UNIVERSITY OF GENOVA

PHD PROGRAM IN BIOENGINEERING AND ROBOTICS

Toward AI-Assisted Laryngology: Real-Time Lesion Screening, Risk Classification, and Tumor Margin Assessment

by

Chiara Baldini

Thesis submitted for the degree of *Doctor of Philosophy* (38° cycle)

February 2026

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

Thesis Jury:

Prof. Stefanie Speidel, *NCT Dresden*	External examiner
Dr. Stamatia Giannarou , *Imperial College London*	External examiner

Dibris

Department of Informatics, Bioengineering, Robotics and Systems Engineering

iit ADVR ISTITUTO ITALIANO DI TECNOLOGIA

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

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Chiara Baldini
February 2026

Acknowledgements

I would like to express my deepest gratitude to the many people who supported me throughout this research.

First and foremost, I thank my supervisor, Dr. Leonardo De Mattos, for his invaluable mentorship, guidance, and constant encouragement throughout this project, and for giving me the opportunity to pursue this stimulating PhD journey.

I would also like to thank my reviewers, Prof. Stefanie Speidel and Dr. Stamatia Giannarou, for their time, insightful feedback, and thoughtful suggestions, which greatly improved the quality of this work.

I am especially grateful to Dr. Sara Moccia, Dr. Lucia Migliorelli, and Dr. Veronica Penza for their technical guidance, stimulating discussions, and their constant presence and support throughout this work.

I would like to thank Claudio Sampieri, Giorgio Peretti, Francesco Mora, and all the other clinical collaborators for the constructive exchange of perspectives between clinicians and engineers, for their patience in reviewing my results, and for their invaluable contributions to data collection, curation, and endoscopy domain knowledge.

A special thank you goes to Kaisar, Richard, Oliver, Simone, Karim, and all the colleagues of the BCN AIM Lab for the enriching experience abroad, the inspiring work environment, and the personal and professional growth made possible by this collaboration.

I am also thankful to the Istituto Italiano di Tecnologia for the institutional support, and to all members of the Biomedical Robotics team and other colleagues at IIT for their collaboration, assistance, and for being by my side throughout this journey.

Finally, I wish to thank my family and friends for their patience and encouragement, which have been a constant source of strength and motivation.

Abstract

Objective: Disorders of the upper aerodigestive tract, especially laryngeal heterogeneous lesions, present key clinical challenges for early diagnosis and therapeutic planning. Endoscopic examination is the current gold standard; however, operator-related factors and image quality have a significant impact on diagnostic accuracy. Moreover, intraoperative identification of tumor margins is often imprecise, leading to incomplete resections and increased recurrence risk.

This research introduces a comprehensive Artificial Intelligence (AI)-assisted vision system designed to support clinicians in laryngeal lesion screening, diagnosis, and surgical margin assessment. The ultimate goal is to enhance the accuracy, speed, and objectivity of laryngeal cancer diagnosis and treatment planning, while reducing operator dependence and improving clinical workflow efficiency.

Approach: The proposed framework combines multiple AI modules to autonomously process and interpret endoscopic data collected in standard clinical settings during laryngeal examinations, both in-office and intraoperatively. Deep Learning (DL) approaches were utilized for addressing four main tasks:

(i) *Informative frame selection* to identify diagnostically relevant frames in endoscopic videos;

(ii) *Lesion detection* using YOLO-based architectures, eventually enhanced with a super-resolution branch (SRE-YOLO);

(iii) *Binary low or high risk classification and multi-class lesion discrimination* through convolutional and transformer-based networks;

(iv) *Laryngeal tumor margin segmentation* taking advantage of foundation models such as the Segment Anything Model (SAM).

A Latent Diffusion Model (LDM) was also designed to generate clinically guided synthetic laryngeal data, while a custom user-friendly graphical user interface (GUI) was

implemented to integrate all AI modules for real-time clinical use and validation at the Unit of Otolaryngology and Head and Neck Surgery of the IRCCS San Martino Hospital (Genova, Italy). A large multicenter White-Light (WL) and Narrow Band Imaging (NBI) endoscopic dataset was used to train the models, and external validation, both retrospective and prospective, was carried out with data from many international institutions.

Main results: The framework achieved high and robust performance across the four different tasks:

(i) The informative frame classifier reached F1-scores of 93-96% with real-time performance (<0.03 seconds/frame);

(ii) The proposed detection model (SRE-YOLO) achieved a mean Average Precision$_{@\text{Intersection over Union = 0.5}}$ ($\text{mAP}_{@\text{IoU=0.5}}$) > 80% internally and externally, with substantial performance enhancements achieved through the use of synthetic data augmentation.

(iii) To classify lesions, the transformer-based model reached F1-scores of 85–89% for *Low-risk*/*High-risk* differentiation and up to 76% for multiclass lesion type classification.

(iv) Segmentation experiments with finetuned SAM obtained a Dice Similarity Coefficient (DSC) above 90%, closely matching expert annotations.

Integration into the VERA prototype enabled validation on > 100 patients, achieving 85% F1-score and strong concordance with clinicians in lesion risk prediction at the patient-level.

Significance: This work introduces one of the first comprehensive AI pipelines for real-time laryngeal assessment, spanning from image quality evaluation to intraoperative tumor margin definition. Its clinical validation demonstrated that AI can provide objective, reproducible support to clinicians with different levels of expertise in both outpatient and surgical settings. Future directions include releasing the collected anonymized datasets to the public to ensure reproducibility and further research, incorporation of patient metadata for personalized treatment suggestions, models' optimization, and large-scale clinical validation for accelerating the adoption of AI in laryngology.

Key-words: Deep Learning; Laryngeal Endoscopy; Lesion Detection; Segmentation; Data Synthesis; Clinical Validation

Table of contents

List of figures **ix**

List of tables **xiii**

Nomenclature **xvi**

1 Introduction **1**

1.1 Clinical context 1

1.1.1 Anatomy and Physiology of the Larynx 1

1.1.2 Epidemiology and Clinical Impact of Laryngeal Pathologies 2

1.2 Current Clinical Workflow and Related Challenges 3

1.2.1 In-office Laryngeal Evaluation 3

1.2.2 Intraoperative Assessment and Surgical Planning 5

1.3 Artificial Intelligence in Laryngeal Lesion Screening and Treatment 6

1.3.1 Data Scarcity and Limited Generalizability 7

1.4 Aim of the Thesis 8

1.5 Proposed Innovation 9

1.6 Structure of the Thesis 10

2 Literature Review **12**

2.1 Informative Frame Extraction 13

2.2 Laryngeal Lesion Detection 15

2.3 Lesion Characterization 18

2.4 Tumor Margin Delineation 22

2.5 Synthetic Data Generation 26

3 Data Collection and Curation **28**

3.1 Datasets for Frame Quality Assessment 30

3.2 Datasets for Lesion Detection . . . 30
3.3 Datasets for Lesion Characterization . . . 32
3.4 Datasets for Tumor Margin Definition . . . 33

4 Informative Frames Extraction 36
4.1 Overview . . . 36
4.2 Materials and Methods . . . 37
4.2.1 Dataset . . . 37
4.2.2 Architecture and Implementation Details . . . 37
4.3 Experimental Protocol . . . 38
4.3.1 Training Settings . . . 38
4.3.2 Performance Evaluation . . . 38
4.4 Results . . . 40
4.5 Discussion . . . 41

5 Laryngeal Lesion Detection 43
5.1 Overview . . . 43
5.2 Materials and Methods . . . 44
5.2.1 Dataset . . . 44
5.2.2 YOLOv8: Architecture Overview . . . 45
5.2.3 Super-Resolution Branch Integration . . . 46
5.3 Experimental Protocol . . . 48
5.3.1 YOLO Architecture Benchmarking . . . 48
5.3.2 Super Resolution Integration . . . 49
5.3.3 *Full Dataset* Experiments . . . 51
5.3.4 Data Pre-processing . . . 51
5.3.5 Training Settings . . . 52
5.3.6 Performance Metrics . . . 52
5.4 Results . . . 53
5.4.1 YOLO Architecture Benchmarking . . . 53
5.4.2 Super Resolution Integration . . . 54
5.4.3 *Full Dataset* Experiments . . . 58
5.5 Discussion . . . 58

6 Lesion Characterization 61
6.1 Overview . . . 61

6.2 Materials and Methods . . . 62
6.2.1 Dataset . . . 62
6.2.2 Architectures . . . 62
6.3 Experimental Protocol . . . 66
6.3.1 Data Preprocessing . . . 66
6.3.2 Training Settings . . . 66
6.3.3 Performance Metrics . . . 67
6.4 Results . . . 68
6.4.1 Binary Risk Assessment: Performance Evaluation . . . 68
6.4.2 Histological Type Classification: Performance Evaluation . . . 69
6.4.3 Expert-AI Comparative Evaluation . . . 70
6.5 Discussion . . . 71

7 Tumor Margin Definition 74
7.1 Overview . . . 74
7.2 Materials and Methods . . . 75
7.2.1 Dataset . . . 75
7.2.2 Architectures Comparison . . . 75
7.3 Experimental Protocol . . . 79
7.3.1 Data Pre-processing . . . 79
7.3.2 Training Settings . . . 79
7.3.3 Performance Metrics . . . 81
7.4 Results . . . 82
7.5 Discussion . . . 83

8 Synthetic Data Generation 86
8.1 Overview . . . 86
8.2 Materials and Methods . . . 87
8.2.1 Datasets . . . 87
8.2.2 Clinically-guided Latent Diffusion Framework . . . 88
8.3 Experimental Protocol . . . 90
8.3.1 Data preparation . . . 90
8.3.2 Training settings . . . 90
8.3.3 Evaluation Metrics . . . 90
8.4 Results . . . 92
8.4.1 Qualitative Generation Results . . . 92

8.4.2 Impact of Synthetic Data on Lesion Detection 93
8.4.3 Impact of Synthetic Data on Histological Classification 94
8.4.4 Expert-Observer Realism Study 94
8.5 Discussion 94

9 Preliminary Clinical Validation 99
9.1 Overview 99
9.2 Prototyping and GUI Development 100
9.3 Clinical Study Protocol 102
9.4 Data Analysis 103
9.4.1 Lesion Localization Performance 103
9.4.2 *Low-risk* vs *High-risk* Classification Performance 103
9.5 Results 104
9.6 Discussion 106

10 Conclusions and Open Challenges 109
10.1 Thesis Contributions 109
10.1.1 Summary of Key Results 109
10.1.2 Comparison with the State of the Art 110
10.2 Impact of the Proposed Research 112
10.2.1 Scientific Impact 112
10.2.2 Clinical Impact 112
10.2.3 Economical Impact 112
10.2.4 Translational and Social Impact 113
10.3 Limitations and Future Work 113

References 115

List of figures

1.1 Overview of laryngeal diseases and corresponding clinical workflow: on the left, a schematic representation of the Upper Aero-Digestive Tract (UADT) highlights the anatomical position of the larynx. At the bottom, examples of endoscopic views of a normal larynx (both closed and open glottis) and a laryngeal carcinoma under White-Light (WL) and Narrow-Band Imaging (NBI) modalities are illustrated. Flexible endoscopy enables non-invasive in-office visualization but still exhibits a high miss rate, with less than 50% of cases detected at localized stages. In contrast, Trans-Oral Laser Microsurgery (TOLMS), the preferred approach for the treatment of early-stage tumors, offers organ preservation but suffers from a high rate of positive margins. . 5

1.2 Overview of the thesis structure, mapping the main chapters to their corresponding steps within the clinical workflow. Six main chapters are proposed, each addressing a specific aspect of the research conducted during this PhD project, from data collection and curation to AI modules development and clinical validation. 11

3.1 Representative sample images from eight laryngeal endoscopy datasets (Genova-Larynx, Barcelona-Larynx, Lucknow-Larynx, Laryngoscope8, Athens-Larynx, Istanbul-Larynx, Seoul-Larynx, and H&NANCE-Larynx) illustrating the diversity in imaging conditions, viewpoints, and pathological appearances across datasets. 29

4.1 Architecture of the ResNet-50 model employed for informative frame classification. The final fully connected layer was modified using a softmax activation function to output the probabilities of the two classes: *Informative (I)* and *Uninformative (U)*. 39

4.2 (a) Examples of frame predictions from the *external test*. The first column shows the original images along with their corresponding ground-truth labels. In the second column, the prediction is output as a green frame for *informative (I)* images and as a red frame for *uninformative (U)* cases. In the third column, the attention maps used by the model for the prediction are output according to the gradCAM algorithm. (b) Informativeness trends over time and the corresponding final VIL values are illustrated for the 4 tested videos. (c) Model performance according to the Receiver Operating Characteristics (ROC) curves for the internal (blue) and external (orange) test sets. 42

5.1 Architecture of the proposed SRE-YOLO model. The Super-Resolution (SR) branch is highlighted in blue. During inference, the SR branch is removed to maintain the real-time performance of the YOLO baseline. 47

5.2 (a) Scatter plots of the efficacy-efficiency performance from the baselines comparison (i.e., YOLOv5n, YOLOv5s, YOLOv8s, YOLOv8n) against the proposed SRE-YOLO. The plot above shows the Frame Per Second (FPS) against the mean Average Precision (mAP) computed using an Intersection over Union (IoU) threshold of 0.50 ($mAP_{@IoU=0.5}$). The graph below presents the billion floating point operations (GFLOPs) against the $mAP_{@IoU=0.5}$. The $mAP_{@IoU=0.5}$ of SRE-YOLO improved significantly, without causing an increase in GFLOPS or FPS metrics. (b) Samples for qualitative analysis. Row 1-2: Left and right vocal fold polyps. Row 3: Bilateral vocal fold polyps. Row 4-5: Intraoperative Narrow-Band Imaging (NBI) of a left vocal fold polyp and the corresponding White Light (WL)-imaging view. Row 6-7: Squamous Cell Carcinoma (SCC) captured with both WL and NBI modalities. Row 8: Left vocal fold polyp. Row 9: NBI view of a SCC arising from the left laryngeal vestibular fold. Row 10: Model failure case in detecting a left vocal fold polyp. The second and third columns illustrate the detection performance of the YOLOv8n (i.e., the baseline) and the proposed SRE-YOLO, respectively. 57

6.1 EfficientNetV2 architecture overview. The model consists of a series of Fused-MBConv and MBConv blocks, followed by a final classification head. The architecture is designed to optimize both accuracy and training speed through the use of fused operations and progressive scaling. 63

6.2 ViTb16 architecture overview: input image is split into 16×16 patches, embedded and enriched with positional encoding. The resulting token sequence is processed through multiple layers of multi-head self-attention and a feed-forward network. The final representation of the class token is passed to a linear classification head to predict one of the seven histological classes. . . 65
6.3 Confusion matrices for the EfficientNetV2S model when tested for *Low-risk* vs *High-risk* lesion classification on the *internal test set* (left), *external set 1* (center), and *external set 2* (right). 69
6.4 Confusion matrices of the ViTb16 model when tested for seven-class histological lesion classification on the *internal test set* (left), *external set 1* (center), and *external set 2* (right). 71
6.5 (a) ROC curves for clinicians at different levels of expertise, ChatGPT-4o, and the proposed AI model when discriminating between *Low-risk* and *High-risk* lesions on the *Barcelona-Larynx* dataset. (b) Qualitative results of the ViTb16 model for seven-class histological lesion classification on sample images from the *internal* and *external test sets*. The ground-truth and predicted labels are indicated in the last two columns, respectively. The first two rows show misclassifications (red), while the last two rows illustrate correct predictions (green). In the third column, the attention maps highlight the regions of the images that the model focused on for its predictions. . . . 72

7.1 Schematic representation of the Segment Anything Model (SAM) architecture, adapted from (Kirillov et al., 2023). The model consists of three main components: an image encoder based on Vision Transformer (ViT), a prompt encoder that processes user-provided prompts (points, boxes, etc.), and a mask decoder that generates segmentation masks based on the combined information from the image and prompt encoders. 77
7.2 Qualitative segmentation results on representative cases from the *H&NANCE-Larynx* and *Seoul-Larynx* test sets. For each category of conventional, prompt-based, and prompt-less foundation models, segmentation outputs of the best-performing tested methodologies are shown. From left to right: original image with ground truth mask, *SegMENT-Plus* from (Sampieri et al., 2024), *MedSAM* (fine-tuned), and *AutoSAMUS*. 84

8.1 Overview of the proposed clinically-guided generative framework. During training (white arrows), captions containing optical modality and lesion type, together with bounding box masks, are used to condition a Latent Diffusion Model (LDM) via ControlNet. During inference (red arrows), random noise, captions, and masks are combined to generate synthetic laryngeal endoscopic images with a realistic appearance and controllable lesion location. 88

8.2 Qualitative examples of synthetic data generation. Starting from real images (first column), the corresponding captions (second column) and bounding box masks (third column) were used to generate synthetic laryngeal images (fourth column). The model reproduced different lesion types (e.g., squamous cell carcinoma, cyst, leukoplakia) and optical modalities (WL and NBI), while preserving anatomical plausibility. 93

8.3 (a) Evolution of $\text{mAP}_{@\text{IoU=0.5}}$ on the *internal generation* and *external detection test* sets as a function of the percentage of synthetic images added to the real training set (random selection, fold 1). Performance peaks at +10% synthetic data and then reaches a plateau. (b) Human-observer study: responses from otorhinolaryngologists assessing realism. The marker location denotes whether the image was correctly (green) classified as real or synthetic, while (c) illustrates the ROC curves achieved by individual expert users. 97

9.1 Illustration of the AI-assisted laryngoscopy prototype integrated into the clinical workflow. The interface displays a live endoscopic video feed on the top monitor and real-time AI outputs (bounding boxes of the detected lesions) on a secondary monitor not directly visible to the main operator. There is also a foot-pedal mechanism used by the operator for saving relevant frames. 101

9.2 Representative examples of misclassification of failure cases from the clinical study. (A-C) Examples of *High-risk* lesions misclassified as *Low-risk* due to motion blur, underexposure, or lesion ambiguity. (G) Example of a benign lesion (granuloma) misclassified as *High-risk* due to vascular enhancement. Frames (D) and (H) correspond to false-negative cases. 107

List of tables

2.1 Summary of representative studies on informative frame selection in laryngeal endoscopy. Performances are compared using the F1-score metric and the inference time in seconds. 14

2.2 Summary of existing laryngeal lesion detection methods and corresponding main results. Precision, Recall, Sensitivity, mean Average Precision (mAP), and Frames Per Second (FPS) metrics are used to evaluate the detection performance and the inference speed. 19

2.3 Summary of studies on automatic laryngeal lesion diagnosis from endoscopic images and corresponding main results. 23

2.4 Summary of representative deep learning studies on lesion and/or anatomical structure segmentation in laryngeal endoscopy. Performances are compared using the conventional mean Intersection over Union (meanIoU) and mean Dice Coefficient (meanDSC) across all the semantic classes. 25

3.1 Overview of the datasets used in this thesis for the development and evaluation of AI-assisted laryngeal endoscopy modules. 35

4.1 Performances of the Deep Learning (DL) models on the *internal test* and *external test* sets. 40

5.1 Results from the baselines, namely, YOLOv8n, YOLOv8s, YOLOv5n, YOLOv5s, comparison on the *High-Resolution (HR) subset*. Efficacy performance was assessed via the mean Average Precision (mAP) computed using an Intersection over Union (IoU) threshold of 0.50 ($mAP_{@IoU=0.5}$). Efficiency performance was assessed via the billion floating point operations (GFLOPs) and the Frames Per Second (FPS). 53

5.2 Results from the YOLOv8n baseline pre-training when trained and tested on the *High-Resolution (HR) subset*. Efficacy performance was assessed via the Average Precision (mAP) computed using an Intersection over Union (IoU) threshold of 0.50 ($mAP_{@IoU=0.5}$). 54

5.3 Results from different placements of the Super Resolution (SR) branch. The performance of the *configuration 4-8* (YOLOv8n+SR_{48deep}) with that of the *configuration 2-6* (YOLOv8n+SR_{26deep}) were compared. In the first architecture, the SR branch processes feature maps from the fourth and eighth layers of the YOLOv8n backbone, whereas in the second architecture, it processes those from the second and sixth layers. The mean Average Precision (mAP) was computed using an Intersection over Union (IoU) threshold of 0.50 ($mAP_{@IoU=0.5}$). 54

5.4 Results from the comparison of the Super-Resolution (SR) branch architectures: performance ($mAP_{@IoU=0.5}$) of the proposed *deep* (YOLOv8n+SR_{48deep}) version was compared with that of the *light* one (YOLOv8n+$SR_{48light}$) . . . 55

5.5 Analysis of the pre-training effect on detection performance. The subscripts *labeled* and *unlabeled* were included to improve readability and indicate whether the Super-Resolution (SR) branch pre-training was conducted (i) on the labeled portion or (ii) on the unlabeled and subsequently labeled portion of the HyperKvasir dataset. 55

5.6 Performance evaluation in terms of the mean Average Precision (mAP) computed using an Intersection over Union (IoU) threshold of 0.50 ($mAP_{@IoU=0.5}$) for the proposed SRE-YOLO, YOLOv8n + $SR_{48light}$, and YOLOv8n + SR_{26deep} models on the *ENDO-LC ext* test set, divided into small (bounding box area $< 32 \times 32$) and medium-sized ($32 \times 32 <$ bounding box area $< 96 \times 96$) lesions. 56

5.7 Results from the YOLOv8n baseline training on the *Full dataset* for general *lesion* detection and *Low-risk* versus *High-risk* lesion detection. Efficacy performance was assessed via Precision, Recall, F1-score, and mean Average Precision (mAP), computed using an Intersection over Union (IoU) threshold of 0.50 ($mAP_{@IoU=0.5}$). 58

6.1 Performance metrics of the EfficientNetV2S model for *Low-risk* vs *High-risk* lesion classification on the *internal* and *external* test sets. 68

6.2 Performance of the ViTb16 model for seven-class histological lesion classification obtained during *internal* and *external testing*. 70
6.3 Comparison between the proposed AI model, ChatGPT-4o, and clinicians with varying levels of expertise on the *external test set 2* from *Barcelona-Larynx* data. The Accuracy and p-values from McNemar's test comparing each rater group to the proposed AI module are reported. 71

7.1 Summary of training settings for the investigated segmentation models: *SegMENT-Plus, nnU-Net, SAM, MedSAM, SAMUS, AutoSAMUS, and SAMed*. 80
7.2 Segmentation performance across all test sets: *internal (Genova-Larynx), external (Seoul-Larynx, H&NANCE-Larynx)*. 85

8.1 Impact of adding 10% synthetic images to the real training dataset on YOLOv8-based lesion detection. Results are reported for the *internal generation* and *external detection test* sets. 96
8.2 Histological classification tests (*In-office dataset* and external data from *Barcelona-Larynx*): comparison between the classifier trained only on real data and the classifier trained on real data plus approximately 20% synthetic samples of minority classes. 98
8.3 Quantitative analysis of the expert-observer realism study results across the five participants. 98

9.1 Distribution of lesion types in the clinical study cohort. 105
9.2 Patient-level performance of the VERA system for *Low-risk* versus *High-risk* classification. Metrics were computed from the aggregated confusion matrix of 110-patient analyses. 106

