman, A. (2009). The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision 2009 88:2*, *88*(2), 303–338. https://doi.org/10.1007/S11263-009-0275-4

Farah, C. S., Dalley, A. J., Nguyen, P., Batstone, M., Kordbacheh, F., Perry-Keene, J., & Fielding, D. (2016). Improved surgical margin definition by narrow band imaging for resection of oral squamous cell carcinoma: A prospective gene expression profiling study. *Head & Neck*, *38*(6), 832–839. https://doi.org/10.1002/HED.23989

Fehling, M. K., Grosch, F., Schuster, M. E., Schick, B., & Lohscheller, J. (2020). Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep Convolutional LSTM Network. *PLoS ONE*, *15*(2), 1–29. https://doi.org/10.1371/journal.pone.0227791

Filauro, M., Paderno, A., Perotti, P., Marchi, F., Garofolo, S., Peretti, G., & Piazza, C. (2018). Role of narrow-band imaging in detection of head and neck unknown primary squamous cell carcinoma. *Laryngoscope*, *128*(9), 2060–2066. https://doi.org/10.1002/lary.27098

Fiz, I., Mazzola, F., Fiz, F., Marchi, F., Filauro, M., Paderno, A., Parrinello, G., Piazza, C., & Peretti, G. (2017a). Impact of close and positive margins in transoral laser microsurgery for TIS-T2 glottic cancer. *Frontiers in Oncology*, *7*(OCT). https://doi.org/10.3389/fonc.2017.00245

Fiz, I., Mazzola, F., Fiz, F., Marchi, F., Filauro, M., Paderno, A., Parrinello, G., Piazza, C., & Peretti, G. (2017b). Impact of close and positive margins in transoral laser microsurgery for TIS-T2 glottic cancer. *Frontiers in Oncology*, *7*(OCT), 1–9. https://doi.org/10.3389/fonc.2017.00245

Fritz, C., & Rajasekaran, K. (2023). Delayed in diagnosis of upper aerodigestive tract cancers: A comprehensive review of medical malpractice cases. *Head and Neck*, *45*(7), 1782–1789. https://doi.org/10.1002/hed.27390

Galli, J., Settimi, S., Mele, D. A., Salvati, A., Schiavi, E., Parrilla, C., & Paludetti, G. (2021). Role of Narrow Band Imaging Technology in the Diagnosis and Follow up of Laryngeal Lesions: Assessment of Diagnostic Accuracy and Reliability in a Large Patient Cohort. *Journal of Clinical Medicine*, *10*(6). https://doi.org/10.3390/jcm10061224

Gamez, M. E., Blakaj, A., Zoller, W., Bonomi, M., & Blakaj, D. M. (2020). Emerging concepts and novel strategies in radiation therapy for laryngeal cancer management. *Cancers*, *12*(6), 1–21. https://doi.org/10.3390/cancers12061651

Garofolo, S., Piazza, C., Del Bon, F., Mangili, S., Guastini, L., Mora, F., Nicolai, P., & Peretti, G. (2015). Intraoperative narrow band imaging better delineates superficial resection margins during transoral laser microsurgery for early glottic cancer. *The Annals of Otology, Rhinology, and Laryngology*, *124*(4), 294–298. https://doi.org/10.1177/0003489414556082

Gong, J., Holsinger, F. C., Noel, J. E., Mitani, S., Jopling, J., Bedi, N., Koh, Y. W., Orloff, L. A., Cernea, C. R., & Yeung, S. (2021). Using deep learning to identify the recurrent laryngeal nerve during thyroidectomy. *Scientific Reports*, *11*(1), 14306. https://doi.org/10.1038/s41598-021-93202-y

Gorphe, P., & Simon, C. (2019a). A systematic review and meta-analysis of margins in transoral surgery for oropharyngeal carcinoma. In *Oral Oncology* (Vol. 98, pp. 69–77). Elsevier Ltd. https://doi.org/10.1016/j.oraloncology.2019.09.017

Gorphe, P., & Simon, C. (2019b). A systematic review and meta-analysis of margins in transoral surgery for oropharyngeal carcinoma. *Oral Oncology*, *98*, 69–77. https://doi.org/10.1016/J.ORALONCOLOGY.2019.09.017

Hamad, A., Haney, M., Lever, T. E., & Bunyak, F. (n.d.). *Automated Segmentation of the Vocal Folds in Laryngeal Endoscopy Videos using Deep Convolutional Regression Networks*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-December*, 770–778. https://doi.org/10.1109/CVPR.2016.90

He, Y., Cheng, Y., Huang, Z., Xu, W., Hu, R., Cheng, L., He, S., Yue, C., Qin, G., Wang, Y., & Zhong, Q. (2021). A deep convolutional neural network-based method for laryngeal squamous cell carcinoma diagnosis. *Annals of Translational Medicine*, *9*(24), 1797–1797. https://doi.org/10.21037/atm-21-6458

Heer, J. (2021). Fast & Accurate Gaussian Kernel Density Estimation. *In 2021 IEEE Visualization Conference (VIS)*, 11–15.