Nomenclature

Acronyms / Abbreviations

AI	Artificial Intelligence
AUC	Area under the ROC Curve
CADe	Computer-Aided Detection
CADx	Computer-Aided Diagnosis
CLS	Classification
CNN	Convolutional Neural Network
CSP	Cross Stage Partial
CT	Computed Tomography
DDPM	Denoising Diffusion Probabilistic Model
DL	Deep Learning
DM	Diffusion Model
DSC	Dice Similarity Coefficient
EDSR	Enhanced Deep Super-Resolution
ENT	Ear, Nose, and Throat
FFN	Feed-Forward Network
FID	Fréchet Inception Distance
FN	False Negatives

FP False Positives

FPN Feature Pyramid Network

FPR False Negative Rate

FPS Frames Per Second

GAN Generative Adversarial Network

GAW Glottal Area Waveform

GDPR General Data Protection Regulation

GPT Generative Pre-trained Transformer

gradCAM gradient-weighted Class Activation Mapping

GUI Graphical User Interface

H&N Head & Neck

HPV Human Papilloma-Virus

HR High-Risk

IoU Intersection over Union

IPCL Intrapapillary Capillary Loops

IS Inception Score

LDM Latent Diffusion Model

LoRA Low-Rank Adaptation

LR Low-Risk

mAP mean Average Precision

MBConv Mobile Inverted Bottleneck Convolution

MEAL Multitask Efficient trAnsformer for Laryngoscopy

MHSA Multi-Head Self-Attention

MLP Multi-Layer Perceptron

MRI Magnetic Resonance Imaging

MSEC Multiscale feature Extraction Capabilities

NBI Narrow Band Imaging

NMS Non-Maximum Suppression

ORL Otorhinolaryngology

PANet Path Aggregation Network

QUADAS Quality Assessment of Diagnostic Accuracy Studies

RCNN Region-based Convolutional Neural Network

ROC Receiver Operating Characteristic

RoI Region Of Interest

SAM Segment Anything Model

SCC Squamous Cell Carcinoma

SPPF Spatial Pyramid Pooling Fast

SRE-YOLO Super-Resolution Enhanced YOLO

SR Super Resolution

SSD Single-Shot Detector

SVM Support Vector Machine

TN True Negatives

TOLMS Transoral Laser Microsurgery

TPR True Positive Rate

TP True Positives

UADT Upper Aero-Digestive Tract

UE Uncertainty Estimation

VAE Variational AutoEncoder

VFP Vocal Fold Paralysis

VIL Video Informativeness Level

ViT Vision Transformer

WL White Light

YOLO You Only Look Once

Chapter 1

Introduction

1.1 Clinical context

1.1.1 Anatomy and Physiology of the Larynx

The larynx (voice box) is a complex organ located in the anterior compartment of the neck, spanning between C3 and C7 vertebrae (Suárez-Quintanilla et al., 2023). It is a component of the Upper Aero-Digestive Tract (UADT), with its superior region, called the epiglottis, attached to the inferior part of the pharynx, while inferiorly it is continuous with the trachea. The principal function of the larynx is to protect the lower airways by preventing the aspiration of food and liquids into the trachea during swallowing. Additionally, it houses the vocal folds and serves as the phonatory organ responsible for sound production. Structurally, it is composed of a cartilaginous skeleton, connected by ligaments and muscles that modulate vocal folds tension and position, and a mucous membrane. The internal space of the larynx is wide in the superior and inferior parts but narrows in the middle, forming a section named glottis and dividing all the spaces into three sections: supraglottic, glottic, and infraglottic. The glottic space comprises the vocal cords and the opening between them, known as rima glottidis. The vocal cords are four folds of fibro-elastic tissue, two superior and two inferior. They can also be divided into true and false vocal folds: the true vocal folds are the primary structures involved in phonation, whereas the false vocal folds provide protective and supportive functions. The interior surface of the larynx is covered by a stratified squamous epithelium in the vocal folds and a ciliated respiratory epithelium elsewhere, which together protect the underlying tissue from mechanical stress and pathogens. During respiration, the vocal folds remain abducted to allow airflow, while during phonation, they adduct and vibrate to generate sound as air passes through. Reflexive closure of the glottis prevents aspiration

during swallowing, ensuring airway protection. The intricate coordination between neural control, muscular dynamics, and vibratory behavior of vocal folds is crucial to guarantee the efficiency of related vital functionalities.

1.1.2 Epidemiology and Clinical Impact of Laryngeal Pathologies

Laryngeal diseases encompass a wide spectrum of benign and malignant conditions that affect both the structure and functions of the vocal folds and surrounding tissues.

Laryngitis refers to inflammation of the larynx. Symptoms usually include a hoarse voice, pain, and coughing (Jaworek et al., 2018). Acute laryngitis is most often caused by a viral infection, while chronic forms are mainly associated with smoking, allergies, and gastroesophageal reflux.

Vocal fold paralysis (VFPs) is the result of laryngeal nerve paralysis, and may arise from congenital conditions, trauma, infections, tumors, and systemic neurologic disorders such as multiple sclerosis (Walton et al., 2017). Depending on its severity, VFPs can significantly compromise phonation, airway protection, and breathing, often leading to dysphonia, aspiration, and even life-threatening respiratory distress.

Among structural lesions, Squamous Cell Carcinoma (SCC) represents the most common malignant tumor of the UADT, accounting for approximately one-third of all Head and Neck cancers (H&N) worldwide. According to recent global estimates (Bray et al., 2024), more than 180,600 new cases and 99,800 deaths of laryngeal cancer were reported globally in 2022, with a predominance in males and strong associations with tobacco use, alcohol consumption, and Human Papilloma-Virus (HPV) infection (Dhull et al., 2018). Despite advances in surgical and non-surgical therapies, early detection remains a key determinant of prognosis: in US population-based data, only 49% of laryngeal cancers were diagnosed at a localized stage (Bray et al., 2024), for which the overall 5-year relative survival was 79% during the period 2015–2021. In contrast, cases with distant metastases exhibited a dramatically lower survival rate of only 35.2% (National Cancer Institute, 2025).

In addition, benign lesions in the H&N region refer to growths that do not spread to other parts of the body, and include nodules, polyps, cysts, granulomas, papillomas, and Reinke's edema. Though non-life-threatening, they can significantly impair voice quality, breathing, and swallowing, thereby affecting patients' communication ability, social interaction, and overall quality of life.

The clinical management of these disorders often requires repeated endoscopic assessments, where diagnostic accuracy heavily depends on the examiner's expertise. Overall,

these findings highlight that early detection and precise lesion characterization are essential for effective care of UADT pathologies, leading to improved patient outcomes and reduced morbidity rates.

1.2 Current Clinical Workflow and Related Challenges

The clinical workflow associated with laryngeal disorders typically starts with a detailed patient history examination, focusing on symptoms such as hoarseness, throat pain, dysphagia, or dyspnea. This first phase is typically followed by endoscopic visualization of the larynx, allowing assessment of mucosal integrity, the presence of structural lesions, and eventually vocal fold mobility. Stroboscopic examination provides additional insight into vibratory function and abnormalities. Biopsy and histopathological analysis confirm the nature of suspicious structural lesions. Imaging modalities such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and neck ultrasound can be employed to evaluate the extension of the previously identified anatomical anomalies, while laryngeal electromyography assists in paralysis evaluation. At this point, treatment decisions are guided by lesion type and severity and can involve pharmacologic therapy for inflammatory conditions, microsurgical excision for benign lesions, and combined oncologic approaches for malignancies. Moreover, a multidisciplinary framework can be identified with the aim of restoring laryngeal function and preserving voice quality.

1.2.1 In-office Laryngeal Evaluation

Endoscopic evaluation constitutes a crucial step in the clinical workflow, as the visual and functional assessment of the larynx directly influences diagnostic accuracy and subsequent therapeutic planning (Davaris et al., 2019). It allows clinicians to visualize organs and internal tissues through cameras mounted on the tip of flexible or rigid instruments. In Ear, Nose, and Throat (ENT) medicine, the three principal ways to examine the internal structures of the larynx are:

1. Mirror laryngoscopy, which consists of placing a small angled mirror into the back of the mouth;
2. Flexible fiberoptic laryngoscopy, in which an endoscope measuring less than 4mm in diameter is inserted through one nostril and guided through the nose to the throat;
3. Rigid transoral laryngoscopy with a rigid tool inserted into the mouth.

Flexible endoscopy is considered the current state-of-the-art technique for laryngeal assessment, as it provides a rapid, real-time, and minimally invasive dynamic visualization of the larynx during phonation and respiration (Rosen et al., 2009). However, several challenges are associated with the visual evaluation of anatomical regions of interest. Not all frames acquired during the examination necessarily contain diagnostically relevant information. Due to the presence of artifacts, saliva, and patient movements related to discomfort, reaching a clear view of the vocal folds often requires multiple attempts and considerable time. As a result, a large proportion of videoendoscopic recordings includes non-informative frames, which can increase clinicians' cognitive workload and lead to diagnostic errors or oversight, especially for early-stage and small lesions. Hence, the accuracy and efficacy of endoscopic diagnosis can be improved by focusing only on good-quality and informative frames. In relation to this domain, advanced optical methodologies have been developed and incorporated into commercial endoscopic instruments to augment visualization capabilities. Beyond the implementation of high-definition imaging systems, the application of Narrow Band Imaging (NBI) has proven advantageous in the detection of laryngeal lesions (Davaris et al., 2019; Vilaseca et al., 2017). This is attributed to its enhancement of vessel and anomaly visualization, which is achieved through the optimized absorption of green and blue light by hemoglobin, as compared to conventional White Light (WL) endoscopy (Kuznetsov et al., 2006). Such enhanced visualization can also help in reducing the miss rate of laryngeal cancer, which has recently been observed to measure 41% in (Kumazawa et al., 2025). However, the current exploitation of flexible transnasal endoscopy, especially combined with NBI modality, in less experienced centers is burdened by its operator-dependent nature, being influenced by the frequency of its use and related factors. It also suffers from a relatively long learning curve and is hampered by intrinsic limitations such as subjectivity in interpretation, attention, and visual inspection capabilities (Żurek et al., 2019). In the standard clinical workflow, endoscopic assessment is often complemented by biopsy, which represents the objective diagnostic step to further analyze the detected lesions and confirm or exclude malignancy. Following missed or incorrect diagnoses, limited expertise among laryngologists may also result in unnecessary repeated biopsies. However, biopsy is a costly and time-demanding procedure, often requiring to be undertaken under anesthesia with related risks and pain for the patient.

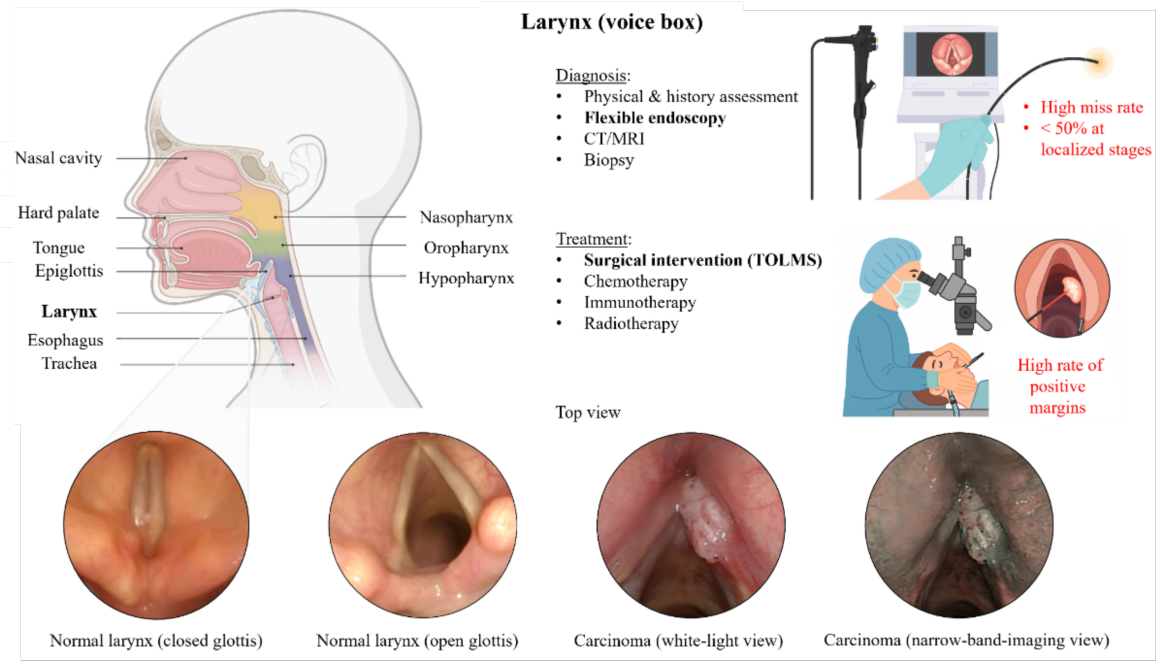


Figure 1.1 Overview of laryngeal diseases and corresponding clinical workflow: on the left, a schematic representation of the Upper Aero-Digestive Tract (UADT) highlights the anatomical position of the larynx. At the bottom, examples of endoscopic views of a normal larynx (both closed and open glottis) and a laryngeal carcinoma under White-Light (WL) and Narrow-Band Imaging (NBI) modalities are illustrated. Flexible endoscopy enables non-invasive in-office visualization but still exhibits a high miss rate, with less than 50% of cases detected at localized stages. In contrast, Trans-Oral Laser Microsurgery (TOLMS), the preferred approach for the treatment of early-stage tumors, offers organ preservation but suffers from a high rate of positive margins.

1.2.2 Intraoperative Assessment and Surgical Planning

Trans-Oral Laser Microsurgery (TOLMS) enables precise, minimally invasive, and cost-effective resection of laryngeal and other H&N tumors through transoral access. Based on the use of a microscope and a micromanipulator-controlled surgical laser, TOLMS offers enhanced intraoperative visualization, superior organ preservation, and reduced morbidity. Currently, this is the preferred approach for treating early-stage tumors of the UADT. This technique is used after a malignant or high-risk lesion is diagnosed, as it provides good oncological outcomes. However, studies report a high rate of positive margins, reaching up to 50% (Fiz et al., 2018), mainly due to the limited precision of intraoperative histopathology and direct visual inspection by the surgeon, used to determine the extent of the tumor.

Figure 1.1 illustrates an overview of the laryngeal anatomy and clinical workflow for laryngeal disease management, highlighting the challenges associated with both in-office

endoscopic evaluation and intraoperative assessment during TOLMS procedures. Therefore, there is an urgent need for automated, objective, and real-time methodologies that can support clinicians during endoscopic evaluation and democratize access to advanced diagnosis. Integrating computer-assisted and real-time imaging analysis into the clinical workflow represents a promising strategy to overcome the limitations and subjectivity of laryngeal visual assessment and lesion management. This can be applied both during the in-office endoscopic exploration phase in order to improve the detection rate of abnormal growths and the accuracy in characterizing them and suggesting appropriate treatment plans, and intraoperatively to improve surgical precision in TOLMS procedures.

1.3 Artificial Intelligence in Laryngeal Lesion Screening and Treatment

Artificial Intelligence (AI) is becoming a constant in medical imaging, enabling automated interpretation and quantitative analysis of complex visual data that traditionally relies on expert evaluation. Within this field, Deep Learning (DL), mainly Convolutional Neural Networks (CNNs) and more recently Vision Transformers (ViTs), has revolutionized how medical images are processed and analyzed. Mimicking the human brain, neural networks consist of multiple interconnected layers of artificial neurons that learn to recognize visual features through iterative exposure to large datasets. By adjusting the weights of these connections during training, the network gradually refines its ability to extract discriminative and hierarchical representations of visual patterns from raw input data. This allows accurate detection, classification, and segmentation of anatomical structures and pathological findings across diverse imaging modalities such as radiology, endoscopy, and histopathology.

The rapid advancement of AI in the medical field has been promoted by access to extensive annotated datasets, improvements in computational power, and the creation of architectures that can capture, correlate, and analyze spatial and contextual information. In clinical practice, AI-based systems are increasingly recognized as valuable decision-support tools, improving diagnostic accuracy, reducing inter-observer variability, and enhancing workflow efficiency.

Several recent studies have explored the current applications and emerging trends of AI in laryngeal disease examination, consistently recognizing its growing impact on the broader field of Otorhinolaryngology (ORL). A thorough literature investigation that included 41 studies was conducted by (Sampieri et al., 2023), where AI and computer vision techniques

were divided based on three fundamental tasks in laryngeal endoscopy: classification for automatic evaluation of image quality and prediction of lesions' histology or risk of malignancy, detection for lesion localization, and segmentation for automatically delineating the boundaries of UADT structures or lesions. Among the main findings, the authors highlighted the importance of utilizing extensive, heterogeneous, and well-curated datasets during training and exhaustively testing real-time performance to face the complexity of real-world clinical use and obtain reliable AI systems.

The concept of "optical biopsy", which refers to the use of AI algorithms to automatically identify and classify tissue abnormalities, without the need for a physical biopsy, was further addressed by (Żurek et al., 2022). In this study, the authors reviewed 11 previous studies regarding the diagnostic utility of AI in laryngeal endoscopy, revealing overall accuracy values ranging from 80% to 99%. Similarly, in (Du et al., 2025), the performances achieved in 9 eligible studies involving Asian and Middle Eastern cohorts were compared, resulting in a pooled sensitivity in diagnosing laryngeal cancer of 95%.

From a clinical impact point of view, the Quality Assessment of Diagnostic Accuracy Studies (QUADAS)-2 tool was used in a literature review by (Wilmes et al., 2025) for the evaluation of quality and risk of bias in 83 studies concerning AI use in UADT endoscopy. Accuracy, sensitivity, and specificity of AI algorithms, clinicians, and their combination in detecting and classifying UADT lesions were computed by the authors, demonstrating that AI performance is comparable to experienced clinicians and exhibits a higher inference speed, even though all included studies were single-center, retrospective, and originated from economically developed nations, where resources for algorithm development, high-quality image equipment, and data storage are possible.

To summarize, despite such analysis demonstrating remarkable achievements at the academic level, they also agree in acknowledging that real-world applications of AI in laryngology and UADT imaging remain relatively underexplored compared to other domains such as radiology or oncology, primarily due to limited data availability and variability in acquisition and validation protocols.

1.3.1 Data Scarcity and Limited Generalizability

Although the development of Computer-Aided Diagnosis and Detection (CADx/e) systems has benefited from large annotated datasets in other fields, laryngology has experienced limited adoption due to the scarcity and limited scope of available data. As observed both in (Sampieri et al., 2023) and (Wilmes et al., 2025), existing public laryngeal endoscopic

datasets are often small, retrospectively obtained from a single center, and tailored for narrow tasks. This restricts their usefulness for real clinical applications, which require broader anatomical and pathological variability. Indeed, notable laryngeal endoscopic datasets include Laryngoscope8 (Yin et al., 2021), Zenodo's Laryngeal Dataset (Moccia et al., 2017), NBI-InfFrames (Moccia et al., 2018b), and CE-NBI (Esmaeili et al., 2022). Despite these resources and the recognized urgent need to establish publicly available open resources, many studies rely on private datasets, eventually supplemented with open-source data, as public datasets are often specialized in a single imaging modality, e.g., NBI, or a single task of the full clinical pipeline, e.g., frame quality assessment in the case of NBI-InfFrames. Moreover, collective data sharing to accelerate AI development and improve its generalizability is further complicated due to the General Data Protection Regulation (GDPR).

Data augmentation techniques can be implemented during AI algorithm design to mitigate data scarcity. However, the problem of dataset imbalance remains a major limitation: reflecting the natural incidence of laryngeal pathologies in clinical practice, in most cases, available datasets are biased toward the most prevalent lesion types or even toward healthy subjects. For instance, in the Laryngoscope8 dataset, nearly 50% of the samples correspond to normal tissue. This imbalance significantly affects model training and generalization, resulting in robust performance for common malignant lesions such as SCC, but reduced accuracy for underrepresented conditions such as granulomas.

1.4 Aim of the Thesis

This research is based on the hypothesis that incorporating AI-driven algorithms into the clinical workflow for laryngeal assessment can substantially improve diagnostic accuracy, currently constrained by human factors, and enable advanced healthcare solutions that enhance care in less specialized centers. The primary objective is to develop and clinically validate a comprehensive AI-assisted system for laryngeal lesion screening, diagnosis, and margin assessment during endoscopic examination and TOLMS procedures. The proposed system aims to address the main challenges associated with variability in laryngeal disease presentation, interpretation, and management by integrating multiple AI modules into a unified framework that supports clinicians throughout the entire clinical workflow. Specifically, the main objectives include:

- Automated selection of high-quality, relevant frames from endoscopic video sequences to reduce cognitive workload and improve diagnostic accuracy during in-office laryngeal evaluations;

- Real-time detection and localization of laryngeal lesions, leveraging both WL and NBI imaging modalities to minimize the abnormality miss rate;
- Risk classification of detected lesions and fine-grained histopathological diagnosis to assist in decision-making;
- Tumor margin delineation to support treatment planning and surgical precision during TOLMS procedures.

The system is intended to operate in real time, providing immediate feedback during endoscopic examinations and surgical interventions. The ultimate goal is to improve patient outcomes by enabling earlier detection of laryngeal pathologies, reducing unnecessary biopsies, and enhancing surgical accuracy.

1.5 Proposed Innovation

Throughout this PhD research, several complementary strategies have been developed to address the main challenges of AI-assisted laryngology, encompassing the entire clinical workflow of laryngeal assessment. During the first phase, a comprehensive multicenter dataset combining WL and NBI endoscopic frames was collected and annotated, enabling the development of lightweight, real-time CNN-based algorithms for informative frame selection and lesion detection. Subsequent work focused on the external validation and optimization of these models through innovative architectural modules to enhance the identification of small and early-stage abnormalities, as well as through data expansion, standardization, and augmentation techniques such as synthetic data generation. In parallel, attention shifted toward lesion risk classification, where self-supervised pretraining and image preprocessing strategies were introduced to mitigate dataset imbalance and variability across clinical centers. The research then progressed toward fine-grained histopathological diagnosis and tumor margin delineation, leveraging both real and synthetic data and exploring the potential of foundation models. Finally, all AI modules were integrated into a physical prototype, called VERA, equipped with a clinically viable graphical User Interface (GUI), and designed to provide a comprehensive AI-assisted vision system for laryngeal lesion screening, diagnosis, and margin assessment in real time. The system, developed within a clinicians-in-the-loop framework, is currently undergoing clinical validation through a prospective study at the Unit of Otolaryngology and Head and Neck Surgery of IRCCS San Martino Hospital (Genova, Italy).

1.6 Structure of the Thesis

Figure 1.2 provides an overview of the thesis structure, mapping the main chapters to their corresponding steps within the clinical workflow. The manuscript is organized into six core chapters, each addressing a specific topic of the research carried out during this PhD project. Chapter 2 presents a comprehensive review of the literature on AI applications in laryngeal endoscopy, outlining existing methodologies, their limitations, and the open challenges that motivated this work. Chapter 8 introduces the proposed framework for synthetic data generation and discusses the perceptual realism evaluation performed with expert clinicians, together with related limitations and future research directions. Following the standard clinical diagnostic workflow, Chapters 4 and 5 detail the development of AI algorithms for informative frame selection and lesion detection, including model architectures, training strategies, and performance assessments. Chapter 6 addresses lesion risk classification, while examining challenges linked to dataset imbalance. This chapter also covers fine-grained histopathological diagnosis. Moving toward surgical planning and treatment, the tumor margin delineation task is examined in Chapter 7, with a focus on the opportunities and challenges of employing foundation models. Finally, Chapter 9 describes the design and implementation of the VERA prototype, including its GUI and clinical validation process.