Heo, J., Lim, J. H., Lee, H. R., Jang, J. Y., Shin, Y. S., Kim, D., Lim, J. Y., Park, Y. M., Koh, Y. W., Ahn, S. H., Chung, E. J., Lee, D. Y., Seok, J., & Kim, C. H. (2022). Deep learning model

for tongue cancer diagnosis using endoscopic images. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-10287-9

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2020). Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(8), 2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372

Inaba, A., Hori, K., Yoda, Y., Ikematsu, H., Takano, H., Matsuzaki, H., Watanabe, Y., Takeshita, N., Tomioka, T., Ishii, G., Fujii, S., Hayashi, R., & Yano, T. (2020). Artificial intelligence system for detecting superficial laryngopharyngeal cancer with high efficiency of deep learning. *Head and Neck*, *42*(9), 2581–2592. https://doi.org/10.1002/hed.26313

Irem Turkmen, H., Elif Karsligil, M., & Kocak, I. (2015). Classification of laryngeal disorders based on shape and vascular defects of vocal folds. *Computers in Biology and Medicine*, *62*, 76–85. https://doi.org/10.1016/j.compbiomed.2015.02.001

Irjala, H., Matar, N., Remacle, M., & Georges, L. (2011). Pharyngo-laryngeal examination with the narrow band imaging technology: Early experience. In *European Archives of Oto-Rhino-Laryngology* (Vol. 268, Issue 6, pp. 801–806). https://doi.org/10.1007/s00405-011-1516-z

Jha, D., Ali, S., Tomar, N. K., Johansen, H. D., Johansen, D., Rittscher, J., Riegler, M. A., & Halvorsen, P. (2021). Real-Time Polyp Detection, Localization and Segmentation in Colonoscopy Using Deep Learning. *IEEE Access*, *9*, 40496–40510. https://doi.org/10.1109/ACCESS.2021.3063716

Jha, D., Riegler, M. A., Johansen, D., Halvorsen, P., & Johansen, H. D. (2020a). DoubleU-Net: A deep convolutional neural network for medical image segmentation. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, *2020-July*, 558–564. https://doi.org/10.1109/CBMS49503.2020.00111

Jha, D., Riegler, M. A., Johansen, D., Halvorsen, P., & Johansen, H. D. (2020b). DoubleU-Net: A deep convolutional neural network for medical image segmentation. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, *2020-July*, 558–564. https://doi.org/10.1109/CBMS49503.2020.00111

Ji, B., Ren, J., Zheng, X., Tan, C., Ji, R., Zhao, Y., & Liu, K. (2020a). A multi-scale recurrent fully convolution neural network for laryngeal leukoplakia segmentation. *Biomedical Signal Processing and Control*, *59*. https://doi.org/10.1016/j.bspc.2020.101913

Ji, B., Ren, J., Zheng, X., Tan, C., Ji, R., Zhao, Y., & Liu, K. (2020b). A multi-scale recurrent fully convolution neural network for laryngeal leukoplakia segmentation. *Biomedical Signal Processing and Control*, *59*. https://doi.org/10.1016/j.bspc.2020.101913

Ji, B., Ren, J., Zheng, X., Tan, C., Ji, R., Zhao, Y., & Liu, K. (2020c). A multi-scale recurrent fully convolution neural network for laryngeal leukoplakia segmentation. *Biomedical Signal Processing and Control*, *59*. https://doi.org/10.1016/j.bspc.2020.101913

Jia Deng, Wei Dong, Socher, R., Li-Jia Li, Kai Li, & Li Fei-Fei. (2009). *ImageNet: A large-scale hierarchical image database*. 248–255. https://doi.org/10.1109/cvprw.2009.5206848

Khanna, A., Londhe, N. D., Gupta, S., & Semwal, A. (2020). *ScienceDirect A deep Residual U-Net convolutional neural network for automated lung segmentation in computed tomography images*. *40*, 1314–1327. https://doi.org/10.1016/j.bbe.2020.07.007

Kim, G. H., Sung, E. S., & Nam, K. W. (2021). Automated laryngeal mass detection algorithm for home-based self-screening test based on convolutional neural network. *BioMedical Engineering Online*, *20*(1). https://doi.org/10.1186/s12938-021-00886-4

Kudo, S. E., Mori, Y., Abdel-Aal, U. M., Misawa, M., Itoh, H., Oda, M., & Mori, K. (2021). Artificial intelligence and computer-aided diagnosis for colonoscopy: Where do we stand now? In *Translational Gastroenterology and Hepatology* (Vol. 6). AME Publishing Company. https://doi.org/10.21037/tgh.2019.12.14

La Vecchia, C., Lucchini, F., Negri, E., & Levi, F. (2004). Trends in oral cancer mortality in Europe. *Oral Oncology*, *40*(4), 433–439. https://doi.org/10.1016/j.oraloncology.2003.09.013

Laves, M. H., Bicker, J., Kahrs, L. A., & Ortmaier, T. (2019). A dataset of laryngeal endoscopic images with comparative study on convolution neural network-based semantic segmentation. *International Journal of Computer Assisted Radiology and Surgery*, *14*(3), 483–492. https://doi.org/10.1007/s11548-018-01910-0

Li, B., Bobinski, M., Gandour-Edwards, R., Farwell, D. G., & Chen, A. M. (2011). Overstaging of cartilage invasion by multidetector CT scan for laryngeal cancer and its potential effect on the use of organ preservation with chemoradiation. *British Journal of Radiology*, *84*(997), 64–69. https://doi.org/10.1259/bjr/66700901

Li, C., Jing, B., Ke, L., Li, B., Xia, W., He, C., Qian, C., Zhao, C., Mai, H., Chen, M., Cao, K., Mo, H., Guo, L., Chen, Q., Tang, L., Qiu, W., Yu, Y., Liang, H., Huang, X., … Lv, X. (2018). Development and validation of an endoscopic images-based deep learning model for

detection with nasopharyngeal malignancies 08 Information and Computing Sciences 0801 Artificial Intelligence and Image Processing. *Cancer Communications*, *38*(1). https://doi.org/10.1186/s40880-018-0325-9

Li, Y., Gu, W., Yue, H., Lei, G., Guo, W., Wen, Y., Tang, H., Luo, X., Tu, W., Ye, J., Hong, R., Cai, Q., Gu, Q., Liu, T., Miao, B., Wang, R., Ren, J., & Lei, W. (2023). Real-time detection of laryngopharyngeal cancer using an artificial intelligence-assisted system with multimodal data. *Journal of Translational Medicine*, *21*(1). https://doi.org/10.1186/s12967-023-04572-y

Lin, J., Walsted, E. S., Backer, V., Hull, J. H., & Elson, D. S. (2019). Quantification and Analysis of Laryngeal Closure from Endoscopic Videos. *IEEE Transactions on Biomedical Engineering*, *66*(4), 1127–1136. https://doi.org/10.1109/TBME.2018.2867636

Lin, Y. C., Watanabe, A., Chen, W. C., Lee, K. F., Lee, I. L., & Wang, W. H. (2010). Narrowband Imaging for Early Detection of Malignant Tumors and Radiation Effect After Treatment of Head and Neck Cancer. *Archives of Otolaryngology–Head & Neck Surgery*, *136*(3), 234–239. https://doi.org/10.1001/ARCHOTO.2009.230

Madana, J., Lim, C. M., & Loh, K. S. (2015). Narrow band imaging of nasopharynx to identify specific features for possible detection of early nasopharyngeal carcinoma. *Head and Neck*, *37*(8), 1096–1101. https://doi.org/10.1002/hed.23705

Mahendran, N., Vincent, D. R., Srinivasan, K., Chang, C. Y., Garg, A., Gao, L., & Reina, D. G. (2019). Sensor-assisted weighted average ensemble model for detecting major depressive disorder. *Sensors (Switzerland)*, *19*(22). https://doi.org/10.3390/s19224822

Marie Schrynemackers, R. K. (2013). On protocols and measures for the validation of supervised methods for the inference of biological networks. *Frontiers in Genetics*, *4*(DEC). https://doi.org/10.3389/FGENE.2013.00262

Mascharak, S., Baird, B. J., & Holsinger, F. C. (2018). Detecting oropharyngeal carcinoma using multispectral, narrow-band imaging and machine learning. *Laryngoscope*, *128*(11), 2514–2520. https://doi.org/10.1002/lary.27159

Matava, C., Pankiv, E., Raisbeck, S., Caldeira, M., & Alam, F. (2020). A Convolutional Neural Network for Real Time Classification, Identification, and Labelling of Vocal Cord and Tracheal Using Laryngoscopy and Bronchoscopy Video. *Journal of Medical Systems*, *44*(2). https://doi.org/10.1007/s10916-019-1481-4

Mendel, R., Ebigbo, A., Probst, A., Messmann, H., & Palm, C. (2017). Barrett's Esophagus Analysis Using Convolutional Neural Networks. In K. H. Maier-Hein geb. Fritzsche, T. M. Deserno geb. Lehmann, H. Handels, & T. Tolxdorff (Eds.), *Bildverarbeitung für die Medizin 2017* (pp. 80–85). Springer Berlin Heidelberg.