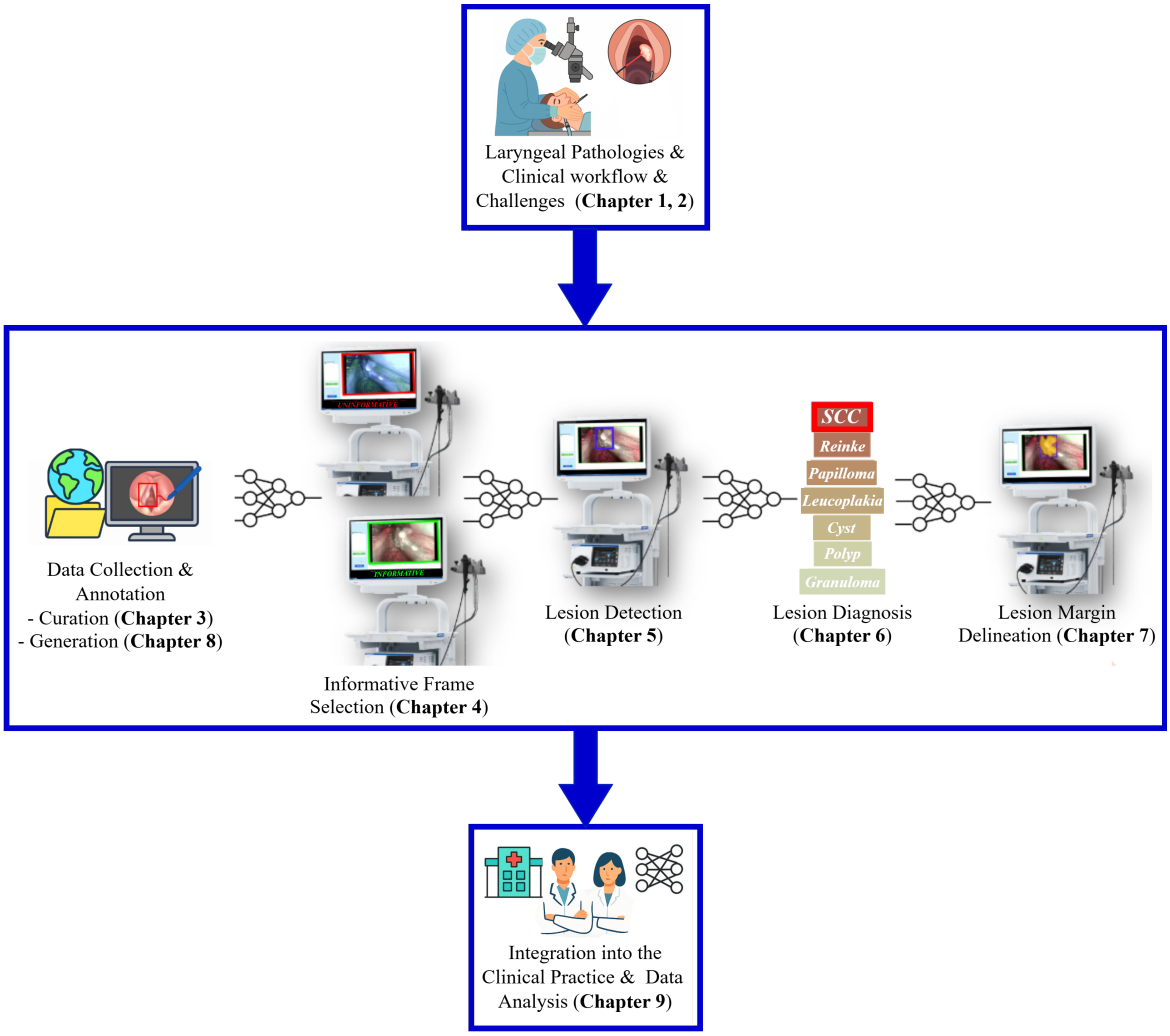


Figure 1.2 Overview of the thesis structure, mapping the main chapters to their corresponding steps within the clinical workflow. Six main chapters are proposed, each addressing a specific aspect of the research conducted during this PhD project, from data collection and curation to AI modules development and clinical validation.

Chapter 2

Literature Review

This chapter reviews the current state of the art in the application of AI to endoscopic evaluation of the larynx. The analysis is structured according to the main tasks that typically compose the clinical diagnostic and surgical workflow:

1. Section 2.1 examines the literature in automatically extracting informative frames from laryngeal endoscopic videos.

2. Section 2.2 analyzes the publications that have addressed the topic of detecting laryngeal lesions.

3. In Section 2.3, a literature review about the methodologies proposed for laryngeal lesion classification is carried out.

4. Section 2.4 explores the results achieved to date for automatic delineation of laryngeal tumor margins during endoscopy;

Lastly, Section 2.5 is dedicated to the emerging topic of synthetic data generation in laryngology, highlighting its potential for overcoming current data limitations and improving AI model generalization. Each section presents key methodologies, datasets, and performance trends reported in the literature.

The work described in this chapter has been published in the following peer-reviewed journal article:

1. Sampieri, C., Baldini, C., Azam, M. A., Moccia, S., Mattos, L. S., Vilaseca, I., Peretti, G. & Ioppi, A. (2023). Artificial intelligence for upper aerodigestive tract endoscopy and laryngoscopy: a guide for physicians and state-of-the-art review. Otolaryngology–Head and Neck Surgery, 169(4), 811-829. https://doi.org/10.1002/ohn.343

2.1 Informative Frame Extraction

The task of selecting frames acquired during endoscopic examinations that contain diagnostically relevant information plays a crucial role in computer-assisted laryngoscopy. Automatically filtering out blurred, poorly illuminated, or occluded frames allows clinicians to focus on the most meaningful views, reducing the duration of the decision-making process and the cognitive load during diagnosis, especially for less-experienced operators. This can also facilitate the creation of good-quality curated datasets for annotation and AI training. To quantitatively assess the performance of algorithms designed for this task, standard metrics such as Precision, Recall, also called Sensitivity or True Positive Rate (TPR), and F1-score are commonly employed (Eq. 2.1, 2.2, and 2.3, respectively). Considering the number of correctly identified informative or positive instances as True Positives (TP), the instances incorrectly identified as positive as False Positives (FP), the positive instances incorrectly classified as negative as False Negatives (FN), and the correctly identified negative instances as True Negatives (TN), the evaluation metrics are computed as follows:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2.1}$$

$$\text{Recall or Sensitivity or TPR} = \frac{TP}{TP+FN} \tag{2.2}$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.3}$$

In addition, inference speed (i.e., the time required to classify a single frame) is a critical parameter, as real-time operation is essential for the integration of such automatic processing into routine clinical workflows.

Moccia et al.((Moccia et al., 2018a)) proposed a classification approach that combines several image feature descriptors, among which image entropy, key points, and intensity variance, and a Support Vector Machine (SVM) to automatically classify 720 NBI frames into informative (Moccia et al., 2018b), blurred, underexposed, and with saliva or specular reflections frames, within an inference time of about 0.030 seconds. Median Recall, Precision, and F1-score values of 84%, 77%, and 80% were achieved after 3-fold cross-validation.

The same dataset, made available on an open repository, was also considered by (Galdran et al., 2019). In this study, the authors compared the effectiveness of three different state-of-the-art CNNs, namely Inceptionv2, ResNet-50, and SqueezeNet, all pre-trained on the ImageNet dataset of natural images. They found the last one the most successful in classifying

Table 2.1 Summary of representative studies on informative frame selection in laryngeal endoscopy. Performances are compared using the F1-score metric and the inference time in seconds.

Study	**Dataset**	**Method**	**Classes**	**F1-score (%)**	**Inference Time (s)**
(Moccia et al., 2018a)	720 NBI frames (NBI-InfFrames)	Handcrafted features + SVM	Informative / blurred / underexposed / saliva or reflections	91/ 79 / 80 / 69	0.030
(Galdran et al., 2019)	NBI-InfFrames	SqueezeNet	Informative / blurred / underexposed / saliva or reflections	98 / 94 / 95 / 91	0.004
(Patrini et al., 2020)	NBI-InfFrames	VGG16 + SVM	Informative / blurred / underexposed / saliva or reflections	97 / 94 / 93 / 91	–
(Yao et al., 2022)	22,000 WL frames	ResNet-18	Informative / non-informative	92 / 69	–
(Zhang et al., 2023c)	NBI-InfFrames	VGG16 + UMAP + Agglomerative clustering	Informative / blurred / underexposed / saliva or reflections	97 / 93 / 95 / 94	–

the informative frame, outperforming the performance achieved in (Moccia et al., 2018a) and obtaining median Precision, Recall, and F1-score amounts ranging from 94% to 95% with 0.004 seconds required to infer a new frame.

Patrini et al. (Patrini et al., 2020) investigated the integration of the two approaches utilized in the two previously mentioned studies. The proposed framework performed feature extraction by means of different CNNs pre-trained on natural images, followed by SVMs or CNN classifiers. When tested on the NBI-InfFrames dataset, the framework's average performance was around 93% for all the metrics.

More recently, the NBI-INfFrames dataset was employed by Zhang et al. (Zhang et al., 2023c) for the actualization of an unsupervised scheme composed by feature extraction via VGG16, dimensionality reduction with a method called uniform manifold approximation and projection, and agglomerative clustering, reaching average Precision of 97%, Recall of 96%, and F1-score of 95%.

The main issues of the NBI-InfFrames dataset are the inclusion of only NBI frames and the small dataset size. A new dataset containing over 22,000 WL frames was introduced by Yao et al. (Yao et al., 2022) to train the ResNet-18 model, resulting in 90% Recall, 94% Precision, and 92% F1-score for informative frames. However, the corresponding metrics for the uninformative frames fell to 76%, 63%, and 69%.

Table 2.1 summarizes the main characteristics and results of the reviewed studies on informative frame selection in laryngeal endoscopy. In general, the studies confirmed that the DL approaches, particularly CNN-based models, outperform traditional handcrafted feature-based methods in terms of accuracy and inference speed. The reported results also reveal that distinguishing between different types of non-informative frames is challenging, suggesting that binary classification (informative vs. non-informative) may be more robust in clinical practice. However, to date, no AI-assisted videoendoscopic system has been deployed and tested in clinical practice. Moreover, the existing datasets introduced and used in such studies are often limited in size, variability, and modality, typically containing only NBI or only WL frames. To address these limitations, this thesis introduces a new multicenter dataset including both WL and NBI laryngoscopic videos, designed to support the development of a generalizable and clinically reliable frame selection algorithm.

2.2 Laryngeal Lesion Detection

Automated detection of laryngeal lesions from informative endoscopic images represents a critical step toward early cancer identification and treatment. Accurate and efficient

lesion localization can support clinicians during the examinations, especially in cases of limited expertise, improve diagnostic consistency, and enable scalable AI-assisted screening programs. To objectively evaluate lesion detection models, Precision is used to quantify the proportion of predicted lesions that are truly lesions, and Recall, also called Sensitivity, to measure the model's ability to identify all real lesions. Contrary to what was explained for the informative frame selection, during the detection task, TP are defined when the Intersection over Union (IoU) metric between the true and predicted object bounding box (Eq. 2.4) exceeds a user-defined threshold.

$$\text{Intersection over Union (IoU)} = \frac{|BoundingBox_{true} \cap BoundingBox_{predicted}|}{|BoundingBox_{true} \cup BoundingBox_{predicted}|} \tag{2.4}$$

The mean Average Precision (mAP) represents a widely used metric in object detection tasks as it summarizes Precision and Recall into a single score. Given the AP of the class i (2.5), which corresponds to the area under the Precision-Recall curve for the class i, the mAP can be obtained as in Eq. 2.6.

$$\text{Average Precision (AP)}_i = \int_0^1 Precision_i(Recall_i)\, dRecall_i \tag{2.5}$$

$$\text{mean Average Precision (mAP)}_{@\text{IoU=0.5}} = \frac{1}{N_{\text{classes}}} \sum_{i=1}^{N_{\text{classes}}} AP_i \tag{2.6}$$

Inference speed is often reported in Frames Per Second (FPS). In 2019, Cen et al. (Cen et al., 2019) exploited the performance of Faster Region-based CNN (Faster RCNN) (Precision=85%, Recall=90%, $AP_{IoU=0.7}$=0.90, and 9 FPS) in localizing visible laryngeal tumors, but only on 54 WL frames. Luan et al. (Luan et al., 2019) achieved a $mAP_{IoU=0.5}$=0.72 with RCNN and a not well-described dataset of 3,223 images with laryngeal structures and six different lesions. Inaba et al. (Inaba et al., 2020) opted for RetinaNet in order to classify 2,400 normal and with superficial cancer NBI images, reaching a Sensitivity of 95%. Faster RCNN and Single-Shot Detector (SSD) were considered by Matava et al. (Matava et al., 2020) for the detection of different anatomical sites, i.e., vocal cords and tracheal rings, in 457 images from bronchoscopy and laryngoscopy (best Sensitivity = 89%). Kim et al. (Kim et al., 2021) obtained Precision=88%, Recall=83%, and F1-score=85% with Mask RCNN in detecting laryngeal masses on 1,224 images. The last three mentioned studies did not report the mAP metric, as the authors were more interested in classification results.

The first study that combined WL and NBI modalities was (Azam et al., 2022a), in which You Only Look Once (YOLO) (version 5) was used to detect laryngeal carcinoma in 624 frames (Precision=68%, Recall=62%, $mAP_{IoU=0.5}$=0.63). In 2023, the number of published papers exploded. Yan et al. (Yan et al., 2023) proposed a Faster RCNN-based approach for laryngeal carcinoma detection in 2,179 WL images collected in 6 different hospitals and with several endoscopic systems (Sensitivity=64%). YOLOv5 was also applied by Wellenstein et al. (Wellenstein et al., 2023), and they found it was capable of detecting and distinguishing carcinoma, anomalies, and normal larynxes with Precision=72%, Recall=75%, and F1-score=74% (FPS=63) in 4,488 WL images acquired in the authors' department and from the public Laryngoscope8 dataset (Yin et al., 2021).

Benign and suspicious for malignancy lesions were detected with 4 different detection models based on Feature Pyramid Networks (FPNs) in 8,172 stroboscopic video frames in Bur et al. (Bur et al., 2023). However, the $mAP_{IoU=0.5}$ did not exceed 0.51. Kim et al. (Kim et al., 2023) resumed YOLOv4 for the detection and classification of benign lesions, including cysts, granuloma, leucoplakia, nodules, and polyps, in 2,183 WL laryngeal endoscopic images with Precision=88%, Recall=82%, and F1-score of 85%. Nie et al. (Nie et al., 2025) optimized the YOLOv8 architecture by incorporating Multiscale feature Extraction Capabilities (MSEC-YOLO), achieving a $mAP_{IoU=0.5}$ of 0.84 when training and testing on 1,353 single-center WL and NBI images that covered 11 categories of laryngeal lesions with a notable imbalance towards malignant tumors. (Dao et al., 2024) introduced the private VoFoCD dataset, containing 1,724 annotated WL laryngoscopy images, and proposed the Multitask Efficient trAnsformer for Laryngoscopy (MEAL) network. It integrates global contextual information with local features to simultaneously classify images as containing or not containing vocal folds and detect both glottic landmarks and lesions. The model achieved an overall detection $mAP_{IoU=0.5}$ of 0.87; however, lesion detection performed the worst among the evaluated tasks.

In recent years, YOLO-based architectures have become the most widely adopted approach in addressing object detection, offering an effective compromise between accuracy and computational efficiency (Table 2.2). Nevertheless, a comprehensive and validated detection model capable of handling multicenter data acquired under heterogeneous optical modalities (WL and NBI) and clinical environments, encompassing both benign and malignant lesion types, is still lacking. In this thesis, these limitations were addressed by developing a YOLO-based detection framework trained on a large multicenter varied dataset that includes both WL and NBI images of benign and malignant laryngeal lesions at differ-

ent sizes and stages. The proposed model was further enhanced with super-resolution and synthetic data augmentation strategies to improve its robustness and generalization.

2.3 Lesion Characterization

This task allows clinicians to distinguish between benign, premalignant, and malignant laryngeal conditions. Accurate classification directly influences patient management, guiding the need for biopsy, surgical intervention, or follow-up. Standard metrics such as Accuracy (Eq. 2.7), Precision (Eq. 2.1), Recall (Eq. 2.2), and F1-score (Eq. 2.3) are commonly used to evaluate classification models.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.7}$$

Regarding the classification of H&N lesions, early contributions are those by Xiong et al. (Xiong et al., 2019), where authors employed the pre-trained InceptionV3 to classify cancer, precancerous lesions, benign lesions, and normal tissues on 14,897 WL images from 5 tertiary hospitals in China with a maximum Accuracy value of 74%, and Araújo et al. (Araujo et al., 2019), which reported an F1-score of 98% achieved by using a SVM to classify the feature extracted via image descriptors and CNN.

To diagnose healthy tissue, hypertrophic vessels, leucoplakia, and dot-like vessels, 1,320 patches were obtained from 33 NBI videos (Moccia et al., 2017). A more comprehensive study, in terms of dataset size and involved type of lesions, was proposed by Ren et al. (Ren et al., 2020). The authors evaluated the performance of a ResNet-based CNN for the automatic classification of laryngeal lesions on 24,667 WL images collected from 9,231 patients. The system achieved an overall Accuracy of 91% in distinguishing normal tissue, polyps, nodules, leucoplakia, and cancer.

The spectrum of possible classification categories was further explored by Yin et al. (Yin et al., 2021): here, the authors exploited FastRCNN to detect lesions and DenseNet121 to further classify the identified Region Of Interest (ROI) into Reinke's edema, glottic cancer, granuloma, normal tissue, vocal cord cyst, nodule, and polyp. Using 3,057 H&N images, mainly belonging to healthy patients, and acquired in WL modality, they achieved 73% of Accuracy.

Palsies were also included in the lesion taxonomy by Cho et al. (Cho et al., 2021), who trained an EfficientNet-B0 network on 4,106 WL images from rigid endoscopes to classify normal tissue and benign lesions with an overall F1-score of 88%.

Table 2.2 Summary of existing laryngeal lesion detection methods and corresponding main results. Precision, Recall, Sensitivity, mean Average Precision (mAP), and Frames Per Second (FPS) metrics are used to evaluate the detection performance and the inference speed.

Study	**Dataset**	**Method**	**Detected objects**	**Performance Metrics**	**FPS**
(Cen et al., 2019)	54 WL images	Faster R-CNN	Laryngeal carcinoma	$mAP_{IoU=0.7}$=0.90	9
(Luan et al., 2019)	3,223 throat images	R-FCN	Cyst, polyp and anatomical structures	$mAP_{IoU=0.5}$=0.72	–
(Inaba et al., 2020)	2,400 NBI images	RetinaNet	Laryngopharyngeal cancer	Sensitivity=95%	–
(Matava et al., 2020)	457 WL laryngocopic + bronchoscopic images	Faster R-CNN, SSD	Vocal cords and tracheal rings	Sensitivity=89%	5
(Kim et al., 2021)	1,224 WL images	Mask R-CNN	Vocal cords and laryngeal mass	Precision=88%, Recall=83%	–
(Azam et al., 2022a)	624 WL + NBI images	YOLOv5	Laryngeal carcinoma	$mAP_{IoU=0.5}$=0.63	38
(Yan et al., 2023)	2,179 WL images (6 centers)	Faster R-CNN	Laryngeal carcinoma	Sensitivity=74%	–
(Wellenstein et al., 2023)	4,488 WL images (2 sources)	YOLOv5	Laryngeal carcinoma and anomaly	Precision=72%, Recall=75%	63
(Bur et al., 2023)	8,172 stroboscopic frames	Feature Pyramid Network (FPN)	Benign and suspicious for malignancy lesions	$mAP_{IoU=0.5}$=0.51	–
(Kim et al., 2023)	2,183 WL images	YOLOv4	Laryngeal cyst, granuloma, leucoplakia, nodule, and polyp	Precision=88%, Recall=82%	–
(Nie et al., 2025)	1,353 WL + NBI images	MSEC-YOLO (YOLOv8 variant)	Benign and malignant laryngeal lesions	$mAP_{IoU=0.5}$=0.84	16
(Dao et al., 2024)	1,724 WL images	MEAL Transformer	Benign lesions, malignant lesions and anatomical structures	$mAP_{IoU=0.5}$=0.87	–

Interestingly, the studies of Dunham et al. (Dunham et al., 2022), Zhao et al. (Zhao et al., 2022), and Esmaeili et al. (Esmaeili et al., 2021) indicated that DL models can obtain better results in a binary-classification task, rather than in multiclass differentiation. In fact, Dunham et al. reported an Accuracy of 93% when their VGG16-based model classified lesions in 19,353 WL laryngeal images as benign or premalignant/malignant, which dropped to 81% when the classes were 5 (normal, nodule, papilloma, polyp, web). Likewise, Zhao and colleagues showed a loss in Accuracy (80% versus 94%) when training their model to recognize 4 classes (normal, polyp, keratinization, and carcinoma) instead of binary urgent/non-urgent classification. They opted for the pre-trained Mobilenet-v2 architecture, and a dataset consisted of 5,122 WL frames acquired from a single general hospital.

A pretrained ResNet-50 model and a dataset of 8,181 vascular patches from (Esmaeili et al., 2022) were used in (Esmaeili et al., 2021), achieving a test Accuracy of 83% on the unseen data during benign versus malignant classification. Other examples of recent works for binary or three-category classification are described below:

- Bhattacharjee et al. (Bhattacharjee et al., 2023) introduced a new deep ensemble model-based laryngeal cancer detection system, which employed pre-trained MobileNetV2, EfficientNet, and DenseNet121, and demonstrated an F1-score of 98% in diagnosing laryngeal cancer from 3,000 WL healthy and malignant frames.

- Yan et al. (Yan et al., 2023) conducted a multicenter case-control study in which they used a Faster R-CNN detector with a convolutional classification head to distinguish between malignant and benign vocal lesions in 2,179 laryngoscopy data from 6 hospitals. They achieved 78% of Accuracy.

- Wei et al. (Wei et al., 2024) developed a fusion network guided by the Segment Anything Model (SAM) for laryngopharyngeal tumor detection with a dual-branch network architecture that concatenates both local and global features. Tested on two datasets of 25,256 and 2,788 WL and NBI frames obtained from 2 Chinese hospitals, the F1-scores of the proposed method were 83.6% and 89.1%.

- Similarly to (Yin et al., 2021), Srivastava et al. (Srivastava et al., 2025) implemented a ResNet-based approach for the binary classification of images' ROIs into benign and malignant classes. The model, trained and tested on lesion and healthy ROIs from 1,290 WL images collected from a single center, achieved an Accuracy of 91%.

Joseph et al. (Joseph and Vidyarthi, 2024) presented a hybrid method based on deep transfer learning and statistical features extraction. This method classified binary abnormalities and

differentiated between healthy tissue, hypertrophic tissue, leucoplakia, and intrapapillary capillary loops (IPCLs) with 99% Accuracy in NBI patches from (Moccia et al., 2017). Vision Transformers (ViTs) were exploited by:

- (Kang et al., 2025), who reached 87% F1-score in a four-class classification task (normal, benign, premalignant, malignant) on 5,768 WL laryngoscopic images from two centers.
- (Zhang et al., 2025), achieving cancer diagnosis results that outperformed physician visual evaluations in 3,140 images. The overall F1-scores were 92%, 95%, and 95% for benign lesions, leucoplakia, and cancer, respectively.