Misra, D. (2019). *Mish: A Self Regularized Non-Monotonic Activation Function*. http://arxiv.org/abs/1908.08681

Moccia, S., De Momi, E., Guarnaschelli, M., Savazzi, M., & Laborai, A. (2017). Confident texture-based laryngeal tissue classification for early stage diagnosis support. *Journal of Medical Imaging*, *4*(03), 1. https://doi.org/10.1117/1.jmi.4.3.034502

Nair, D., Qayyumi, B., Sharin, F., Mair, M., Bal, M., Pimple, S., Mishra, G., Nair, S., & Chaturvedi, P. (2021). Narrow band imaging observed oral mucosa microvasculature as a tool to detect early oral cancer: an Indian experience. *European Archives of Oto-Rhino-Laryngology : Official Journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS) : Affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery*, *278*(10), 3965–3971. https://doi.org/10.1007/s00405-020-06578-4

Ni, X. G., He, S., Xu, Z. G., Gao, L., Lu, N., Yuan, Z., Lai, S. Q., Zhang, Y. M., Yi, J. L., Wang, X. L., Zhang, L., Li, X. Y., & Wang, G. Q. (2011). Endoscopic diagnosis of laryngeal cancer and precancerous lesions by narrow band imaging. *Journal of Laryngology and Otology*, *125*(3), 288–296. https://doi.org/10.1017/S0022215110002033

Nourani, V., Elkiran, G., & Abba, S. I. (2018). Wastewater treatment plant performance analysis using artificial intelligence - An ensemble approach. *Water Science and Technology*, *78*(10), 2064–2076. https://doi.org/10.2166/wst.2018.477

Odell, E., Eckel, H. E., Simo, R., Quer, M., Paleri, V., Klussmann, J. P., Remacle, M., Sjögren, E., & Piazza, C. (2021). European Laryngological Society position paper on laryngeal dysplasia Part I: aetiology and pathological classification. In *European Archives of Oto-Rhino-Laryngology* (Vol. 278, Issue 6, pp. 1717–1722). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s00405-020-06403-y

Paderno, A., Holsinger, F. C., & Piazza, C. (2021). Videomics: bringing deep learning to diagnostic endoscopy. In *Current Opinion in Otolaryngology and Head and Neck Surgery* (Vol. 29, Issue 2, pp. 143–148). Lippincott Williams and Wilkins. https://doi.org/10.1097/MOO.0000000000000697

Paderno, A., Piazza, C., Del Bon, F., Lancini, D., Tanagli, S., Deganello, A., Peretti, G., De Momi, E., Patrini, I., Ruperti, M., Mattos, L. S., & Moccia, S. (2021a). Deep Learning for Automatic Segmentation of Oral and Oropharyngeal Cancer Using Narrow Band Imaging: Preliminary Experience in a Clinical Perspective. *Frontiers in Oncology*, *11*. https://doi.org/10.3389/fonc.2021.626602

Paderno, A., Piazza, C., Del Bon, F., Lancini, D., Tanagli, S., Deganello, A., Peretti, G., De Momi, E., Patrini, I., Ruperti, M., Mattos, L. S., & Moccia, S. (2021b). Deep Learning for Automatic Segmentation of Oral and Oropharyngeal Cancer Using Narrow Band Imaging: Preliminary Experience in a Clinical Perspective. *Frontiers in Oncology*, *11*. https://doi.org/10.3389/fonc.2021.626602

Paderno, A., Villani, F. P., Fior, M., Berretti, G., Gennarini, F., Zigliani, G., Ulaj, E., Montenegro, C., Sordi, A., Sampieri, C., Peretti, G., Moccia, S., & Piazza, C. (2023). Instance segmentation of upper aerodigestive tract cancer: site-specific outcomes. *Acta Otorhinolaryngologica Italica : Organo Ufficiale Della Societa Italiana Di Otorinolaringologia e Chirurgia Cervico-Facciale*, *43*(4), 283–290. https://doi.org/10.14639/0392-100X-N2336

Pan, X., Bai, W., Ma, M., & Zhang, S. (2022). RANT: A cascade reverse attention segmentation framework with hybrid transformer for laryngeal endoscope images. *Biomedical Signal Processing and Control*, *78*. https://doi.org/10.1016/j.bspc.2022.103890

Parker, F., Brodsky, M. B., Akst, L. M., & Ali, H. (2021). Machine Learning in Laryngoscopy Analysis: A Proof of Concept Observational Study for the Identification of Post-Extubation Ulcerations and Granulomas. *Annals of Otology, Rhinology and Laryngology*, *130*(3), 286–291. https://doi.org/10.1177/0003489420950364

Pham, V. T., Tran, C. M., Zheng, S., Vu, T. M., & Nath, S. (2021). Chest X-ray abnormalities localization via ensemble of deep convolutional neural networks. *International Conference on Advanced Technologies for Communications*, *2021-Octob*, 125–130. https://doi.org/10.1109/ATC52653.2021.9598342

Piazza, C., Bon, F. D., Peretti, G., & Nicolai, P. (2011). “Biologic endoscopy”: Optimization of upper aerodigestive tract cancer evaluation. In *Current Opinion in Otolaryngology and Head and Neck Surgery* (Vol. 19, Issue 2, pp. 67–76). https://doi.org/10.1097/MOO.0b013e328344b3ed

Piazza, C., Cocco, D., De Benedetto, L., Del Bon, F., Nicolai, P., & Peretti, G. (2010a). Narrow band imaging and high definition television in the assessment of laryngeal cancer: A prospective study on 279 patients. *European Archives of Oto-Rhino-Laryngology*, *267*(3), 409–414. https://doi.org/10.1007/s00405-009-1121-6

Piazza, C., Cocco, D., De Benedetto, L., Del Bon, F., Nicolai, P., & Peretti, G. (2010b). Narrow band imaging and high definition television in the assessment of laryngeal cancer: a prospective study on 279 patients. *European Archives of Oto-Rhino-Laryngology*, *267*(3), 409–414. https://doi.org/10.1007/S00405-009-1121-6

Piazza, C., Cocco, D., Del Bon, F., Mangili, S., Nicolai, P., Majorana, A., Bolzoni Villaret, A., & Peretti, G. (2010). Narrow band imaging and high definition television in evaluation of oral and oropharyngeal squamous cell cancer: A prospective study. *Oral Oncology*, *46*(4), 307–310. https://doi.org/10.1016/j.oraloncology.2010.01.020

Qiu, B., van der Wel, H., Kraeima, J., Glas, H. H., Guo, J., Borra, R. J. H., Witjes, M. J. H., & van Ooijen, P. M. A. (2021). Automatic Segmentation of Mandible from Conventional Methods to Deep Learning-A Review. *Journal of Personalized Medicine*, *11*(7). https://doi.org/10.3390/jpm11070629

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. *New England Journal of Medicine*, *380*(14), 1347–1358. https://doi.org/10.1056/nejmra1814259

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 779–788. https://doi.org/10.1109/CVPR.2016.91

Ren, J., Jing, X., Wang, J., Ren, X., Xu, Y., Yang, Q., Ma, L., Sun, Y., Xu, W., Yang, N., Zou, J., Zheng, Y., Chen, M., Gan, W., Xiang, T., An, J., Liu, R., Lv, C., Lin, K., … Zhao, Y. (2020). Automatic Recognition of Laryngoscopic Images Using a Deep-Learning Technique. *Laryngoscope*, *130*(11), E686–E693. https://doi.org/10.1002/lary.28539

Ronneberger O, Fischer P, B. T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention 2015*, 234–241. https://doi.org/10.1007/978-3-319-24574-4

Rosenthal, E. L. (2014). Optical Imaging of Head and Neck Cancer: Opportunities and Challenges. *JAMA Otolaryngology–Head & Neck Surgery*, *140*(2), 93–94. https://doi.org/10.1001/JAMAOTO.2013.6166

Saito, Y., Oka, S., Kawamura, T., Shimoda, R., Sekiguchi, M., Tamai, N., Hotta, K., Matsuda, T., Misawa, M., Tanaka, S., Iriguchi, Y., Nozaki, R., Yamamoto, H., Yoshida, M., Fujimoto, K., & Inoue, H. (2021). Colonoscopy screening and surveillance guidelines. *Digestive Endoscopy : Official Journal of the Japan Gastroenterological Endoscopy Society*, *33*(4), 486–519. https://doi.org/10.1111/den.13972

Sampieri, C., Baldini, C., Azam, M. A., Moccia, S., Mattos, L. S., Vilaseca, I., Peretti, G., & Ioppi, A. (2023). State of the Art Review Artificial Intelligence for Upper Aerodigestive Tract Endoscopy and Laryngoscopy: A Guide for Physicians and State-of-the-Art Review. *Otolaryngology-Head and Neck Surgery*, *2023*(00), 1–19. https://doi.org/10.1002/ohn.343

Sampieri, C., Costantino, A., Spriano, G., Peretti, G., De Virgilio, A., & Kim, S. H. (2022a). Role of surgical margins in transoral robotic surgery: A question yet to be answered. In *Oral Oncology* (Vol. 133). Elsevier Ltd. https://doi.org/10.1016/j.oraloncology.2022.106043

Sampieri, C., Costantino, A., Spriano, G., Peretti, G., De Virgilio, A., & Kim, S.-H. (2022b). Role of surgical margins in transoral robotic surgery: A question yet to be answered. *Oral Oncology*, *133*, 106043. https://doi.org/10.1016/J.ORALONCOLOGY.2022.106043