For the first time, the study by (Qiu et al., 2024) explicitly incorporated the temporal dimension of endoscopic imaging, capturing motion and contextual information across consecutive frames to improve diagnostic consistency. High values of Accuracy and F1-score (92% and 95%, respectively) were obtained using the Video-Swin-Transformer for laryngeal cancer identification in 1,109 laryngoscopic videos of normal, benign, and malignant cases.

Lastly, an interesting multimodal methodology was introduced in (Kwon et al., 2025). The authors aimed to improve the classification accuracy of a CNN in identifying glottic cancer from endoscopic images by combining it with a voice CNN trained to recognize faulty voice signals. The authors collected images and voice signals from over 400 men undergoing laryngeal examinations, some of whom had cancer. The decoupled networks achieved F1-scores of 90% for image only and 89% for voice only, whereas the proposed ensemble decision system for images and voices improved performance up to 96%.

These findings emphasize the inherent complexity of laryngeal lesion classification and confirm the necessity of large and heterogeneous datasets for effective multiclass model training. Although existing studies confirm that deploying a binary tool that reliably characterizes lesions at high risk of malignancy would already constitute a significant advancement in clinical routine, similar to the detection task, most have been trained on single-center, single-modality datasets (e.g., WL or NBI) and preprocessed, e.g., ROI-based cropped, high-quality data (Table 2.3). This often results in the overlooking of early-stage and underrepresented lesion types, which reduces the studies' generalizability and applicability in real clinical workflows.

In this work, DL approaches for both binary and multi-pathology lesion recognition across heterogeneous, multicenter endoscopic datasets were implemented. To improve model robustness, domain augmentation and self-supervised pretraining strategies were introduced, mitigating issues related to class imbalance and acquisition variability. Following

the literature studies (Kang et al., 2025; Xiong et al., 2019; Zhang et al., 2025; Zhao et al., 2022), comparative experiments were also conducted between the proposed binary classification model and clinicians with different levels of expertise, demonstrating that the AI system can effectively support laryngeal diagnosis by achieving diagnostic performance comparable to that of the most experienced specialists.

2.4 Tumor Margin Delineation

During surgical planning, accurate delineation of tumor margins ensures complete resection with maximal tissue preservation, positively influencing postoperative voice and swallowing functions. This influence can be translated into better patient quality of life and reduced healthcare costs associated with rehabilitation, additional interventions, and adjuvant therapies. Even during in-office diagnostic processes, lesion segmentation further supports the quantitative evaluation of lesion shape and vascular architecture, providing additional markers of malignancy risk. Over the last 5 years, the number of studies that focus on using DL to segment laryngeal images has increased. However, most efforts still focus on anatomical structure segmentation, i.e., vocal fold and/or glottis delineation, rather than on the clinically crucial task of estimating laryngeal tumor margins. In addition to the IoU (Eq. 2.9), the Dice Similarity Coefficient (DSC) is a more suitable metric for evaluating segmentation performance. Unlike IoU, the DSC places a slightly higher emphasis on correctly segmented regions and remains reliable in medical imaging scenarios affected by class imbalance. The DSC is defined as:

$$\text{Dice Similarity Coefficient (DSC)} = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{2.8}$$

$$\text{Intersection over Union (IoU)} = \frac{|A \cap B|}{|A \cup B|} \tag{2.9}$$

where A is the set of pixels in the ground truth segmentation, and B is the set of pixels in the predicted segmentation.

Among the few studies targeting pathological delineation, Laves et al. (Laves et al., 2019) proposed a weighted ensemble network of UNet and ErfNet for laryngeal tissue segmentation. The network was trained using 536 stereo endoscopic images from 2 patients undergoing laser incisions, which were manually segmented and categorized into void, vocal folds, other tissue, glottal space, pathology, surgical tool, and intubation. They reported an average IoU across the classes of 0.85, with inference rates of approximately 20 FPS for each network.

Table 2.3 Summary of studies on automatic laryngeal lesion diagnosis from endoscopic images and corresponding main results.

Study	**Dataset**$_{\text{(\# of Clinics)}}$	**Method**	**Classes**	**Results (%)**	**FPS**
(Xiong et al., 2019)	14,897 WL images$_{(5)}$	InceptionV3	Cancer/precancerous/benign/normal	Accuracy = 74	–
(Araujo et al., 2019)	(Moccia et al., 2017)	Descriptors + CNN + SVM	Healthy/hypertrophic/leucoplakia/ dot-like	F1-score = 98	41
(Ren et al., 2020)	24,667 WL images	ResNet-based CNN	Normal/polyp/nodule/leucoplakia/ cancer	Accuracy = 91	–
(Yin et al., 2021)	(Yin et al., 2021)	FastRCNN + DenseNet121	Reinke's edema/glottic cancer/granuloma/normal/cyst/nodule/polyp	Accuracy = 73	–
(Cho et al., 2021)	4,106 WL images	EfficientNet-B0	Reinke's edema/granuloma/normal/cyst/ nodule/polyp/leucoplakia/papilloma/ palsies	F1-score = 88	–
(Esmaeili et al., 2021)	(Esmaeili et al., 2022)	ResNet-50	Benign/malignant	Accuracy = 83	–
(Dunham et al., 2022)	19,353 WL images	VGG16	Benign/premalignant/malignant	Accuracy = 93	–
(Zhao et al., 2022)	5,122 WL images	Mobilenet-v2	Urgent/non-urgent	Accuracy = 94	86
(Bhattacharjee et al., 2023)	3,000 WL images	Ensemble model	Healthy/cancer	F1-score = 98	–
(Yan et al., 2023)	2,179 WL images$_{(6)}$	Faster R-CNN + classificator	Benign/malignant	Accuracy = 78	–
(Wei et al., 2024)	28,044 WL + NBI frames$_{(2)}$	SAM-guided fusion network	Normal/benign/malignant	F1-score = 89	–
(Joseph and Vidyarthi, 2024)	(Moccia et al., 2017)	SqueezeNet + statistical features	Normal/abnormal	F1-score = 99	–
(Qiu et al., 2024)	1,109 videos	Video Swin-Transformer	Normal/benign/malignant	F1-score = 95	–
(Srivastava et al., 2025)	1,290 WL + NBI images	ResNet-50	Benign/malignant	Accuracy = 91	–
(Kang et al., 2025)	5,768 WL images$_{(2)}$	Swin-Transformer	Normal/benign/premalignant /malignant	F1-score = 87	340
(Zhang et al., 2025)	3,140 WL images	Vision Transformer	Benign/leucoplakia/cancer	F1-score = 95	–
(Kwon et al., 2025)	471 WL images & voice signals	Image & voice CNNs + decision tree	Normal/cancer	F1-score = 96	–

Ji et al. (Ji et al., 2020) demonstrated that a multi-scale, recurrent, fully CNN was able to segment instances of leucoplakia in 649 WL frames of the glottic area acquired with flexible endoscopes in 0.205 seconds per frame and with a mean IoU value of 0.83.

Zhou et al. (Zhou et al., 2023) further advanced the field with a point-wise spatial network capable of accurately identifying carcinoma in 2,026 images of the UADT, achieving a mean IoU of 0.86. Similarly, Pan et al. (Pan et al., 2022) introduced custom attention frameworks with a hybrid transformer, which improved object boundary definition in laryngeal endoscopy. Testing the proposed model on the same dataset as (Laves et al., 2019), plus on an additional external collection of 822 laryngeal endoscopic images annotated with structural labels, revealed mean IoU and mean DSC values ranging from 0.77 to 0.89 and 0.83 to 0.93, respectively.

In contrast to anatomical-tailored methodologies or those restricted to specific clinical settings, a dedicated network designed to precisely segment laryngeal carcinoma in both WL and NBI images was developed in (Azam et al., 2022b; Sampieri et al., 2024). These studies provided a comprehensive validation on several sets of data, consisting of more than 4,000 images of laryngeal, oropharyngeal, and oral cavity carcinoma. They also conducted an inference speed analysis on five unseen videolaryngoscopies, and reported mean IoU and DSC values between 0.81 and 0.84, as well as an inference speed of 25-26 FPS.

Recently, the widely used SAM was applied to this specific domain by (Wei et al., 2024). Nevertheless, the focus of the mentioned work was on analyzing classification Accuracy rather than SAM's performance for precise tumor delineation.

Table 2.4 highlights that most existing studies have been conducted retrospectively on data from one or few clinical centers and under controlled conditions, overlooking the variability and challenges inherent to real-world endoscopic imaging and clinical deployment of AI systems. In this thesis, such limitations are addressed by investigating the feasibility of adopting SAM, after fast and efficient fine-tuning steps, for real-time malignant lesions segmentation in diverse endoscopic examination settings. Compared to custom-trained models, SAM offers a generalizable and adaptable framework that can be transferred across domains with minimal fine-tuning. Its architecture enables prompt-based precise segmentation, while the availability of variants already pre-trained on diverse medical databases reduces the dependence on large, well-annotated datasets for training (Ma et al., 2024). These features make SAM particularly advantageous for laryngeal imaging, where data are limited and heterogeneous, supporting robust and transferable tumor margin delineation across different clinical settings.

Table 2.4 Summary of representative deep learning studies on lesion and/or anatomical structure segmentation in laryngeal endoscopy. Performances are compared using the conventional mean Intersection over Union (meanIoU) and mean Dice Coefficient (meanDSC) across all the semantic classes.

Study	**Dataset**	**Method**	**Task**	**Results**	**FPS**
(Laves et al., 2019)	536 stereo endoscopic images acquired during laser incisions (2 patients)	UNet + ErfNet	Segmentation of void, vocal folds, other tissue, glottal space, pathology, surgical tool, and intubation instances	meanIoU=0.85	20 for each network
(Ji et al., 2020)	649 flexible endoscopic WL frames	Multi-scale recurrent CNN	Leucoplakia segmentation	meanIoU=0.83	≈ 5
(Zhou et al., 2023)	2,026 WL images of the UADT	Point-wise spatial network	Carcinoma segmentation	meanIoU=0.86	–
(Pan et al., 2022)	822 WL laryngeal images / Data from (Laves et al., 2019)	Attention hybrid transformer	Segmentation of epiglottis, glottal space, pyriform fossa, vocal folds and other tissues / Same as in (Laves et al., 2019)	meanIoU=0.77 / 0.89, meanDSC=0.83 / 0.93	221
(Sampieri et al., 2024)	>4,000 WL + NBI images from 3 centers	Custom EfficientNet-based CNN	Laryngeal, oral and oropharyngeal carcinoma segmentation	meanIoU=0.81 – 0.84, meanDSC=0.81 - 0.83	20-25
(Wei et al., 2024)	28,044 WL + NBI frames from 2 centers	Segment Anything Model (SAM)	Benign and malignant lesion delineation	Focus on classification metrics	–

2.5 Synthetic Data Generation

Researchers have increasingly used synthetic data generation techniques to augment existing datasets and overcome variability limitations. Recent advances in generative models, such as Variational AutoEncoders (VAEs) and Diffusion Models (DMs), have demonstrated remarkable success in producing realistic synthetic images for medical domains, including general endoscopy. The Fréchet Inception Distance (FID) (Heusel et al., 2017) and Inception Score (IS) (Salimans et al., 2016) are conventionally used to quantitatively assess the realism and diversity of the generated images, often complemented by reader studies involving expert clinicians or final users for perceptual validation. The FID measures the distance between feature distributions of real and generated images in the same embedding space, with low values corresponding to higher similarity. The IS evaluates the quality and diversity of generated images by analyzing the confidence and variety of class predictions from a pretrained Inception network: higher IS values indicate more diverse samples.

Several studies have explored generative approaches for gastrointestinal endoscopy or laparoscopy, mainly targeting lower digestive tract imaging. Methods based on VAEs, Generative Adversarial Networks (GAN), and DMs have been proposed to synthesize realistic endoscopic images and videos, enhancing data diversity and improving the performance of downstream tasks such as lesion detection. Works by (Bajhaiya et al., 2023; Diamantis et al., 2022; Li et al., 2024; Liu et al., 2025) demonstrated that generated endoscopic samples can achieve visual quality comparable to real data and boost diagnostic model performance by providing spatially and temporally coherent synthetic content.

Focusing on laryngeal applications, Darvish et al. (Darvish and Kist, 2024) investigated the use of VAEs to synthesize high-speed videoendoscopy frames. The proposed method allowed for precise control over the synthetic closure and opening of the vocal folds by adjusting the latent space values. Then, to quantify vocal fold oscillations, the authors reconstructed Glottal Area Waveform (GAW) biosignals by segmenting the opening between the vocal folds. To this end, they exploited the large-scale, publicly available BAGLS dataset of endoscopic high-speed video with glottal segmentation annotations (Gómez et al., 2020). Since their primary interest was in oscillation analysis, they assessed the performance of the proposed methodology through the normalized mean absolute error of the reconstructed GAW signals.

Khazrak et al. (Khazrak et al., 2025) extracted 404 frames from stroboscopic videos of laryngoscopic assessment, including normal folds, nodules, polyps, cysts, Reinke's edema, sulcus vocalis, granulomas, and keratoses. Then, a Denoising Diffusion Probabilistic Model

(DDPM) was trained to generate synthetic images for different pathology classes, and unrealistic samples were manually removed based on the opinions of experts. The final goal was to train binary and multi-class classification VGG16-based models on 4,180 synthetic samples and synthetic plus real samples datasets. Generated samples resulted in FID scores ranging from 104.70 (nodule) to 227.31 (sulcus vocalis), while classification models achieved the best results when trained on a dataset combining synthetic images with 50% original data: F1-score of 68% and 82%, and Accuracy equal to 71% and 83% for binary and multi-class tasks, respectively. Such performance was compared with that of training exclusively on synthetic images (F1-score=66% and 78%) or real data (Accuracy values < 40%).

As image generation is still an emerging topic in laryngology, no studies have yet addressed realistic synthesis of data for lesion localization or diagnosis using heterogeneous data acquired across different clinical settings, such as in-office or intraoperative examinations with WL or NBI modalities. The laryngeal region's complex anatomical and pathological characteristics make it a particularly challenging area for synthetic approaches. Advanced generative models are required to produce clinically relevant images that capture inherent pathological variability. To reduce the risk of generating unrealistic samples and to avoid asking experts to discard them manually, the approach proposed in this thesis adopted conditioning mechanisms to control the generation process with clinical observations, such as acquisition parameters or spatial information of the lesion. Furthermore, how to maximize the potential of synthetic data by automatically and optimally selecting the relevant samples was explored to improve performance on subsequent tasks.

Chapter 3

Data Collection and Curation

Deep learning models for laryngeal endoscopy critically depend on the availability of large, well-curated, and heterogeneous datasets. Throughout this PhD project, multiple datasets were collected, curated, and annotated in close collaboration with clinicians from several international centers. These datasets include WL and NBI recordings, cover a wide spectrum of benign and malignant lesions, and differ in terms of acquisition devices, clinical environments, and annotation types (bounding boxes, segmentation masks, and histological labels). To ensure consistency and clinical validity, all datasets were manually revised by experienced otolaryngology collaborators, who reviewed every frame, verified lesion categories, and refined or added annotations where needed. This harmonization step was essential to mitigate inter-center variability and establish a unified annotation standard across datasets originating from different hospitals. In addition, all data were standardized in terms of metadata structure and annotation format. Bounding boxes, segmentation masks, classification labels, and acquisition metadata were consolidated into a consistent schema and exported into JSON files associated with each image. These standardized annotations, together with the corresponding endoscopic frames, were then stored on the specifically created institutional data platform, enabling secure access, reproducibility, and traceability for all subsequent AI experiments. Table 3.1 summarizes all the datasets employed in this thesis, providing an overview of their main characteristics and their specific role in the different AI tasks addressed in the thesis and described in the following chapters.

For each specific task, a dedicated collection of data was created by selecting and splitting frames from the above-mentioned datasets. The selection criteria and splitting strategies are detailed in the following sections, along with the specific preprocessing steps applied to optimize model performance for the targeted tasks.

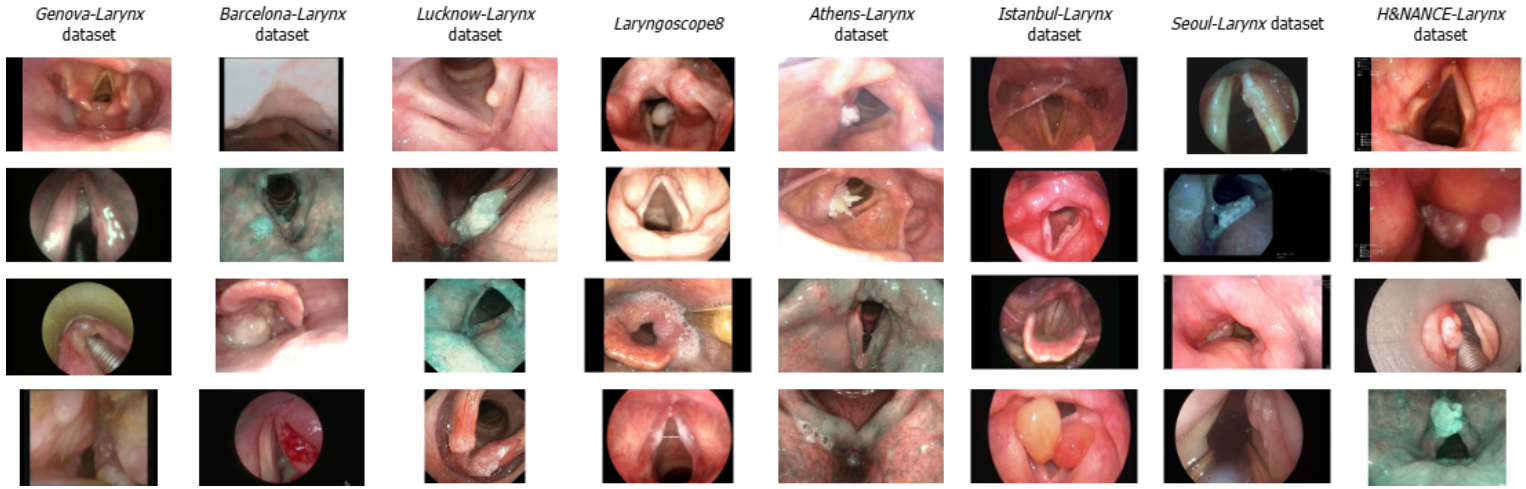


Figure 3.1 Representative sample images from eight laryngeal endoscopy datasets (Genova-Larynx, Barcelona-Larynx, Lucknow-Larynx, Laryngoscope8, Athens-Larynx, Istanbul-Larynx, Seoul-Larynx, and H&NANCE-Larynx) illustrating the diversity in imaging conditions, viewpoints, and pathological appearances across datasets.

3.1 Datasets for Frame Quality Assessment

The public NBI-InfFrames dataset from Moccia et al. (Moccia et al., 2018b) was initially used for automatic extraction of informative frames. It comprises 720 frames from NBI videos of 18 laryngeal SCC cases: 180 informative frames with proven good-quality information, 180 dark frames identified as underexposed, 180 blurred frames, and 180 frames that exhibit either saliva or specular reflections. The images were captured with a resolution of 1,920×1,072 pixels, using an Olympus Visera Elite S190 video processor and an Olympus ENF-VH HD video rhino-laryngoscope[1]. However, this dataset presents some weaknesses: it consists of only NBI frames, and it is based only on a few patients, which might reduce the generalization capability of a model trained only on this dataset. Moreover, prospectively considering screening campaigns, the automated informative frame selection should be exploited not only for patients with laryngeal carcinoma but also for other diseases and for healthy subjects.

To address these limitations, a new collection of both WL and NBI frames from the *Genova-Larynx* and *Barcelona-Larynx* datasets (Table 3.1), indicated as the *Frame quality* dataset, was added to the existing set of images. The recordings of the videolaryngoscopies on 51 patients with a healthy larynx or affected by several laryngeal conditions (SCC, papillomatosis, polyps, Reinke's edema, granulomas, and nodules) from the Units of Otolaryngology and Head and Neck Surgery of the IRCCS San Martino Hospital, University of Genova (Italy) and Hospital Clínic de Barcelona (Spain) were retrieved[2]. Videos were captured in the office using a flexible ENF-VH laryngoscope[1] through a transnasal route. The examinations were performed first using WL and then switching to NBI. VideoLAN VLC media player v.3.0.18 software was used for the sampling of the frames, that were then divided into two categories according to the following criteria: frames with a close and clear view of the larynx as a whole or of its subsites without artifacts alterations such as blur, saliva, reflexes or underexposure were classified as *informative (I)*. All the other extracted frames were classified as *uninformative (U)*. At both centers, junior otolaryngologists independently labeled the frames while a senior laryngologist reviewed the dataset.

3.2 Datasets for Lesion Detection

A multi-center dataset of endoscopic images containing different benign and malignant laryngeal lesions, as well as healthy cases, was collected for the lesion detection task.

[1] Olympus Medical System Corporation, Tokyo, Japan

[2] The Institutional Review Board approval was obtained at both institutions (CER Liguria: 169/2022; Reg. HCB/2023/0897)

Originally, endoscopic videos were captured at the Units of Otolaryngology and Head and Neck Surgery of the IRCCS San Martino Hospital, University of Genova (Italy)[3] during outpatient assessments using a High-Definition (HD) video rhino-laryngoscope ENF-VH by Olympus[1] through a transnasal route or during intraoperative examinations with 0°, 30°, or 70° rigid endoscopes with HD camera head connected to a Visera Elite CLV-S190 light source[1]. The extraction of frames from recorded videos was conducted by clinical partners, leading to the *Genova-Larynx* dataset (Table 3.1). Specifically, they consistently saved at least 5 relevant images in WL and 5 in NBI per patient, using a video player software. A collection of 4,214 frames with diverse appearances and resolutions, representing a wide spectrum of healthy and pathological cases, was selected for detection.

In order to reduce bias caused by a single-center source, two other sets of data were integrated with the previous collection.

The first additional set was the open-access Laryngoscope8 dataset (Yin et al., 2021) (Table 3.1). It consists of 3,057 WL laryngeal images collected from H&N surgery procedures in the Department of Otorhinolaryngology of the Sixth Medical Center of the Chinese PLA General Hospital, Beijing (China). The images, mostly with a resolution of 1,920×1,080 pixels, were classified into eight categories - Reinke's edema, glottic cancer, granuloma, leucoplakia, cyst, nodules, polyps, and normal VCs - by the otolaryngologists in the hospital. Two laryngoscope devices were used during the data acquisition process: Xion Matrix HD3[4] and Delon HD380B[5].

The second additional set includes images obtained from the ENT Center, Voice & Airway Clinic of the Sushrut Institute of Plastic Surgery & Super Specialty Hospital, Lucknow (India), referred to as *Lucknow-Larynx* dataset (Table 3.1). A total amount of 1,216 WL and NBI endoscopic frames with resolutions ranging from 496×508 to 1,080×1,728 were acquired from both healthy patients and patients with abnormalities through an HD video rhino-laryngoscope[1].