Scholman, C., Westra, J. M., Zwakenberg, M. A., Dikkers, F. G., Halmos, G. B., Wedman, J., Wachters, J. E., van der Laan, B. F. A. M., & Plaat, B. E. C. (2020). Differences in the diagnostic value between fiberoptic and high definition laryngoscopy for the characterisation of pharyngeal and laryngeal lesions: A multi-observer paired analysis of videos. *Clinical Otolaryngology*, *45*(1), 119–125. https://doi.org/10.1111/coa.13476

Shephard, A. J., Mahmood, H., Raza, S. E. A., Araujo, A. L. D., Santos-Silva, A. R., Lopes, M. A., Vargas, P. A., McCombe, K., Craig, S., James, J., Brooks, J., Nankivell, P., Mehanna, H., Khurram, S. A., & Rajpoot, N. M. (2023). *Transformer-based Model for Oral Epithelial Dysplasia Segmentation*. http://arxiv.org/abs/2311.05452

Tama, B. A., Kim, D. H., Kim, G., Kim, S. W., & Lee, S. (2020). Recent advances in the application of artificial intelligence in otorhinolaryngology-head and neck surgery. In *Clinical and Experimental Otorhinolaryngology* (Vol. 13, Issue 4, pp. 326–339). Korean Society of Otolaryngology. https://doi.org/10.21053/ceo.2020.00654

Tamashiro, A., Yoshio, T., Ishiyama, A., Tsuchida, T., Hijikata, K., Yoshimizu, S., Horiuchi, Y., Hirasawa, T., Seto, A., Sasaki, T., Fujisaki, J., & Tada, T. (2020). Artificial intelligence-

based detection of pharyngeal cancer using convolutional neural networks. *Digestive Endoscopy*, *32*(7), 1057–1065. https://doi.org/10.1111/den.13653

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *36th International Conference on Machine Learning, ICML 2019*, *2019-June*, 10691–10700.

Tan, M., & Le, Q. V. (2021). *EfficientNetV2 : Smaller Models and Faster Training*.

Valls-Mateus, M., Nogués-Sabaté, A., Blanch, J. L., Bernal-Sprekelsen, M., Avilés-Jurado, F. X., & Vilaseca, I. (2018). Narrow band imaging for head and neck malignancies: Lessons learned from mistakes. *Head and Neck*, *40*(6), 1164–1173. https://doi.org/10.1002/HED.25088

Van Der Sommen, F., De Groof, J., Struyvenberg, M., Van Der Putten, J., Boers, T., Fockens, K., Schoon, E. J., Curvers, W., De With, P., Mori, Y., Byrne, M., & Bergman, J. J. G. H. M. (2020). Machine learning in GI endoscopy: Practical guidance in how to interpret a novel field. *Gut*, *69*(11), 2035–2045. https://doi.org/10.1136/gutjnl-2019-320466

Vilaseca, I., Valls-Mateus, M., Nogués, A., Lehrer, E., López-Chacón, M., Avilés-Jurado, F. X., Blanch, J. L., & Bernal-Sprekelsen, M. (2017a). 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 and Neck*, *39*(9), 1854–1863. https://doi.org/10.1002/hed.24849

Vilaseca, I., Valls-Mateus, M., Nogués, A., Lehrer, E., López-Chacón, M., Avilés-Jurado, F. X., Blanch, J. L., & Bernal-Sprekelsen, M. (2017b). 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*, *39*(9), 1854–1863. https://doi.org/10.1002/HED.24849

Vlantis, A. C., Wong, E. W. Y., Ng, S. K., Chan, J. Y. K., & Tong, M. C. F. (2019). Narrow Band Imaging Endoscopy of the Nasopharynx for Malignancy: An Inter- and Intraobserver Study. *Laryngoscope*, *129*(6), 1374–1379. https://doi.org/10.1002/lary.27483

Vlantis, A. C., Woo, J. K. S., Tong, M. C. F., King, A. D., Goggins, W., & van Hasselt, C. A. (2016). Narrow band imaging endoscopy of the nasopharynx is not more useful than white light endoscopy for suspected nasopharyngeal carcinoma. *European Archives of Oto-Rhino-Laryngology*, *273*(10), 3363–3369. https://doi.org/10.1007/s00405-016-3940-6

Wallner, J., Schwaiger, M., Hochegger, K., Gsaxner, C., Zemann, W., & Egger, J. (2019). A review on multiplatform evaluations of semi-automatic open-source based image segmentation for cranio-maxillofacial surgery. *Computer Methods and Programs in Biomedicine*, *182*, 105102. https://doi.org/10.1016/j.cmpb.2019.105102

Wang, C. Y., Mark Liao, H. Y., Wu, Y. H., Chen, P. Y., Hsieh, J. W., & Yeh, I. H. (2020). CSPNet: A new backbone that can enhance the learning capability of CNN. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, *2020-June*, 1571–1580. https://doi.org/10.1109/CVPRW50498.2020.00203

Wang, P., Xiao, X., Glissen Brown, J. R., Berzin, T. M., Tu, M., Xiong, F., Hu, X., Liu, P., Song, Y., Zhang, D., Yang, X., Li, L., He, J., Yi, X., Liu, J., & Liu, X. (n.d.). Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nature Biomedical Engineering*. https://doi.org/10.1038/s41551-018-0301-3

Wang, P., Xiao, X., Glissen Brown, J. R., Berzin, T. M., Tu, M., Xiong, F., Hu, X., Liu, P., Song, Y., Zhang, D., Yang, X., Li, L., He, J., Yi, X., Liu, J., & Liu, X. (2018). Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nature Biomedical Engineering*, *2*(10), 741–748. https://doi.org/10.1038/s41551-018-0301-3

Wang, W. H., Lin, Y. C., Chen, W. C., Chen, M. F., Chen, C. C., & Lee, K. F. (2012). Detection of mucosal recurrent nasopharyngeal carcinomas after radiotherapy with narrow-band imaging endoscopy. *International Journal of Radiation Oncology Biology Physics*, *83*(4), 1213–1219. https://doi.org/10.1016/j.ijrobp.2011.09.034

Wang, Y. Y., Hamad, A. S., Palaniappan, K., Lever, T. E., & Bunyak, F. (2022). LARNet-STC: Spatio-temporal orthogonal region selection network for laryngeal closure detection in endoscopy videos. *Computers in Biology and Medicine*, *144*. https://doi.org/10.1016/j.compbiomed.2022.105339

Winawer, S. J., Zauber, A. G., Fletcher, R. H., Stillman, J. S., O'brien, M. J., Levin, B., Smith, R. A., Lieberman, D. A., Burt, R. W., Levin, T. R., Bond, J. H., Brooks, D., Byers, T., Hyman, N., Kirk, L., Thorson, A., Simmang, C., Johnson, D., & Rex, D. K. (2006). Guidelines for colonoscopy surveillance after polypectomy: a consensus update by the US Multi-Society Task Force on Colorectal Cancer and the American Cancer Society. *CA: A Cancer Journal for Clinicians*, *56*(3), 143–145. https://doi.org/10.3322/canjclin.56.3.143

Witt, D. R., Chen, H., Mielens, J. D., McAvoy, K. E., Zhang, F., Hoffman, M. R., & Jiang, J. J. (2014). Detection of chronic laryngitis due to laryngopharyngeal reflux using color and

texture analysis of laryngoscopic images. *Journal of Voice*, *28*(1), 98–105. https://doi.org/10.1016/j.jvoice.2013.08.015

Woo, S., Park, J., Lee, J., & Kweon, I. S. (2018). Convolutional_Block_Attention. *Eccv*, 17. http://files/737/Woo et al. - Convolutional Block Attention Module.pdf

Wu, Z., Ge, R., Wen, M., Liu, G., Chen, Y., Zhang, P., He, X., Hua, J., Luo, L., & Li, S. (2021). ELNet:Automatic classification and segmentation for esophageal lesions using convolutional neural network. *Medical Image Analysis*, *67*, 101838. https://doi.org/10.1016/j.media.2020.101838

Xception, C. F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. *InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017*, 1251–1258. https://doi.org/10.4271/2014-01-0975

Xiong, H., Lin, P., Yu, J. G., Ye, J., Xiao, L., Tao, Y., Jiang, Z., Lin, W., Liu, M., Xu, J., Hu, W., Lu, Y., Liu, H., Li, Y., Zheng, Y., & Yang, H. (2019). Computer-aided diagnosis of laryngeal cancer via deep learning based on laryngoscopic images. *EBioMedicine*, *48*, 92–99. https://doi.org/10.1016/j.ebiom.2019.08.075

Xu, J., Wang, J., Bian, X., Zhu, J. Q., Tie, C. W., Liu, X., Zhou, Z., Ni, X. G., & Qian, D. (2022). Deep Learning for nasopharyngeal Carcinoma Identification Using Both White Light and Narrow-Band Imaging Endoscopy. *Laryngoscope*, *132*(5), 999–1007. https://doi.org/10.1002/lary.29894

Yadav, G., Maheshwari, S., & Agarwal, A. (2014). Contrast limited adaptive histogram equalization-based enhancement for real time video system. *Proceedings of the 2014 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2014*, 2392–2397. https://doi.org/10.1109/ICACCI.2014.6968381

Yang, M. (n.d.). *DenseASPP for Semantic Segmentation in Street Scenes*.