To combine the datasets cohesively and ensure uniform standards, 5 laryngologists with more than 5 years of experience in the field first conducted a pre-evaluation of all frames. This step is aimed at excluding those with a bad quality or ambiguous existing annotations. Exclusion criteria for image quality included underexposure, blurred frames, and the presence of saliva or specular reflections, following the same criteria utilized in the previous frame selection task described in Section 3.1. Following this preliminary screening, they used the

[3]CER Liguria: 169/2022

[4]Xion GmbH, Berlin, Germany

[5]Beijing Fanxing Guangdian Medical Treatment Equipment Co., Ltd, Beijing, China

CVAT software tool[6] to annotate bounding boxes around each identified lesion and assign labels to them based on pathological findings, with the corresponding lesion distribution detailed in Table 3.1.

3.3 Datasets for Lesion Characterization

To address the risk assessment and diagnosis of laryngeal lesions, data were collected at 5 different tertiary referral centers for H&N:

1. *Genova-Larynx set*: This set of images was collected at the Units of Otolaryngology and Head and Neck Surgery of the IRCCS San Martino Hospital, University of Genova (Italy) with local Ethics Committee approval[7]. It comprises 7,836 images collected from 889 different patients.

2. *Lucknow-Larynx dataset*: A subset of 899 images was selected from the full collection described in 3.4, by excluding the normal cases and those with nodules. These images were acquired from different patients at the ENT Center, Voice & Airway Clinic of the Sushrut Institute of Plastic Surgery & Super Specialty Hospital, Lucknow (India) with local Ethics Committee approval.

3. *Barcelona-Larynx dataset*: A total of 150 images, each from a different patient, were selected from data collected at the Otolaryngology Unit of Hospital Clínic de Barcelona (Spain) with local Ethics Committee approval[8]. Among all frames, 13 frames belonging to healthy cases were considered only for the comparisons of the proposed AI solution with physicians and ChatGPT.

4. Athens-Larynx dataset: It comprises 929 images collected from 133 different patients at the Otolaryngology Unit of the Oncologic Hospital Saint Savvas, Athens (Greece), with local Ethics Committee approval.

5. *Laryngoscope8 dataset* (Yin et al., 2021): As previously described in Section 3.2, data were collected at the Department of Otorhinolaryngology of the Sixth Medical Center of PLA General Hospital, Beijing (China). This publicly available dataset comprises 3,057 images from 1,950 patients, of which only 1,400 non-healthy case images

[6]https://www.cvat.ai/
[7]CER Liguria: 169/2022
[8]Reg. HCB/2023/0897

were selected for lesion characterization after excluding the strongly underrepresented *Nodule* class.

The videos collected at all participating centers were recorded during normal clinical practice using a flexible or rigid endoscope. From these videos, images were manually extracted, pursuing good-quality and clear visualization of the larynx constraints, and annotated to build the datasets mentioned above. The image selection process included both WL and NBI imaging modalities and different imaging angulations, while the annotation was performed by expert physicians with a minimum of 5 years of experience in laryngology and NBI image interpretation at each clinical center. All data were subsequently revised by a single otolaryngologist for consistency. As in Section 3.2, each laryngeal image was marked with a bounding box encompassing the entire lesion's visible surface and labeled according to the lesion's type obtained from biopsy or expert confirmation. If multiple lesions were present, multiple boxes were added. If no lesion was present, no box was added. The possible lesion types included eight categories: *cyst, granuloma, leucoplakia, polyp, papilloma, Reinke's edema, healthy, and carcinoma*. Laryngeal frames characterized by the absence of such lesions were organized into a single category named *Healthy*.

3.4 Datasets for Tumor Margin Definition

The segmentation task designed to precisely identify tumor margins was conducted using three clinically annotated datasets comprising laryngeal SCC frames:

1. *Genova-Larynx* subset of SCC data annotated with tumor masks: This large and varied collection of images, arising from the full collection in Table 3.1, was used for training and internal testing. It consists of 3,639 WL and NBI frames with pixel-level tumor annotations collected from the Unit of Otolaryngology and Head and Neck Surgery at IRCCS San Martino Hospital, University of Genova (Italy)[9].

2. *Seoul-Larynx* dataset: Used for external validation, this dataset contains 156 WL and NBI SCC frames independently annotated by otolaryngology experts at the Department of Otolaryngology of Severance Hospital, Yonsei University (South Korea).

3. *H&NANCE-Larynx* segmentation samples: An additional prospective set of 134 frames from ongoing clinical examinations was used for real-world performance evaluation. Such data were collected during online endoscopies at the Unit of Otolaryngology

[9]CER Liguria: 169/2022

and Head and Neck Surgery at IRCCS San Martino Hospital (Italy)[10], directly saved and annotated by the clinical team without any other processing or specific selection criteria. They are characterized by high variability in terms of image quality and annotation complexity.

[10]CET Liguria: 508/2023

Table 3.1 Overview of the datasets used in this thesis for the development and evaluation of AI-assisted laryngeal endoscopy modules.

Dataset	Center	Country	Settings	$\text{Lesion}_{\text{instances}}$	#Frames	Used for
Genova-Larynx	Units of Otolaryngology and H&N Surgery of the IRCCS San Martino Hospital	Italy	WL + NBI; In-office + Intraoperative	$\text{SCC}_{(3,298)}$; $\text{Cyst}_{(425)}$; $\text{Granuloma}_{(143)}$; $\text{Leucoplakia}_{(590)}$; $\text{Papilloma}_{(791)}$; Reinke's $\text{edema}_{(1,070)}$; $\text{Polyp}_{(1,527)}$; $\text{Nodue}_{(36)}$; $\text{Normal}_{(5,030)}$	12,910	Frame selection, Detection, Classification, Segmentation, Generation
Barcelona-Larynx	Hospital Clínic de Barcelona	Spain	WL + NBI; In-office + Intraoperative	$\text{SCC}_{(406)}$; $\text{Cyst}_{(14)}$; $\text{Granuloma}_{(16)}$; $\text{Leucoplakia}_{(14)}$; $\text{Papilloma}_{(17)}$; Reinke's $\text{edema}_{(13)}$; $\text{Polyp}_{(21)}$; $\text{Normal}_{(1,388)}$	1,889	Frame selection, Classification
Lucknow-Larynx	ENT Center, Voice & Airway Clinic of the Sushrut Institute of Plastic Surgery & Super Specialty Hospital	India	WL + NBI; In-office	$\text{SCC}_{(365)}$; $\text{Cyst}_{(66)}$; $\text{Granuloma}_{(60)}$; $\text{Leucoplakia}_{(108)}$; $\text{Papilloma}_{(54)}$; Reinke's $\text{edema}_{(10)}$; $\text{Polyp}_{(236)}$; $\text{Nodule}_{(82)}$; $\text{Normal}_{(235)}$	1,216	Detection, Classification
Laryngoscope8 (Yin et al., 2021)	Department of Otorhinolaryngology of the Sixth Medical Center of the Chinese PLA General Hospital	China	WL; In-office	$\text{SCC}_{(22)}$; $\text{Cyst}_{(59)}$; $\text{Granuloma}_{(493)}$; $\text{Leucoplakia}_{(181)}$; Reinke's $\text{edema}_{(108)}$; $\text{Polyp}_{(537)}$; $\text{Nodule}_{(172)}$	1,572	Detection, Classification
Athens-Larynx	Otolaryngology Unit of the Oncologic Hospital Saint Savvas	Greece	WL + NBI; In-office	$\text{SCC}_{(384)}$; $\text{Cyst}_{(61)}$; $\text{Granuloma}_{(31)}$; $\text{Leucoplakia}_{(152)}$; $\text{Papilloma}_{(37)}$; Reinke's $\text{edema}_{(96)}$; $\text{Polyp}_{(168)}$	929	Generation, Detection, Classification
Istanbul-Larynx	University Health Sciences	Turkey	WL + NBI; In-office	$\text{Cyst}_{(36)}$; $\text{Granuloma}_{(33)}$; $\text{Leucoplakia}_{(41)}$; $\text{Papilloma}_{(5)}$; Reinke's $\text{edema}_{(5)}$; $\text{Polyp}_{(163)}$; $\text{Nodule}_{(50)}$	333	Detection
Seoul-Larynx	Department of Otolaryngology of Severance Hospital	South Korea	WL + NBI; In-office + Intraoperative	$\text{SCC}_{(156)}$	156	Segmentation
H&NANCE-Larynx	Same as Genova-Larynx	Italy	WL + NBI; In-office + Intraoperative	$\text{SCC}_{(849)}$; $\text{Cyst}_{(141)}$; $\text{Granuloma}_{(6)}$; $\text{Leucoplakia}_{(401)}$; $\text{Papilloma}_{(29)}$; Reinke's $\text{edema}_{(49)}$; $\text{Polyp}_{(182)}$	1,657	Segmentation, Clinical validation

Chapter 4

Informative Frames Extraction

4.1 Overview

Informative image selection in laryngoscopy has the potential for improving automatic good-quality data extraction from full video laryngoscopies, allowing selective data storage and a faster review and decision-making process. Such functionality can also help to support dataset curation and, when combined with other AI detection or diagnosis models, to achieve real-time processing during deployment. This work explored the development and validation of a lightweight DL model designed to face the automatic identification of informative frames in laryngeal endoscopy. We demonstrated the feasibility and efficacy of AI integration into clinical practice with the aim of providing visual feedback to guide the otolaryngologist during the examination.

The structure of this chapter is organized as follows. Materials and methods are described in Section 4.2: the Subsection 4.2.1 presents the datasets and annotation criteria employed in this study, while a description of the architecture of the proposed model and the implementation details is available in Subsections 4.2.2 and 4.3.2, respectively. Sections 4.4 and 4.5 introduce the experimental results and their interpretation.

The work described in this chapter has been published in the following peer-reviewed journal article:

- Baldini, C., Azam, M.A., Sampieri, C., Ioppi, A., Ruiz-Sevilla, L., Vilaseca, I., Alegre, B., Tirrito, A., Pennacchi, A., Peretti, G., Moccia, S., & Mattos, L.S. (2024). An automated approach for real-time informative frames classification in laryngeal endoscopy using deep learning. European Archives of Oto-Rhino-Laryngology, 281(8), 4255-4264.

4.2 Materials and Methods

4.2.1 Dataset

As described in Section 3.1 of Chapter 3, for this task, the dataset used was obtained by merging data from the publicly available image collection (Moccia et al., 2018b) and the images provided by the IRCCS San Martino Hospital (Italy) included a total of 5,147 frames, and was considered for training and internal validation. It was reorganized with a patient-wise allocation method into 3,126 frames for the training phase, 1,317 frames for validation, and the remaining 704 frames (346 *I* and 358 *U*) for *internal test* set. Data were then further augmented by applying 10% vertical and horizontal shifts, 10% zoom, rotations ranging from 0° to 10° and horizontal flip. The collection of 646 NBI and WL laryngeal frames (306 *I*, and 340 *U*) retrieved from the Hospital Clínic de Barcelona (Spain) was used as an *external test* set to validate the model performance on data acquired and annotated independently from the training set, allowing for evaluating the robustness of the proposed method. Data variability is associated with the geographical diversity of the involved population, the endoscopic technique, and the annotation process conducted by different physicians.

4.2.2 Architecture and Implementation Details

A careful analysis of the literature related to the processing of laryngeal endoscopic frames was conducted. This included the work by Yao et al. (Yao et al., 2022), who presented a very detailed comparison of the performance of various state-of-the-art networks. Based on their results, ResNet-50 and MobileNetv2 were exploited for the task of informative frame extraction due to their computational efficiency. ResNet50, introduced by He et al. (He et al., 2016), is based on the concept of residual learning or residual blocks, which enable the mitigation of the vanishing-gradient problem during training through the use of identity shortcut connections. This design allows the model to learn complex functions while maintaining stable training dynamics. Opting for one of the compact versions of ResNet that counts only 50 layers (Fig. 4.1), the focus is on accelerating the inference process. MobileNetV2, proposed by Sandler et al. (Sandler et al., 2018), is instead optimized for computational efficiency: it employs inverted residual blocks and depthwise separable convolutions, significantly reducing the number of parameters and operations without sacrificing performance. This makes MobileNetV2 particularly suitable for deployment on resource-constrained devices, such as mobile phones and embedded systems. Both architectures

were adapted for binary classification by modifying the top classification layer to output the probability of the frame belonging to each of the *I* and *U* classes, using a softmax function.

4.3 Experimental Protocol

4.3.1 Training Settings

Instead of the standard ImageNet-based pre-training, the pretext task strategy introduced by Gidaris et al. (Gidaris et al., 2018) was adopted to speed up the training and enhance its performance. After randomly sampling 232 frames from the *Frame quality* dataset, patches measuring 512×512 pixels were cropped, and subjected to rotations of 0°, 90°, 180°, and 270°. The CNNs were pre-trained for 200 epochs on the rotated patches to predict the correct rotation angle, with the categorical cross-entropy loss, the Adam optimizer, a batch size of 8, an initial learning rate of 1e-4, and a decay rate of 0.9. The best-performing weights obtained during this "pretext" phase were used to initialize the weights of the corresponding models for the real task, which involved training the CNNs for 200 epochs using the same hyperparameters as those of the pretext task for the purpose of classifying the frames as *I* or *U*. The performance of the developed models was tested both internally and externally. Additionally, four unedited preoperative video laryngoscopies not used for the model's training were selected for testing the computational speed of the model and simulating real-time frame quality classification during examinations.

4.3.2 Performance Evaluation

Classification Metrics

The outcomes of the DL models were evaluated by comparing the predicted classes with the ground-truth classes. TP represents the number of frames accurately predicted as belonging to the class *I*, while TN accounts for the number of *uninformative* frames correctly predicted. FN corresponds to an *informative* frame not identified by the model, and conversely, FP refers to frames wrongly categorized as part of that class. Based on these definitions, the models' performance was assessed by calculating Precision, Recall, F1-score, and the Receiver Operating Characteristic (ROC) curve. The Receiver Operating Characteristic (ROC) curve is obtained by plotting the True Positive Rate (TPR) (Eq. 2.2) against the False Positive Rate (FPR) for all possible decision thresholds of the binary classifier. The TPR and FPR are

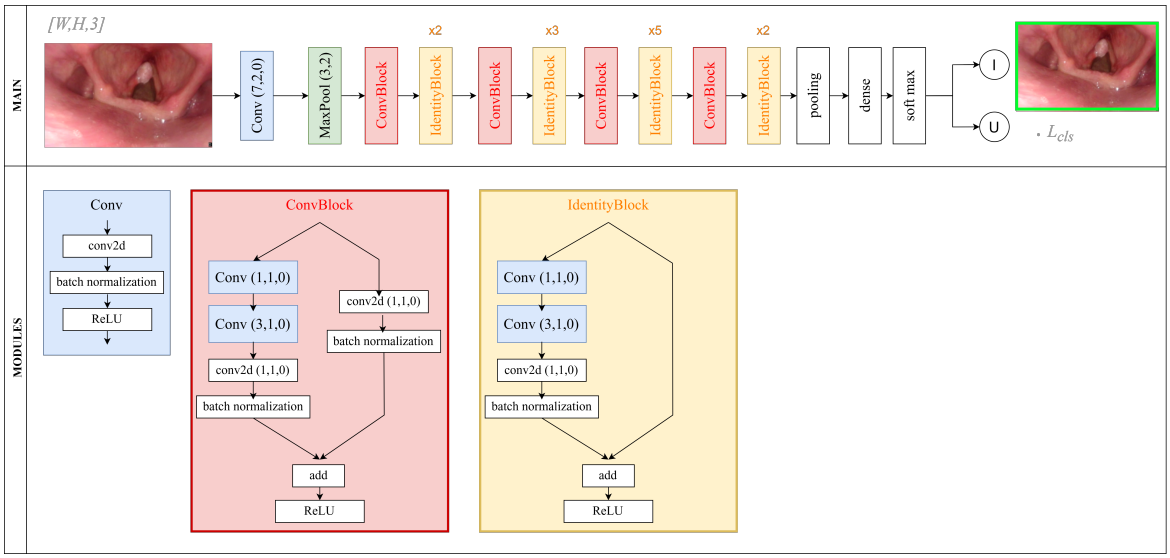


Figure 4.1 Architecture of the ResNet-50 model employed for informative frame classification. The final fully connected layer was modified using a softmax activation function to output the probabilities of the two classes: *Informative (I)* and *Uninformative (U)*.

defined as follows:

$$\text{FPR} = \frac{FP}{FP+TN} \tag{4.1}$$

To provide better insights into the attention of the CNNs, the gradient-weighted Class Activation Mapping (gradCAM) algorithm (Selvaraju et al., 2017) was implemented. This algorithm outputs a heatmap of the important regions in an image used by the network to make its predictions. The map is obtained by computing the gradients of the predicted class with respect to the final convolutional layer's feature maps. It permits users to understand the DL model decisions, thereby facilitating debugging and improvements.

Regarding the real-time performance assessment on videos, the informativeness of the selected video was analyzed in a dynamic context by measuring the Video Informativeness Level (VIL) (Eq. 4.2), which is equivalent to the fraction of frames predicted as informative ($I(t)$) with regard to the totality of examined frames ($I(t)+U(t)$) at each instant t:

$$\text{Video Informativeness Level (VIL) (t)} = \frac{I(t)}{\text{I(t)} + \text{U(t)}} \times 100 \tag{4.2}$$

VIL values approaching the upper range, i.e., 100%, indicate that most frames are informative, reflecting an endoscopic examination performed under good acquisition conditions. This measure can quantify the part of the video that is really relevant for the examiner with potential implications for selective data storage and a faster review process.

Table 4.1 Performances of the Deep Learning (DL) models on the *internal test* and *external test* sets.

Model	TPs	TNs	FPs	FNs	Precision (P)	Recall (R)	F1-score
Internal test							
MobileNetv2	333	339	19	13	95%	96%	95%
ResNet-50	334	339	19	12	95%	97%	96%
External test							
ResNet-50	272	331	9	34	97%	89%	93%

The classification output of each frame is presented to the user as a colored border around the frame: green for *I* prediction, red for *U* prediction.

Statistical Analysis

The McNemar test was employed to statistically compare the performance of the different evaluated CNNs on the *internal test* set. Then, the discriminative performance of the selected model was compared between the *internal test* and *external test* sets by using the ROC curves and Area Under the ROC Curve (AUC) calculations. Pairwise comparisons of AUCs were conducted using a Z-test. A two-sided $p<0.05$ was considered significant for both tests. Statistical analysis was carried out using Python version 3.9 (packages scipy.stat and statmodels version 0.13.2).

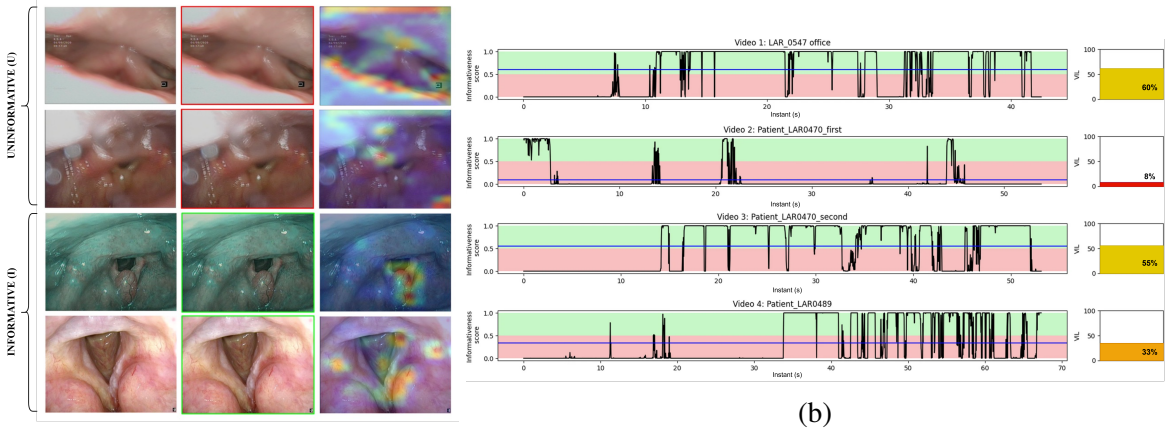
4.4 Results

The classification metrics generated by the CNNs when tested on the *internal test* set after the "pretext" task-based weights initialization and model training are reported in Table 4.1. MobileNetv2 classified frames with a Precision and an F1-score equal to 95%, while the Recall was 96%. Similarly, the ResNet-50 model achieved the best performance during the internal validation. Out of the 704 test frames, 673 were correctly classified, resulting in Precision = 95%, Recall = 97%, and the F1-score = 96%. Even though MobileNetv2 and ResNet-50 performed closely on the *internal set* with no statistical differences (p-value>0.05), the ResNet-50 model was chosen for additional *external test* evaluation and video testing due to the superior F1 score. An ablation study was further conducted to validate the efficiency of the "pretext" task, demonstrating that the ResNet-50 F1-score improved by one percentage point when compared to the results achieved using the standard ImageNet-based weights. Table 4.1 also reports the results of the ResNet-50 model in classifying *informative* frames on

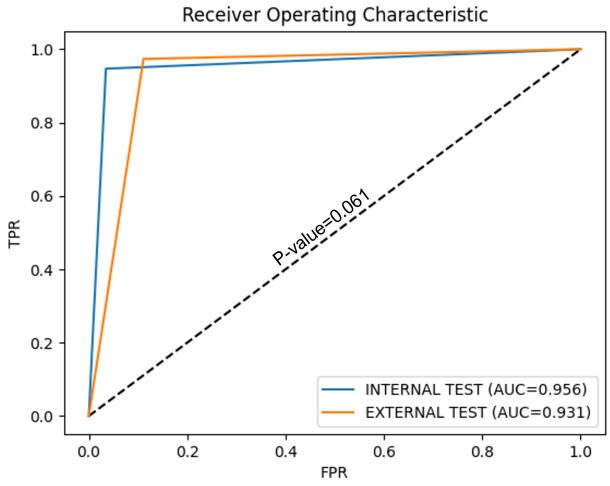
the *external set*: the Precision, Recall, and F1-score were 97%, 89%, and 93%, respectively. Figure 4.2a shows some graphical examples of the output of the ResNet-50 model on the *external set*. Misclassifications mainly occurred with frames in which there is a slight blur but not enough to set the ground truth as *uninformative*, while those that exhibit reflections along with a clear view of the vocal cords can be borderline cases. When comparing the ROC curves of the ResNet-50 model for the *internal* and *external tests*, no statistical difference was found between the two datasets (Fig. 4.2c), validating a high generalization capacity. The robustness of the proposed method was further evaluated on the 4 unseen videos. The mean computing time was 0.015 seconds per frame when the ONNX library was utilized for inference on an NVIDIA RTX A6000 GPU (48 GB of memory), while the VIL values ranged from 8% to 60% (Fig. 4.2b), suggesting that the proposed method can work in real time and it could potentially reduce the size of the data to be stored and subsequently assessed.

4.5 Discussion

The introduction of an automatic informative frame classifier offers several advantages, including the possibility of time savings during endoscopic examinations, facilitating the clinical decision-making process, and the pre-selection of frames to label and process with CADx/e systems. Even though this topic has been studied by a few publications in the literature so far, no attempt has been made to implement these algorithms on real-world procedures. The present work introduces an automated approach for assessing frame quality during laryngoscopic examinations, establishing the first extensively generalized multi-center and multi-modality dataset designed for this specific purpose. With its solid precision and real-time capabilities, the proposed method proved to be suitable for deployment in clinical settings, either autonomously for objective quality control and navigation support or in conjunction with other algorithms. Tests on videos were conducted to simulate clinical use by developing a user interface capable of classifying data in real time and providing immediate feedback to the clinician. Results revealed that VIL never exceeded 60%, suggesting that our system could potentially reduce the size of the data to be stored and analyzed by operators or subsequent AI algorithms.



(a) (b)



(c)

Figure 4.2 (a) Examples of frame predictions from the *external test*. The first column shows the original images along with their corresponding ground-truth labels. In the second column, the prediction is output as a green frame for *informative (I)* images and as a red frame for *uninformative (U)* cases. In the third column, the attention maps used by the model for the prediction are output according to the gradCAM algorithm. (b) Informativeness trends over time and the corresponding final VIL values are illustrated for the 4 tested videos. (c) Model performance according to the Receiver Operating Characteristics (ROC) curves for the internal (blue) and external (orange) test sets.

Chapter 5

Laryngeal Lesion Detection

5.1 Overview

During the examination, when otolaryngologists detect suspect lesions, these are further investigated with a biopsy to determine if they are malignant. Biopsy is a cost and time-demanding procedure, often requiring to be undertaken under general anesthesia in the operating room with related risks for the patient. Moreover, it is often an inefficient procedure, with a non-irrelevant rate of false negatives causing substantial delays in patients' treatment. For these reasons, the literature underscores the importance of adopting prevention-focused approaches to mitigate the need for such interventions (Żurek et al., 2022). In this scenario, computer-assisted systems have demonstrated potential for providing support to physicians and less-experienced personnel in detecting lesions. Nevertheless, they are limited by their focus on only one type of imaging or histological category. Furthermore, current approaches do not consider factors, such as different lesion sizes relative to the endoscope field of view, that strongly affect the detection accuracy. Indeed, the effectiveness of these DL methods increases when the lesion is close to the camera, suggesting the need for improvements to ensure reliable performance, especially when the lesion is small. Given the premises, this study proposes a real-time DL approach for laryngeal lesion detection from multi-center data acquired by flexible or rigid endoscopes in WL or NBI optical mode. To enhance small and early-stage lesion detection, a Super-Resolution Enhanced-YOLO network (SRE-YOLO) was further designed.

This chapter is organized as follows: in the Section 5.2, data curated and used to develop the detection algorithm are described (Subsection 5.2.1). Then Subsections 5.2.2 and 5.2.3 provide an explanation of the architecture of the baseline YOLO model, and the proposed SRE-YOLO variant. Data processing, implementation details, and performance metrics are

documented in Section 5.3. Lastly, detection results and further discussions are introduced in Sections 5.4, 5.5.

The work described in this chapter has been published in the following peer-reviewed journal article:

1. Baldini, C., Migliorelli, L., Berardini, D., Azam, M. A., Sampieri, C., Ioppi, A., Srivastava, R., Peretti, G., & Mattos, L. S. (2025). Improving real-time detection of laryngeal lesions in endoscopic images using a decoupled super-resolution enhanced YOLO. Computer Methods and Programs in Biomedicine, 260, 108539.

The code used for the SRE-YOLO implementation was also released, and it is now available online at https://github.com/ChiaraBaldini/SRE-YOLO.git.

5.2 Materials and Methods

5.2.1 Dataset

The dataset was obtained as described in Section 3.2 of Chapter 3. After annotation and quality checking, two subsets were defined starting from the original collection:

- *High-Resolution (HR) subset*: frames with at least one dimension $\geq 1,280$ pixels, used exclusively for YOLO baseline selection, pre-training analysis, and the SRE-YOLO experiment. It offered sufficient spatial detail to reliably evaluate detection performance, especially for small and early-stage lesions. The final *HR subset* consists of 3,452 HR frames for training, 269 for validation, and 171 for testing;
- *Full dataset*: all available checked frames from *Genova-Larynx, Laryngoscope8, and Lucknow-Larynx*, used for the selected YOLOv8n baseline training and performance analysis on general lesion and *Low-risk* versus *High-risk* lesions localization. The final multi-center dataset consists of 5,357 images with bounding box annotations for the eight possible laryngeal lesions, divided in a patient-wise manner into 4,228 frames for training, 792 for validation, and 337 for testing.

This separation ensured that the SRE-YOLO model was evaluated only under conditions compatible with its architectural requirements (high-resolution frames), while baseline YOLO models remained fully generalizable to all image resolutions.

To test the proposed SRE-YOLO, an external dataset, referred to as *ENDO-LC ext*, consisting of 149 HR frames in WL and NBI modalities, was collected at the University of

Health Sciences, Istanbul (Turkey), and manually annotated by clinical partners from the Units of Otolaryngology and Head and Neck Surgery of the IRCCS San Martino Hospital, University of Genova (Italy). More details about this set of data are available in Table 3.1 at the voice *Istanbul-Larynx*. Following the categorization proposed for the COCO dataset (Lin et al., 2014), this dataset includes small lesions (bounding box area < 32×32 pixels) and medium lesions (32×32 < bounding box area < 96×96).

5.2.2 YOLOv8: Architecture Overview

The YOLOv8 model for object detection is available in several sizes, including nano, small, medium, large, and extra large. Each version differs in terms of network depth and width, and is optimized to offer a specific balance of speed, size, and accuracy. The YOLOv8n is the smallest version and highly prioritizes speed and efficiency over accuracy, making it especially suitable for real-time use in environments with limited resources. The architecture of YOLOv8n can be partitioned into three main components: the backbone, the neck, and the head. The backbone is responsible for analyzing input images to extract both shallow and deep features. As in (Wang et al., 2021), it is based on the CSPDarknet53 architecture, which integrates Cross Stage Partial (CSP) connections into the Darknet53 network (Wang et al., 2020). This network consists of multiple convolutional CSP-sized stages, referred to as C2F stages, each composed of groups of residual bottleneck blocks. Between each stage, there is an interleaving 3×3 convolutional layer with a padding of 1 and a stride of 2. The configuration of these stages defines the main structure of YOLOv8, and the nano YOLOv8n comprises four stages with 1,2,2 and 1 bottleneck blocks, respectively. The backbone ends with the Spatial Pyramid Pooling Fast (SPPF) module, an advanced variant of the traditional SPP module (He et al., 2015). SPPF enhances the YOLOv8n capability to detect objects of various sizes and shapes by pooling the feature map at different scales and fusing them. The neck of the YOLOv8n processes features from three different levels of the backbone within a Path Aggregation Network (PANet)-like structure (Liu et al., 2018) using C2F stages. This structure, aggregating high- and low-level features in a bidirectional way, enriches spatial details and significantly enhances the detection. The head of YOLOv8n outputs detections across small, medium, and large scales. Each scale features two 3×3 convolutional layers, followed by two parallel 1×1 convolutions for bounding box coordinates regression (i.e., x, y, width (W), and height (H)) and object classification into the lesion category. The YOLOv8n predictions are refined with the Non-Maximum Suppression (NMS) algorithm to retain only

the most promising candidate bounding boxes. Figure 5.1 illustrates the overall architecture of the YOLOv8n model.

5.2.3 Super-Resolution Branch Integration

Detecting small objects in real-time applications remains a challenge. Here, an additional high-resolution branch was appended to the YOLOv8n structure to address this problem, inspired by the work in related fields of research (Berardini et al., 2025; Zhang et al., 2023a). The SR branch was used during the training phase and was subsequently discarded during inference. This enhanced lesion detection accuracy without adding to the computational load of the detection architecture. By taking as input feature maps obtained from YOLOv8n, the branch aims to reconstruct the super-resolution frame comparable with the ground truth image. The SRE-YOLO exploits the feature maps provided by the convolutional layers of the backbone and applies two consecutive steps of encoding and decoding (Fig. 5.1). The output of the eighth block of the backbone is upsampled and then combined with the output from the fourth block of the same backbone. By passing through two convolutional operations with kernel and stride equal to 1 to reduce the dimensionality, the encoded feature maps are decoded via a deep structure that evokes the Enhanced Deep Super-Resolution residual network (EDSR) (Lim et al., 2017). The output of the additional SR branch ($X_{SRoutput}$) corresponds with an HR frame of size 2W×2H of the input given to the backbone, whose dimensions are W×H. The similarity between the ground-truth frame (X) and $X_{SRoutput}$ is computed in terms of L1 loss (Eq. 5.1) and used as a learning component during the training phase.

$$L_{SR} = \|X - X_{SRoutput}\|_1 \tag{5.1}$$

Taking into consideration the L_{bbox},L_{dfl}, *and* L_{cls} loss terms implemented in the official Ultralytics GitHub repository[1], the modified loss used during the training of the SRE-YOLO model can be represented as:

$$Loss = c_1 \cdot L_{SR} + c_2 \cdot L_{bbox} + c_3 \cdot L_{dfl} + c_4 \cdot L_{cls} \tag{5.2}$$

where *c1* was set to 0.1 after a proper tuning, and the YOLO default values of 7.5, 1.5, and 0.5 were chosen for *c2, c3, and c4*, respectively. The bounding box regression loss L_{bbox} is formulated as an IoU-based loss between predicted and ground-truth bounding boxes.

[1] https://github.com/ultralytics/yolov5

The distribution focal loss L_{dfl} models bounding box coordinates as discrete probability distributions to improve localization precision. The classification loss L_{cls} is implemented as a binary cross-entropy loss over the predicted class probabilities. These loss terms follow the definitions adopted in the official YOLO implementation. The overall architecture of the proposed SRE-YOLO model is illustrated in Fig. 5.1.

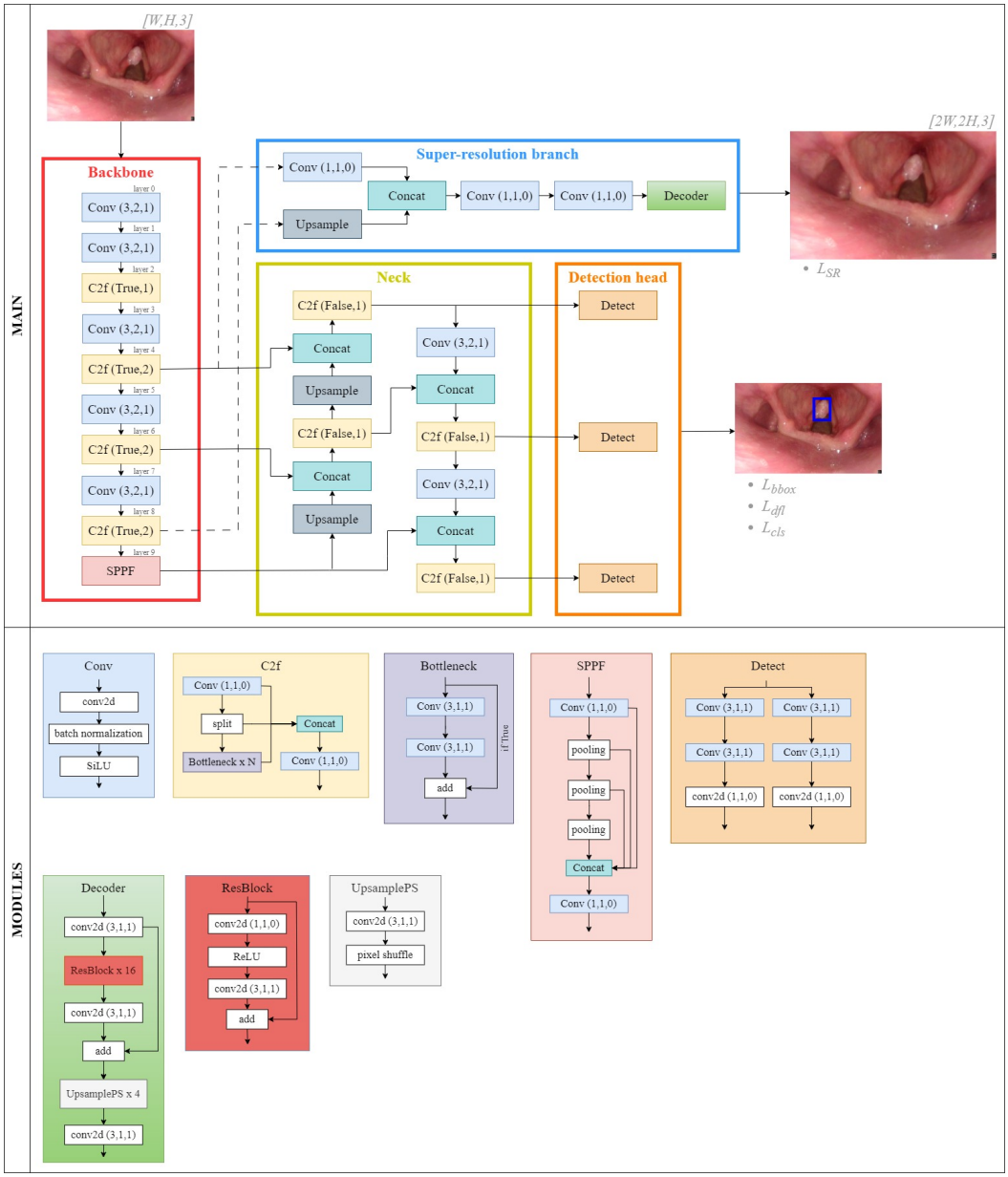


Figure 5.1 Architecture of the proposed SRE-YOLO model. The Super-Resolution (SR) branch is highlighted in blue. During inference, the SR branch is removed to maintain the real-time performance of the YOLO baseline.

5.3 Experimental Protocol

5.3.1 YOLO Architecture Benchmarking

Several YOLOv8 variants and alternative pre-training configurations were compared on the *HR subset* in order to identify the architecture offering the best compromise between accuracy, computational efficiency, and real-time feasibility in generally detecting laryngeal lesions.

Baseline Model Selection

The first experiment dealt with the comparison of the YOLO variants by evaluating different baseline models with varying complexities to determine their effectiveness. These models included the baseline YOLOv8-nano (YOLOv8n) and YOLOv8-small (YOLOv8s) pre-trained on the widely used COCO dataset (Lin et al., 2014), while more complex models were excluded due to the requirements of maintaining the complexity as limited as possible for fast inference. In addition, the YOLOv5 counterparts, i.e., YOLOv5n and YOLOv5s, were evaluated.

Pre-training Analysis

In order to analyze how the pre-training can influence the baselines' performance, an additional inter-domain transfer learning step was implemented and compared with no transfer learning option and the default COCO-based pre-training. Indeed, while the COCO dataset provides a wide variety of visual features and scenarios, it lacks specific representations of medical contexts, particularly in the narrow and homogeneous domains of endoscopic images. Consequently, the HyperKvasir dataset (Borgli et al., 2020), which contains gastrointestinal examinations that closely mimic the textural and color patterns found in laryngeal endoscopy, was downloaded and employed for the pre-training of the detection model. Data were collected using standard endoscopy equipment from Olympus[1] and Pentax[2] at the Department of Gastroenterology, Bærum Hospital, Vestre Viken Hospital Trust, Drammen (Norway). Despite the large size of the dataset, it includes corresponding bounding boxes only for 1,000 JPEG compressed images from the polyp class. Frame sizes are varied, ranging from 332×487 to 1,920×1,072 pixels. Therefore, after selecting the best performing model from the previous experiment as a baseline, its performance was tested by:

[2]Pentax Medical Europe, Germany

1. Initializing its weights using the standard Kaiming initialization (regarding this, we used the symbol "-" in the subsequent tables);
2. Pre-training on the COCO dataset;
3. Pre-training on the HyperKvasir dataset;
4. Pre-training on the combined COCO and HyperKvasir datasets.

5.3.2 Super Resolution Integration

To further improve small-lesion detection, YOLOv8n was extended with a decoupled SR branch, resulting in the SRE-YOLO model. The SR module enhances fine-grained spatial information without increasing inference-time cost, making it suitable for clinical use. Several ablation experiments were conducted to optimize the SR branch design and integration strategy.

Super-Resolution (SR) Branch Placement

To investigate the influence of feature-map depth from the shared backbone on detection performance, the performance of the *configuration 4-8* was compared with that of the *configuration 2-6*. Here, *configuration 4-8* refers to YOLOv8 + SR_{48deep}, where the SR branch processed feature maps from the fourth and eighth backbone layers. The term "deep" denotes the use of the deeper super-resolution branch described in Section 5.2.3, which is based on an EDSR-inspired architecture. Likewise, *configuration 2-6* refers to YOLOv8n + SR_{26deep}, where the SR branch processed feature maps from the second and sixth backbone layers.

Super-Resolution (SR) Branch Depth

To assess the impact of the SR branch depth on detection performance, two different versions of SR branch architecture were proposed and tested: *deep* versus *light*. In the latter, the encoder comprised two 1×1 convolutional layers that supplemented the backbone. The decoding process had three transposed convolutional layers, each with a kernel size of 4×4 and a stride of 2. The main difference between the *deep* and *light* structures of the decoder dealt with the respective presence or absence of the multiple residual blocks, and the upsampling methodology used. EDSR in the *deep* version used multiple residual blocks with short connections to make the optimization easier and removed batch normalization

to retain the range flexibility and reduce memory usage. Here, upsampling was performed via pixel shuffle operations, i.e., efficient sub-pixel convolution layers, which learn an array of filters to upscale the low-resolution feature maps into the HR output. Thus, the results of the YOLOv8n + SR_{48deep} were compared with those of YOLOv8n + $SR_{48light}$ with both baselines pre-trained on the COCO dataset.

SRE-YOLO Pre-training Configuration

In the final experiment, both the SR branch and the YOLOv8n baseline underwent studies to assess the best pre-training configuration. Specifically, the configurations tested included:

(i) Both architectures pre-trained on the COCO dataset, proposed as SRE-YOLO;

(ii) Both architectures pre-trained on COCO and subsequently fine-tuned on the HyperKvasir dataset;

(iii) Both architectures exclusively pre-trained on the HyperKvasir dataset;

(iv) YOLOv8n baseline pre-trained on COCO, while the SR architecture with weights initially set according to the Kaming initialization standard.

For point (ii), two different strategies were investigated, since the HyperKvasir dataset available online includes a labeled portion of the data for segmentation and detection purposes (1,000 frames) and a larger collection of unlabeled frames. Starting from the model pre-trained on COCO (both the baseline and the SR), the effect of the pre-training obtained using only the labeled portion or both the unlabeled, in a self-supervised manner, and labeled data from the HyperKvasir dataset was evaluated. To this specific end, two distinct phases were implemented:

(ii-a) The baseline and the SR branch were initialized with COCO weights, and then the SR branch was trained, due to the lack of ground truth for the YOLO detector training, with 10,000 frames from the HyperKvasir unlabeled portion. This pretraining phase updated the SR branch weights using the unlabeled data in a self-supervised manner. The (supervised) detection-related loss terms were set to zero, as they required bounding-box annotations that this data did not provide.

(ii-b) The entire model was fine-tuned using only the labeled subset of the HyperKvasir dataset ("HyperKvasir labeled"). This fine-tuning phase, built on the model pre-trained with unlabeled data, used labeled data to enhance lesion performance through supervised learning ("HyperKvasir unlabeled + labeled").

5.3.3 *Full Dataset* Experiments

To evaluate the generalization capability of the YOLOv8n baseline, selected based on HR-subset performance, an additional experiment was conducted on the *Full dataset*, which includes all WL and NBI frames regardless of resolution. This allowed validation in a more realistic and heterogeneous setting, reflecting the variability of routine clinical practice, of the model's capacity in generally detecting lesions and classifying them into two risk categories: *Low-risk* and High-risk.

General *Lesion* Detection

The selected YOLOv8n baseline was trained from COCO weights on the *Full dataset* to assess its generalization capability in detecting laryngeal lesions across a wide range of image resolutions and qualities. This experiment aimed to validate the model's effectiveness in a realistic clinical setting, where images may vary significantly due to different acquisition devices and conditions.

Low-risk versus *High-risk Lesion* Detection

Low-risk lesions included benign conditions such as Reinke's edema, granuloma, cyst, nodules, papilloma, and polyps. *High-risk* lesions encompassed malignant or potentially malignant conditions, including SCC, and leucoplakia. Distinguishing these two categories is clinically essential, as the diagnostic pathway, treatment planning, and surgical management differ substantially between benign and malignant or pre-malignant lesions. Early identification of cases at high risk of malignancy directly influences the urgency of intervention or biopsy. After having differentiated the data based on the lesion classes annotated by clinicians into the two main risk categories, the YOLOv8n model was trained starting from COCO weights.

5.3.4 Data Pre-processing

As images from the three *Genova-Larynx, Laryngoscope8, Lucknow-Larynx* datasets had different sizes, the images of the *HR subset* were resized to 1,280×1,280 pixels, and this size was used as ground truth for the SR branch (2H×2W, Fig. 5.1). In parallel, a low-resolution version of each frame was computed by downsampling it to 640×640 (H×W) pixels using bilinear interpolation, which served as the input size for the baseline architectures. To obtain the ground truth for the SR component, this process was also applied to the COCO

and HyperKvasir datasets. Similarly, to ensure uniformity, all images of the *Full dataset* were automatically resized to 640×640 pixels when provided to the YOLO baseline. Data augmentation techniques were applied on-the-fly during training to enhance data variability, including spatial transformations, translations, scaling, horizontal flipping, and mosaic augmentations following the official YOLO implementation[1].

5.3.5 Training Settings

All the training experiments were conducted using PyTorch on a Dell Precision 7820 equipped with a 48 GB NVIDIA RTX A6000 GPU. After a comprehensive hyperparameter tuning, the AdamW optimizer with an initial learning rate (lr_0) equal to 5e-2, a decay of 0.0005, and a minimal learning rate of $0.2 \times lr_0$ were adopted for HR subset-related experiments. The cosine annealing scheduler was employed to adjust the learning rate during training. YOLO base models and SRE-YOLO were trained for 100 epochs using a batch size of 16. However, early stopping after 50 epochs without improvements to the validation metrics was implemented to prevent overfitting. For the experiments on the *Full dataset*, the same optimizer and learning rate settings were used, but the models were trained for longer (1,000 epochs) with the same early stopping rule of HR training, autobatch function, and a lr_0 equal to 1e-3.

5.3.6 Performance Metrics

To assess the detection performance, the mean AP using an IoU threshold of 50% ($\mathrm{mAP}_{\mathrm{IoU=0.5}}$) was adopted, as described in Eq. 2.6. Additionally, the inference speed was measured in terms of FPS and the number of billion floating point operations (GFLOPs) to compare the SRE-YOLO real-time applicability and computational complexity with the other tested approaches. Inference speed evaluations were carried out on the same hardware configuration mentioned in Subsection 5.3.5. The FPS was computed by averaging the time taken to process all images in the test set.

Table 5.1 Results from the baselines, namely, YOLOv8n, YOLOv8s, YOLOv5n, YOLOv5s, comparison on the *High-Resolution (HR) subset*. Efficacy performance was assessed via the mean Average Precision (mAP) computed using an Intersection over Union (IoU) threshold of 0.50 ($mAP_{@IoU=0.5}$). Efficiency performance was assessed via the billion floating point operations (GFLOPs) and the Frames Per Second (FPS).