Yang, X. X., Li, Z., Shao, X. J., Ji, R., Qu, J. Y., Zheng, M. Q., Sun, Y. N., Zhou, R. C., You, H., Li, L. X., Feng, J., Yang, X. Y., Li, Y. Q., & Zuo, X. L. (2021a). Real-time artificial intelligence for endoscopic diagnosis of early esophageal squamous cell cancer (with video). *Digestive Endoscopy*, *33*(7), 1075–1084. https://doi.org/10.1111/DEN.13908

Yang, X. X., Li, Z., Shao, X. J., Ji, R., Qu, J. Y., Zheng, M. Q., Sun, Y. N., Zhou, R. C., You, H., Li, L. X., Feng, J., Yang, X. Y., Li, Y. Q., & Zuo, X. L. (2021b). Real-time artificial intelligence for endoscopic diagnosis of early esophageal squamous cell cancer (with video). *Digestive Endoscopy*, *33*(7), 1075–1084. https://doi.org/10.1111/den.13908

Yao, P., Usman, M., Chen, Y. H., German, A., Andreadis, K., Mages, K., & Rameau, A. (2022). Applications of Artificial Intelligence to Office Laryngoscopy: A Scoping Review. *Laryngoscope*, *132*(10), 1993–2016. https://doi.org/10.1002/lary.29886

Yap, M. H., Hachiuma, R., Alavi, A., Brungel, R., Cassidy, B., Goyal, M., Zhu, H., Ruckert, J., Olshansky, M., Huang, X., Saito, H., Hassanpour, S., Friedrich, C. M., Ascher, D., Song, A., Kajita, H., Gillespie, D., Reeves, N. D., Pappachan, J., … Frank, E. (2020). Deep Learning in Diabetic Foot Ulcers Detection: A Comprehensive Evaluation. *Computers in Biology and Medicine*, 104596. https://doi.org/10.1016/j.compbiomed.2021.104596

Ye, Z., Saraf, A., Ravipati, Y., Hoebers, F., Zha, Y., Zapaishchykova, A., Likitlersuang, J., Tishler, R. B., Schoenfeld, J. D., Margalit, D. N., Haddad, R. I., Mak, R. H., Naser, M., Wahid, K. A., Sahlsten, J., Jaskari, J., Kaski, K., Mäkitie, A. A., Fuller, C. D., … Kann, B. H. (n.d.). *Fully automated sarcopenia assessment in head and neck cancer: development and external validation of a deep learning pipeline*. https://doi.org/10.1101/2023.03.01.23286638

Yue, G., Zhuo, G., Li, S., Zhou, T., Du, J., Yan, W., Hou, J., Liu, W., & Wang, T. (2023). Benchmarking Polyp Segmentation Methods in Narrow-Band Imaging Colonoscopy Images. *IEEE Journal of Biomedical and Health Informatics*. https://doi.org/10.1109/JBHI.2023.3270724

Zafar, A., Aamir, M., Nawi, N. M., Arshad, A., Riaz, S., Alruban, A., Dutta, A. K., & Almotairi, S. (2022). *Applied sciences A Comparison of Pooling Methods for Convolutional Neural Networks*. 1–21.

Zhang, Q., Wang, H., Zhao, Q., Zhang, Y., Zheng, Z., Liu, S., Liu, Z., Meng, L., Xin, Y., & Jiang, X. (2021). Evaluation of Risk Factors for Laryngeal Squamous Cell Carcinoma: A Single-Center Retrospective Study. *Frontiers in Oncology*, *11*. https://doi.org/10.3389/fonc.2021.606010

Zhang, Z., Liu, Q., & Wang, Y. (2018). Road Extraction by Deep Residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, *15*(5), 749–753. https://doi.org/10.1109/LGRS.2018.2802944

Zhao, Q., He, Y., Wu, Y., Huang, D., Wang, Y., Sun, C., Ju, J., Wang, J., & Mahr, J. J. li. (2022). Vocal cord lesions classification based on deep convolutional neural network and transfer learning. *Medical Physics*, *49*(1), 432–442. https://doi.org/10.1002/mp.15371

Zhao, R., Qian, B., Zhang, X., Li, Y., Wei, R., Liu, Y., & Pan, Y. (2020). Rethinking dice loss for medical image segmentation. *Proceedings - IEEE International Conference on Data Mining, ICDM*, *2020-November*, 851–860. https://doi.org/10.1109/ICDM50108.2020.00094

Zhong, Y., Yang, Y., Fang, Y., Wang, J., & Hu, W. (2021). A Preliminary Experience of Implementing Deep-Learning Based Auto-Segmentation in Head and Neck Cancer: A Study on Real-World Clinical Cases. *Frontiers in Oncology*, *11*, 638197. https://doi.org/10.3389/fonc.2021.638197

Zwakenberg, M. A., Westra, J. M., Halmos, G. B., Wedman, J., van der Laan, B. F. A. M., & Plaat, B. E. C. (2023). Narrow-Band Imaging in Transoral Laser Surgery for Early Glottic Cancer: A Randomized Controlled Trial. *Otolaryngology - Head and Neck Surgery (United States)*. https://doi.org/10.1002/OHN.307