Model	Pre-training baseline	$mAP_{@IoU=0.5}$	GFLOPs	FPS
YOLOv5s	COCO	0.71	16	47
YOLOv5n	COCO	0.70	**4**	**65**
YOLOv8s	COCO	0.77	29	38
YOLOv8n	**COCO**	**0.77**	8	59

5.4 Results

5.4.1 YOLO Architecture Benchmarking

Baseline Model Selection

Table 5.1 shows the results from the comparison of baselines (YOLOv8n, YOLOV8s, YOLOv5n, YOLOv5s). The YOLOv8n outperforms YOLOv5s and YOLOv5n. Indeed, the $mAP_{@IoU=0.5}$ is reduced by 7% for YOLOv5n and 6% for YOLOv5s when compared to YOLOv8n. However, YOLOv8n and YOLOv8s achieve the same $mAP_{@IoU=0.5}$ of 0.77.

When assessing the efficiency, YOLOv8n and YOLOv5n have the lowest GFLOPs (equal to 8 and 4, respectively) and the highest FPS (equal to 59 and 65, respectively) with respect to YOLOv5s (GFLOPs=16, FPS=47) and YOLOv8s (GFLOPs=29, FPS=38). Thus, when seeking the optimal trade-off between efficiency and effectiveness, YOLOv8n emerged as the optimal choice.

Pre-training Analysis

Table 5.2 summarizes the results obtained by the YOLOv8n baseline with a different pre-training. The table specifically shows how YOLOv8n performs on the *High-Resolution (HR) test set* under different scenarios: (i) with weights initialized using standard Kaiming initialization, (ii) after pre-training on the HyperKvasir dataset, (iii) after pre-training on COCO, followed by fine-tuning on the HyperKvasir dataset, and (iv) after pre-training only on the COCO dataset. The YOLOv8n pre-trained only on the COCO dataset surpasses the others, giving the highest $mAP_{@IoU=0.5}$ of 0.77.

Table 5.2 Results from the YOLOv8n baseline pre-training when trained and tested on the *High-Resolution (HR) subset*. Efficacy performance was assessed via the Average Precision (mAP) computed using an Intersection over Union (IoU) threshold of 0.50 ($mAP_{@IoU=0.5}$).

Model	**Pre-training baseline**	**$mAP_{@IoU=0.5}$**
YOLOv8n	-	0.71
YOLOv8n	HyperKvasir	0.72
YOLOv8n	COCO+HyperKvasir	0.75
YOLOv8n	**COCO**	**0.77**

Table 5.3 Results from different placements of the Super Resolution (SR) branch. The performance of the *configuration 4-8* (YOLOv8n+SR_{48deep}) with that of the *configuration 2-6* (YOLOv8n+SR_{26deep}) were compared. In the first architecture, the SR branch processes feature maps from the fourth and eighth layers of the YOLOv8n backbone, whereas in the second architecture, it processes those from the second and sixth layers. The mean Average Precision (mAP) was computed using an Intersection over Union (IoU) threshold of 0.50 ($mAP_{@IoU=0.5}$).

Model	**Pre-training baseline**	**$mAP_{@IoU=0.5}$**
YOLOv8n + SR_{26deep}	COCO	0.80
YOLOv8n + SR_{48deep} (SRE-YOLO)	**COCO**	**0.82**

5.4.2 Super Resolution Integration

Super-Resolution Branch Placement

Table 5.3 compares the performance of the SR branch in its *configuration 4-8* (YOLOv8n + SR_{48deep}) with that of the *configuration 2-6* (YOLOv8n + SR_{26deep}). The architectures differ in the feature maps in input from the backbone. YOLOv8n + SR_{48deep} outperformed its akin by 2 percentage points, achieving a $mAP_{@IoU=0.5}$ equal to 0.82.

Super-Resolution Branch Depth

In Table 5.4, the results from the *deep* variant of the SR branch were compared against those of the *light* one. The architectures differ in the encoder-decoder structure. The YOLOv8n + SR_{48deep} outperformed by 2 percentage points the performance of YOLOv8n + $SR_{48light}$, which achieves $mAP_{@IoU=0.5}$ equal to 0.80.

Table 5.4 Results from the comparison of the Super-Resolution (SR) branch architectures: performance ($mAP_{@IoU=0.5}$) of the proposed *deep* (YOLOv8n+SR_{48deep}) version was compared with that of the *light* one (YOLOv8n+$SR_{48light}$)

.

Model	**Pre-training baseline**	**$mAP_{@IoU=0.5}$**
YOLOv8n + $SR_{48light}$	COCO	0.80
YOLOv8n + SR_{48deep} (SRE-YOLO)	**COCO**	**0.82**

Table 5.5 Analysis of the pre-training effect on detection performance. The subscripts *labeled* and *unlabeled* were included to improve readability and indicate whether the Super-Resolution (SR) branch pre-training was conducted (i) on the labeled portion or (ii) on the unlabeled and subsequently labeled portion of the HyperKvasir dataset.

Model	**Pre-training baseline**	**Pre-training SR**	**$mAP_{@IoU=0.5}$**
YOLOv8n + SR_{48deep}	$HyperKvasir_{labeled}$	$HyperKvasir_{labeled}$	0.72
YOLOv8n + SR_{48deep}	COCO	COCO + $HyperKvasir_{unlabeled + labeled}$	0.79
YOLOv8n + SR_{48deep}	COCO + $HyperKvasir_{labeled}$	COCO + $HyperKvasir_{labeled}$	0.80
YOLOv8n + SR_{48deep}	COCO	-	0.82
SRE-YOLO	**COCO**	**COCO**	**0.82**

SRE-YOLO Pre-training Configuration

Table 5.5 displays the outcomes of assessing the most effective pre-training configuration for both the baseline (YOLOv8n) and the SR branch (SR_{48deep}). The lowest performance is achieved when both the baseline and SR were pre-trained on the HyperKvasir ($mAP_{@IoU=0.5}$=0.72). The performance of pre-trained architectures on COCO and fine-tuned on HyperKvasir was slightly better, with an $mAP_{@IoU=0.5}$ equal to 0.80 when only the labeled portion of this dataset was considered, and an $mAP_{@IoU=0.5}$ equal to 0.79 when the model was fine-tuned first on the unlabeled HyperKvasir frames and then on those provided by the official repository with the corresponding labels. Pre-training the baseline only on the COCO dataset enhanced detection performance, achieving an $mAP_{@IoU=0.5}$ of 0.82, regardless of the pre-training applied to the SR branch.

Additional validation and qualitative results

The *ENDO-LC ext* subset of data with small and medium lesions was used to test the enhancement in performance and generalization capabilities of the baseline YOLOv8n model through the integration of the SR branch. The outcomes (Table 5.6) compare the performance of the SRE-YOLO, YOLOv8n + $SR_{48light}$, and YOLOv8n + SR_{26deep} models against their

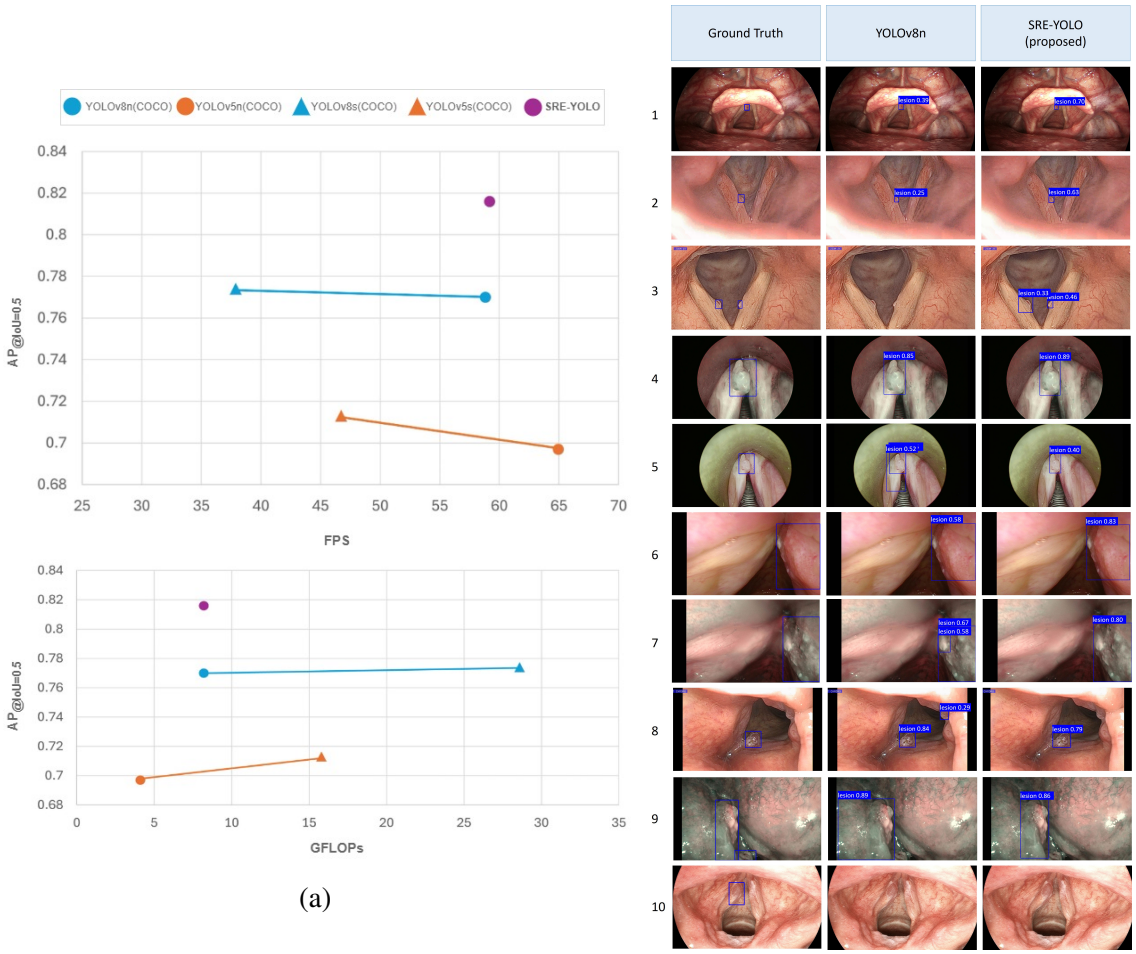
Table 5.6 Performance evaluation in terms of the mean Average Precision (mAP) computed using an Intersection over Union (IoU) threshold of 0.50 ($mAP_{@IoU=0.5}$) for the proposed SRE-YOLO, YOLOv8n + $SR_{48light}$, and YOLOv8n + SR_{26deep} models on the *ENDO-LC ext* test set, divided into small (bounding box area $< 32 \times 32$) and medium-sized ($32 \times 32 <$ bounding box area $< 96 \times 96$) lesions.

Model$_{\text{tested on } ENDO\text{-}LC\ ext}$	**Small lesions**	**Medium lesions**
YOLOv8n	0.66	0.80
YOLOv8n+$SR_{48light}$	**0.82**	0.77
YOLOv8n+SR_{26deep}	0.74	0.82
SRE-YOLO	0.80	**0.82**

baseline YOLOv8n pre-trained on COCO. The *ENDO-LC ext* set, which includes data from patients not part of the *HR subset*, revealed that SRE-YOLO enhanced mainly the small-sized lesion detection (+15% of $mAP_{@IoU=0.5}$), overcoming the baseline even on external frames. The other models also demonstrated improved detection of small lesions relative to the baseline, while gains for medium-sized lesions were modest.

Figure 5.2a displays a comparison of performances between the proposed SRE-YOLO and baseline models, including YOLOv5n, YOLOv5s, YOLOv8s, and YOLOv8n. The two scatterplots illustrate the relation between the $mAP_{@IoU=0.5}$ and efficiency-related metrics, i.e., FPS (upper plot) and GFLOPs (lower plot). The SRE-YOLO (purple circle), achieved a $mAP_{@IoU=0.5}$ equal to 0.82, with 59 FPS (as shown in the upper plot), and requires 8 GFLOPs (as reported in the lower plot). Such results are compared with those of YOLOv8n and YOLOv8s (blue circle and triangle, respectively), and of YOLOv5n and YOLOv5s (orange circle and triangle, respectively). The figure shows that the SRE-YOLO architecture outperformed its counterparts, particularly the YOLOv8n (i.e., SRE-YOLO baseline), in terms of quantitative metrics, while keeping the same effectiveness as YOLOv8n.

Figure 5.2b presents a qualitative comparison of the performance of the SRE-YOLO and YOLOv8n in detecting laryngeal lesions of varying apparent sizes, using frames obtained with different tools and optical modalities. The first column displays the original frames with ground-truth bounding boxes, while the second and third columns show the lesion detection results from YOLOv8n and SRE-YOLO, respectively. The results indicate that SRE-YOLO consistently outperformed YOLOv8n, particularly in detecting small lesions that were challenging to identify. For instance, in rows 1, 2, 4, 5, 8, and 9, SRE-YOLO successfully detected small polyps and SCCs that YOLOv8n missed. However, there were instances where both models failed to detect lesions, as illustrated in row 10.



(a) (b)

Figure 5.2 (a) Scatter plots of the efficacy-efficiency performance from the baselines comparison (i.e., YOLOv5n, YOLOv5s, YOLOv8s, YOLOv8n) against the proposed SRE-YOLO. The plot above shows the Frame Per Second (FPS) against the mean Average Precision (mAP) computed using an Intersection over Union (IoU) threshold of 0.50 ($mAP_{@IoU=0.5}$). The graph below presents the billion floating point operations (GFLOPs) against the $mAP_{@IoU=0.5}$. The $mAP_{@IoU=0.5}$ of SRE-YOLO improved significantly, without causing an increase in GFLOPS or FPS metrics. (b) Samples for qualitative analysis. Row 1-2: Left and right vocal fold polyps. Row 3: Bilateral vocal fold polyps. Row 4-5: Intraoperative Narrow-Band Imaging (NBI) of a left vocal fold polyp and the corresponding White Light (WL)-imaging view. Row 6-7: Squamous Cell Carcinoma (SCC) captured with both WL and NBI modalities. Row 8: Left vocal fold polyp. Row 9: NBI view of a SCC arising from the left laryngeal vestibular fold. Row 10: Model failure case in detecting a left vocal fold polyp. The second and third columns illustrate the detection performance of the YOLOv8n (i.e., the baseline) and the proposed SRE-YOLO, respectively.

Table 5.7 Results from the YOLOv8n baseline training on the *Full dataset* for general *lesion* detection and *Low-risk* versus *High-risk* lesion detection. Efficacy performance was assessed via Precision, Recall, F1-score, and mean Average Precision (mAP), computed using an Intersection over Union (IoU) threshold of 0.50 ($mAP_{@IoU=0.5}$).

YOLOv8n for	**Precision (P)**	**Recall (R)**	**F1-score**	**$mAP_{@IoU=0.5}$**
Lesion detection	75%	**69%**	**72%**	0.73
Low-risk vs. *High-risk* detection	**79%**	65%	71%	**0.74**
· *Low-risk*	82%	74%	78%	0.81
· *High-risk*	75%	55%	63%	0.67

5.4.3 *Full Dataset* Experiments

General *Lesion* Detection

For general *lesion* detection on the full multicenter dataset, YOLOv8n pretrained on the COCO dataset achieved solid performance despite the high variability of the WL and NBI acquisitions, image resolutions and lesion appearances. As demonstrated in Table 5.7, when evaluated on 337 test frames of the *Full dataset*, the model reached an F1-score of 72% and a $mAP50_{@IoU=0.5}$ of 0.73. The relatively low recall reflects the challenging nature of the task and the high incidence of missed detections in low-resolution images, especially where small or poorly illuminated lesions were present.

Low-risk versus *High-risk* Lesion Detection

Using the same dataset as in the previous paragraph, a separate YOLOv8n model was trained to distinguish *Low-risk* from *High-risk* lesions, initially facing the important clinical bifurcation of the management pathways of benign lesions from those for malignant or pre-malignant conditions. The model achieved an F1-score of 71% and a $mAP50_{@IoU=0.5}$ of 0.74 (Table 5.7), confirming the ability to localize and categorize lesions according to their associated clinical risk.

5.5 Discussion

Accuracy in detecting laryngeal lesions during endoscopic examinations is strongly influenced by the clinician's experience and the intrinsic variability of image quality. Even with advanced optical modalities such as NBI, early-stage lesions may be easily overlooked, and less-experienced clinicians often face steep learning curves. In this context, DL-based lesion

detection systems can play a pivotal role by automatically highlighting suspicious regions, thereby reducing missed lesions and supporting consistent, real-time decision-making.

The experiments conducted in this chapter demonstrate that YOLOv8n, selected for its optimal balance between accuracy and efficiency, can already achieve solid performance on the multicenter *Full dataset*. Indeed, it obtained an $mAP50_{@IoU=0.5}$ of 0.73 for general *lesion* detection across both WL and NBI modalities. The relatively low Recall reflects the intrinsic difficulty of detecting small or poorly contrasted lesions in low-resolution or moderate-quality images, yet the model proved capable of identifying the majority of abnormalities in clear frames, with detection becoming increasingly accurate as the endoscope approaches the region of interest. Moreover, YOLOv8n was able to perform a first-pass risk stratification, reaching $mAP50_{@IoU=0.5} = 0.74$ when distinguishing *Low-risk* from *High-risk* lesions. A more detailed and fine-grained analysis of lesion types is addressed in the subsequent chapter on lesion classification. However, this initial risk analysis suggests that *High-risk* lesions, which have lower Recall values than *Low-risk* lesions, present more difficult and ambiguous cases, such as precancerous lesions and leucoplakia occurrences.

Regarding lesion detection, only a limited number of state-of-the-art methods have been evaluated on datasets of comparable scale (Table 2.2). Among these, (Wellenstein et al., 2023) applied a YOLOv5-based detector to 4,488 WL images, achieving an F1-score of 74% with real-time performance, while (Nie et al., 2025) reported a $mAP50_{@IoU=0.5}$ of 0.84 using a YOLOv8-based model trained on 1,353 WL and NBI images. In contrast, the YOLO-based detection framework proposed in this thesis was trained and validated on larger, heterogeneous, multicenter datasets that encompass both WL and NBI modalities, achieving performances comparable to approaches from the literature in both internal and external test sets, while maintaining real-time performance. Furthermore, the proposed super-resolution–enhanced design explicitly targets the detection of small and early-stage lesions, which remains a key open challenge in the current literature.

To further address the limitations of baseline detectors on challenging cases, particularly small lesions, this work introduced SRE-YOLO, a novel architecture in which a decoupled SR branch is integrated into the YOLOv8n backbone. The SR component, used only during training and discarded during inference, enabled the model to refine high-level feature extraction without increasing computational cost at test time. Experiments on the *HR subset* of data, specifically created based on image size to develop the SR branch, indicated that SRE-YOLO outperformed the YOLOv8n baseline, yielding a +5% improvement in $mAP50_{@IoU=0.5}$ and noticeably reducing false negatives for small lesions. This enhancement is further supported by qualitative examples, where SRE-YOLO provided more precise

bounding boxes and higher confidence scores while maintaining identical GFLOPs and FPS to the baseline during inference.

To summarize, training SRE-YOLO required restricting the dataset to *HR* frames, thus reducing the amount of available data. Furthermore, the current detection models do not yet incorporate a final step for assessing lesion severity directly within the detection stage, limiting their ability to perform only general abnormalities detection. Despite such limitations related to the SRE-YOLO implementation, the results of the baseline YOLOv8n on the *Full dataset* are aligned with the values reported in the current literature. Unlike most published approaches, which typically rely on constrained or single-center datasets, the baseline in this work was trained and evaluated on a substantially more heterogeneous and clinically realistic dataset. As such, it can address the lesion detection task more comprehensively, reflecting the true variability encountered in routine clinical practice and establishing a solid foundation for reliable AI-assisted laryngeal lesion detection.

Chapter 6

Lesion Characterization

6.1 Overview

The characterization of laryngeal lesions is a fundamental step in the endoscopic evaluation of the larynx, as it allows benign conditions to be differentiated from potentially malignant lesions and directly guides the diagnostic-therapeutic pathway. Correctly identifying the nature of a lesion influences critical decisions such as the need for biopsy, the frequency of follow-up, the indication for surgery, and the urgency of the intervention. In real clinical workflows, however, visual classification is a highly subjective process that is sensitive to the clinician's experience and can be extremely difficult, especially in the presence of small, early-stage lesions or those with overlapping characteristics. In this context, DL models capable of providing reliable, standardized, real-time classification can support specialists, particularly less experienced practitioners, by improving diagnostic accuracy and reducing inter-observer variability.

This chapter presents the materials and methods used in the development, training, and evaluation of the proposed AI modules for laryngeal lesion classification (Section 6.2). It first describes the employed dataset (Section 6.2.1), including its composition and annotation strategy. The following section (Sec. 6.2.2) details the methodologies adopted for the two main classification tasks addressed in this work: risk assessment and histological-type classification. Then, the experimental protocol is described, including data pre-processing pipelines, training settings, performance metrics, and statistical analysis procedures, in Section 6.3. It also outlines the design of the expert-AI comparative evaluation study conducted to benchmark the AI models against clinicians with varying levels of experience and GPT. Finally, Section 6.4 presents and discusses the results for each experimental test.

6.2 Materials and Methods

6.2.1 Dataset

Laryngeal images and related histopathological data of patients from several centers around the world between 2014 and 2023 were retrospectively selected according to the availability of pre-treatment recorded videolaryngoscopies in local databases. Section 3.3 of Chapter 3 describes the primary dataset used for the classification task. Specifically, these data were obtained from multiple sources, including the *Genova-Larynx* set, the *Lucknow-Larynx* dataset, the *Barcelona-Larynx* dataset, the *Athens-Larynx* set, and the *Laryngoscope8* dataset (Table 3.1). An additional dataset collected for this work was the *Pre-Training* dataset, formed by 8,651 frames automatically extracted, within a window of 10 frames, from the videos of the same patients of the *Genova-Larynx* dataset. This process considered only good-quality frames predicted using the AI model described in Chapter 4. These images were automatically labeled considering the overall diagnosis, i.e., lesion's type, linked to the specific video. These images were used as an in-domain pre-training strategy to enhance the model's performance without requiring manual labeling by the physicians. Finally, the *Genova-Larynx, Lucknow-Larynx, and Laryngoscope8* datasets were combined to create the *internal dataset*, which was utilized for training and internal testing of the CADx model. The *Athens-Larynx* and *Barcelona-Larynx* datasets served as *external set 1* and *external set 2* for assessing the CADx model's generalization capabilities.

6.2.2 Architectures

Binary risk assessment: *Low-risk* vs *High-risk* lesions classification

The risk assessment task aimed at classifying laryngeal lesions into two categories: *Low-risk (LR)* and *High-risk (HR)*. *LR* lesions included benign conditions such as *cysts, granulomas, polyps, papillomas, and Reinke's edema*, while *HR* lesions encompassed *leucoplakia* and *SCC*, which have a higher potential for malignancy and require more urgent clinical attention. The risk-classification model was based on EfficientNetV2-S, which was introduced in (Tan and Le, 2021) and selected for this task due to its strong trade-off between accuracy, number of parameters, and inference efficiency. The main components of the EfficientNetV2 architecture include:

1. MBConv (Mobile Inverted Bottleneck Convolution): This is the primary block used in EfficientNetV1. It features an initial 1×1 convolution to expand the number of

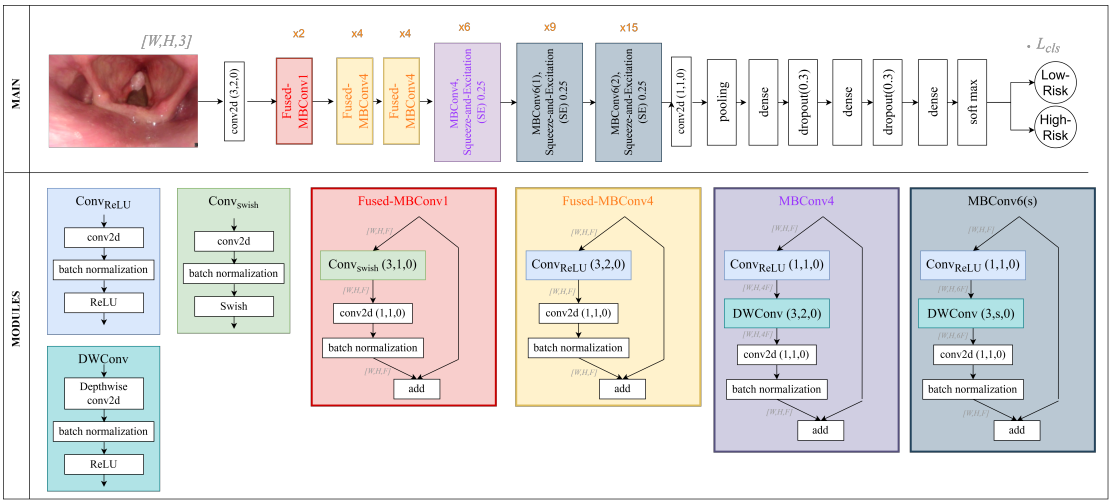


Figure 6.1 EfficientNetV2 architecture overview. The model consists of a series of Fused-MBConv and MBConv blocks, followed by a final classification head. The architecture is designed to optimize both accuracy and training speed through the use of fused operations and progressive scaling.

channels, followed by a depthwise convolution, and then a 1×1 convolution to project back to the desired output channels.

2. Fused-MBConv Blocks: Introduced in EfficientNetV2, this block combines the expansion (point-wise convolutions) and depth-wise convolutions into a single fused operation. It's designed to optimize for training speed by reducing the complexity and number of operations, especially in the early layers.

A non-uniform scaling strategy that gradually increases the depth, width, and image size is employed during training. This progressive scaling helps mitigate memory usage and training time, especially since aggressive scaling of image size can slow down training. A detailed description of the EfficientNetV2 architecture and its components can be found in Fig. 6.1. The model was initialized using ImageNet weights and adapted for the binary classification task by replacing the final fully connected layer with a two-unit softmax output.

Histological Type Classification

The histological type classification task aimed to discriminate between seven fine-grained categories of laryngeal lesions, namely:

1. *Cyst*

2. *Granuloma*
3. *Leucoplakia*
4. *Papilloma*
5. *Polyp*
6. *Reinke's edema*
7. *Squamous Cell Carcinoma (SCC)*

This level of detail goes beyond simple risk stratification and provides more detailed diagnostic information that can further influence treatment planning, follow-up frequency, and the need for biopsy or surgical intervention. Due to the high visual similarity and overlapping among several of these lesions, this task requires feature representations capable of capturing both local and global contextual cues. Hence, the ViTb16 model, a Vision Transformer (ViT) architecture originally introduced in (Dosovitskiy, 2020), was selected for its strong performance in image classification tasks and its ability to model long-range dependencies through self-attention mechanisms. The main components of the ViTb16 architecture include:

1. Patch Embeddings: The input image is divided into fixed-size patches (16×16 pixels for ViTb16), which are flattened and linearly projected into a learned embedding space, similar to tokens in natural language processing.

2. Positional Embeddings: Since transformers lack inherent spatial structure, learnable positional encodings are added to the patch embeddings to maintain information about the relative spatial arrangement of patches.

3. Classification (CLS) Token: A dedicated learnable token associated with the sequence, whose corresponding representation at the final layer is used for image-level classification.

4. Transformer Encoder: The hierarchical transformer encoder consists of multiple layers, each containing:

   Multi-Head Self-Attention (MHSA) Layers: These layers compute relationships between all tokens (patches + class token), capturing global relationships, contextual interactions, and long-range structural patterns that are critical in differentiating visually similar laryngeal lesions.

Feed-Forward Network (FFN): A position-wise Multi-Layer Perceptron (MLP) applied after the attention module, responsible for non-linear feature transformation within each token.

Layer Normalization and Residual Connections: These components help stabilize training and facilitate gradient flow through the deep architecture.

5. Classification Head: The final representation of the CLS token is passed through a linear layer to produce class probabilities for the seven histological categories.

A schematic overview of the ViTb16 architecture is provided in Fig. 6.2.

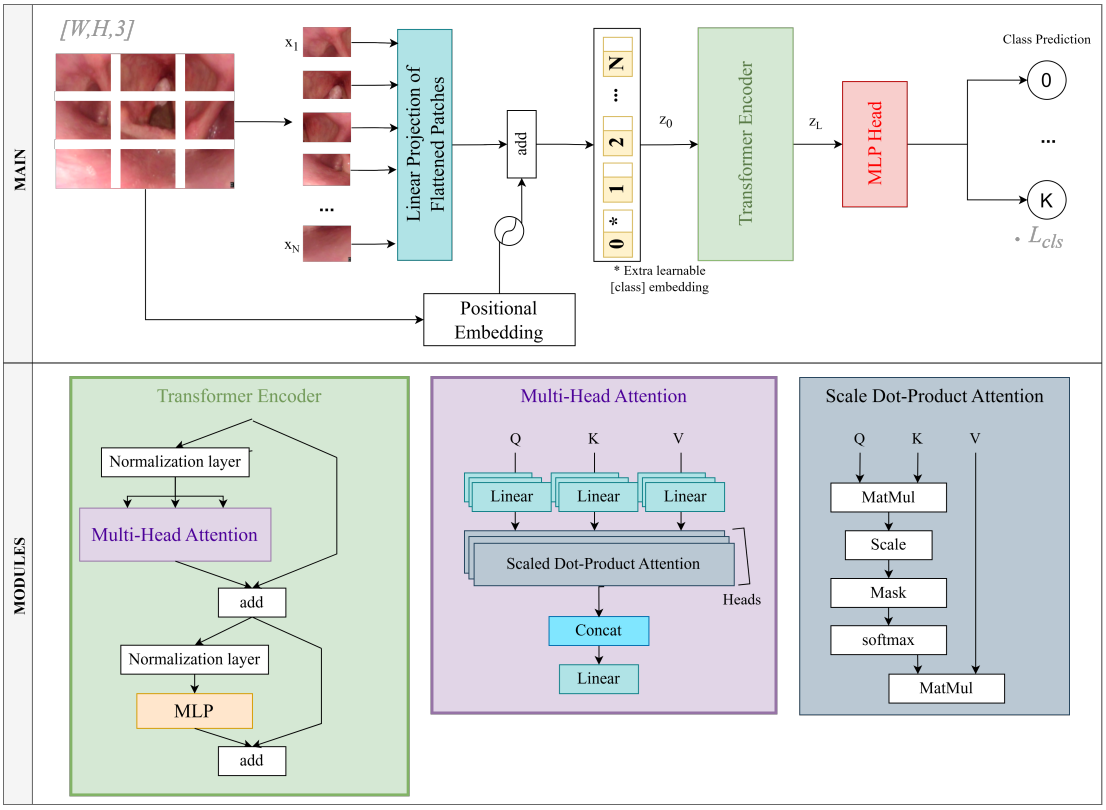


Figure 6.2 ViTb16 architecture overview: input image is split into 16×16 patches, embedded and enriched with positional encoding. The resulting token sequence is processed through multiple layers of multi-head self-attention and a feed-forward network. The final representation of the class token is passed to a linear classification head to predict one of the seven histological classes.

6.3 Experimental Protocol

6.3.1 Data Preprocessing

In line with similar studies (Srivastava et al., 2025; Yin et al., 2021), both model training and inference were performed using frames cropped around the lesion RoI based on the coordinates of annotated bounding boxes. In fact, the classification model is designed to work in a comprehensive pipeline following the previously described detection module (Chapter 5). The latter would flag the lesion on the image and submit the image within the bounding box to the classification model, which would ultimately analyze the lesion type. The cropped images were also resized through a zero-padding operation to 224×224 pixels for consistency, as the models' architecture used predefined input resolutions for their scaling strategies. In addition, data augmentation was used to increase the training dataset size by 30%. This was done using image transformations: rotations up to 10°, shift range up to 0.1 in both directions, shear up to 10%, zoom in the range $[0, 0.1]$, horizontal flipping, and random brightness changes. To develop the AI modules, the pre-processed, cropped frames from the *internal* were reorganized into three different sets for training (8,477), validation (2,415), and testing (1,226). This division followed a patient-level strategy to avoid having images from the same patients in different sets. The same pre-processing strategy was applied to the *external set 1* and *external set 2*, while the *Pre-Training* dataset was only resized to 224×224 pixels without any cropping or augmentation.

6.3.2 Training Settings

The models were initialized using ImageNet weights and further adapted to the laryngeal lesion classification tasks by replacing the final classification head with a fully connected layer of size two for binary and seven for histological characterization. Prior to supervised training, the backbone was also exposed to a pseudo-supervised pretraining phase using 8,651 WL and NBI video frames (*Pre-Training dataset*), which helped the model internalize domain-specific visual patterns and enhance feature extraction capabilities before fine-tuning on the labeled *internal dataset*. The DL models selected, EfficientNetV2S for binary risk assessment and ViTb16 for histological characterization, were trained using the following parameters: 100 epochs, batch size of 8, Adam optimizer with an initial learning rate of 1e-4, and exponential decay at a rate of 0.96. The hardware used for the classification module implementation was identical to that used in the previous chapters.

6.3.3 Performance Metrics

The performance was evaluated by comparing the predicted classes with the ground-truth classes. For every class, a confusion matrix with task-specific definitions of TP, TN, FP, and FN was used. The outcomes of this study were also analyzed using standard diagnostic metrics from the medical image analysis field, including Precision (Eq. 2.1), Recall (Eq. 2.2), F1-score (Eq. 2.3), and Accuracy (Eq. 2.7). These metric evaluates the overall correctness of the model's predictions by measuring the percentage of correctly classified instances across all the images. Moreover, the ROC curve is used to compute the Area Under the Curve (AUC). Such metrics were computed for both the internal and external test sets to assess models' generalizability.

Expert–AI Comparative Evaluation

To interpret the performance of the trained risk assessment model, a comparison with healthcare professionals and the ChatGPT-4o large language model was performed. For this, physicians with varying degrees of expertise in laryngoscopy were enrolled at the Hospital Clínic de Barcelona (Spain). Among the physicians, 3 were senior laryngologists with more than 10 years of experience in laryngoscopy, 3 were general otorhinolaryngologists, 3 were otorhinolaryngology senior residents with 3-4 years of training, and 3 were general practitioners with limited contact with laryngoscopy. None of the physicians had previously been exposed to any case in the test dataset. The 150-image Spanish *external set 2* was used for this purpose. Plane images without bounding boxes were shown to the raters, and they were asked to classify each image as *HR* (lesions that looked like a *leukoplakia* or *SCC*), *LR*, or healthy (no lesion detected). Physicians were divided into four groups: expert laryngologists, general otorhinolaryngologists, residents, and general practitioners. For each group of physicians, the most common class chosen by each category for every case was chosen as the final answer for that group. ChatGPT-4o was tested without previous training with the following English prompt: "If I show you some laryngeal endoscopy images, could you tell me which is the diagnosis among these conditions: laryngeal papillomatosis, Reinke's edema, granuloma, polyp, cyst, leukoplakia, carcinoma or healthy larynx". The images were submitted one by one, and each answer was collected. Since the binary classification model was specifically trained to distinguish between *LR* and *HR* classes, and the Barcelona dataset included healthy cases without annotated bounding boxes, these images were processed using the YOLO-based lesion detection algorithm developed for general lesion identification

Table 6.1 Performance metrics of the EfficientNetV2S model for *Low-risk* vs *High-risk* lesion classification on the *internal* and *external* test sets.

	Accuracy	**Precision**	**Recall**	**F1-score**	**AUC**
Internal test set					
Weighted average	90%	90%	89%	89%	0.89
Low-risk		91%	94%	90%	
High-risk		90%	84%	87%	
External set 1 (Athens-Larynx)					
Weighted average	85%	86%	85%	85%	0.85
Low-risk		79%	88%	83%	
High-risk		90%	83%	87%	
External set 2 (Barcelona-Larynx)					
Weighted average	88%	89%	88%	88%	0.88
Low-risk		92%	88%	90%	
High-risk		83%	89%	86%	

(Subsection 5.3.3 of Chapter 5) to extract potential classifier inputs. The image's final prediction was considered *healthy* if no detection occurred.

Statistical Analysis

Expert-AI performance comparisons based on Accuracy measures between the classification module, physicians, and the ChatGPT-4o model were conducted using McNemar's Test. To statistically compare the ROC curves between the *internal* and the two external datasets, pairwise comparisons of AUCs were conducted using a Z-test. Statistical significance was set at $p<0.05$. Statistical analyses were performed using Python version 3.9 and R Studio version 2024.09.1.

6.4 Results

6.4.1 Binary Risk Assessment: Performance Evaluation

On the *internal test set*, the CADx model performed well, reaching 90% Accuracy and AUC of 0.89, respectively (Table 6.1). When validated externally, the model performed similarly, with an Accuracy of 85% and an AUC of 0.85 ($p=0.07$) on the Greek dataset (*external set 1*), and an Accuracy of 88% and an AUC of 0.88 ($p=0.73$) on the Spanish dataset (*external set 2*). The confusion matrices for the three datasets are shown in Fig. 6.3.

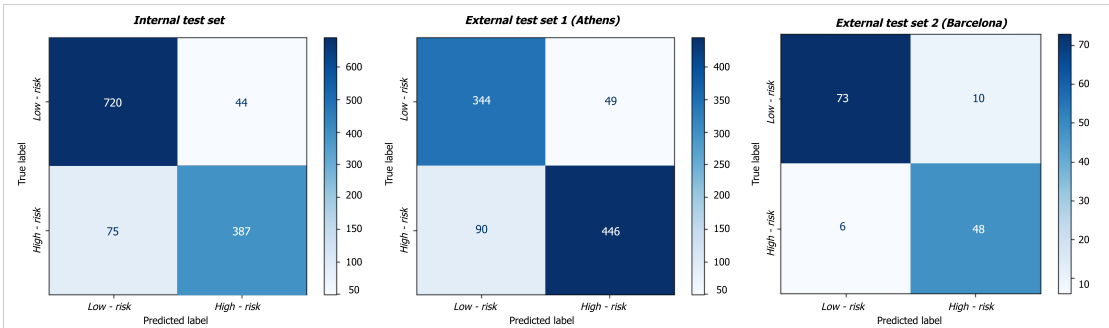


Figure 6.3 Confusion matrices for the EfficientNetV2S model when tested for *Low-risk* vs *High-risk* lesion classification on the *internal test set* (left), *external set 1* (center), and *external set 2* (right).

6.4.2 Histological Type Classification: Performance Evaluation

For the multi-class diagnostic task, the ViTb16 model was evaluated on the three test sets using seven lesion categories. As expected, this task was substantially more challenging than the binary risk assessment, due to class imbalance and high visual similarity among several lesion types. Indeed, on the *internal test set*, the model achieved a weighted F1-score of 65%. When tested externally on the Greek data (*external set 1*), performance remained consistent, with a weighted Precision, Recall, and F1-score of 66%, 67%, and 65%, respectively. On the Spanish dataset (*external set 2*), the Precision slightly increased, exceeding 70%, due to cleaner or less ambiguous data. However, the weighted F1-score reached 66%. Regarding the latter test set, some qualitative results are reported in Figure 6.5b. From the original image shown in the first column, the RoI lesion was extracted as illustrated in the second column. The attention maps are visualized in the third column, highlighting the features that the ViTb16 model focused on to make its predictions. Finally, the predicted histological class is reported in the last column. In the first row, a cyst was wrongly classified as a polyp, with the attention map focusing on the protrusion of the vocal fold without peaks of attention. The second row shows an SCC case, while the model identified features belonging to the leucoplakia class, as indicated by the attention map that highlights white spots of the image. The last two rows illustrate correctly classified lesions: a polyp captured with the NBI modality and an SCC in WL. Class-wise metrics for the two external sets are reported in Table 6.2, and the corresponding confusion matrices are illustrated in Fig. 6.4.

Table 6.2 Performance of the ViTb16 model for seven-class histological lesion classification obtained during *internal* and *external testing*.

	Accuracy	**Precision**	**Recall**	**F1-score**
Internal test set				
Weighted average	66%	65%	66%	65%
Class 0: *Cyst*		52%	71%	60%
Class 1: *Granuloma*		29%	16%	20%
Class 2: *Leucoplakia*		36%	32%	34%
Class 3: *Papilloma*		67%	43%	52%
Class 4: *Polyp*		72%	69%	71%
Class 5: *Reinke's edema*		73%	76%	74%
Class 6: *SCC*		72%	81%	76%
External set 1 (Athens-Larynx)				
Weighted average	67%	66%	67%	65%
Class 0: *Cyst*		69%	30%	41%
Class 1: *Granuloma*		57%	52%	54%
Class 2: *Leucoplakia*		62%	45%	52%
Class 3: *Papilloma*		27%	16%	20%
Class 4: *Polyp*		69%	90%	78%
Class 5: *Reinke's edema*		67%	40%	50%
Class 6: *SCC*		69%	84%	76%
External set 2 (Barcelona-Larynx)				
Weighted average	69%	72%	69%	66%
Class 0: *Cyst*		100%	40%	57%
Class 1: *Granuloma*		100%	82%	90%
Class 2: *Leucoplakia*		82%	64%	72%
Class 3: *Papilloma*		50%	6%	11%
Class 4: *Polyp*		55%	100%	71%
Class 5: *Reinke's edema*		69%	69%	69%
Class 6: *SCC*		66%	88%	66%

6.4.3 Expert-AI Comparative Evaluation

The study carried out to interpret the risk discriminative capability of the proposed solution involved clinicians and two AI systems: the proposed EfficientNetV2S classifier and ChatGPT. Four groups of human raters were considered: *general practitioners*, *otolaryngology residents*, *general otolaryngologists*, and *laryngologists*. Each participant independently classified the frames from the *Barcelona-Larynx* dataset into *Low-risk* or *High-risk*, without access to patient history or additional clinical information.

As shown in Table 6.3, the proposed classifier achieved an AUC of 0.88, performing comparably to expert laryngologists and outperforming general practitioners, residents, and ChatGPT. Figure 6.5a plots the ROC curves for all raters and models. The black point

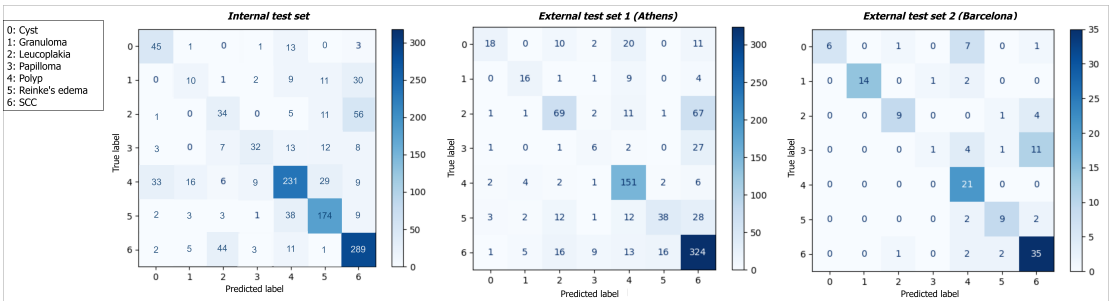


Figure 6.4 Confusion matrices of the ViTb16 model when tested for seven-class histological lesion classification on the *internal test set* (left), *external set 1* (center), and *external set 2* (right).

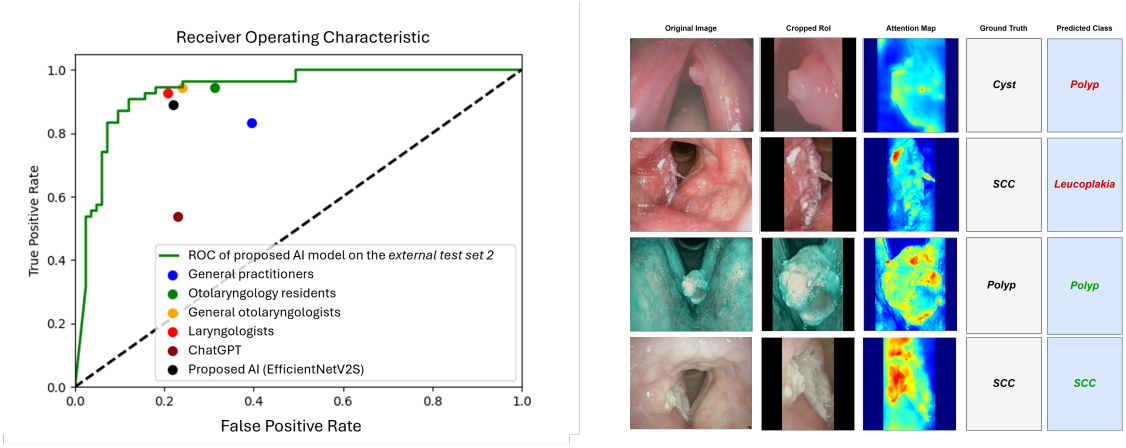
represents the ROC value relative to the proposed AI module: its location near the points related to expert and general otolaryngologists demonstrates that the proposed method can reach expert-level performance and may offer substantial assistance to less-experienced clinicians.

Table 6.3 Comparison between the proposed AI model, ChatGPT-4o, and clinicians with varying levels of expertise on the *external test set 2* from *Barcelona-Larynx* data. The Accuracy and p-values from McNemar's test comparing each rater group to the proposed AI module are reported.

Rater Group	**Accuracy**	**p-value**
Expert Laryngologists	93%	0.28
General Otolaryngologists	89%	1.00
Proposed AI (EfficientNetV2S)	**88%**	
Otolaryngology Residents	69%	0.87
General Practitioners	69%	**<0.05**
ChatGPT-4o	68%	**<0.05**

6.5 Discussion

The results presented in this chapter demonstrated that the proposed classification framework can effectively support diagnostic decision-making across multiple levels of granularity. The binary risk assessment model achieved high performance on both *internal* and *external* test sets, reaching Accuracy values comparable to those of experienced laryngologists, while significantly outperforming less-experienced clinicians and ChatGPT. These findings emphasize the clinical value of automated *High-risk* lesion identification as an early assessment step



(a) (b)

Figure 6.5 (a) ROC curves for clinicians at different levels of expertise, ChatGPT-4o, and the proposed AI model when discriminating between *Low-risk* and *High-risk* lesions on the *Barcelona-Larynx* dataset. (b) Qualitative results of the ViTb16 model for seven-class histological lesion classification on sample images from the *internal* and *external test sets*. The ground-truth and predicted labels are indicated in the last two columns, respectively. The first two rows show misclassifications (red), while the last two rows illustrate correct predictions (green). In the third column, the attention maps highlight the regions of the images that the model focused on for its predictions.

that can meaningfully reduce diagnostic variability and guide towards the correct treatment pathways.

Notably, both classification models also achieved inference speeds compatible with real-world clinical deployment, with average processing times of 0.005 seconds for binary risk classification and 0.007 seconds for the ViTb16–based multiclass task when evaluated on an NVIDIA Tesla T4 GPU.

With more than 20,000 WL and NBI images, our dataset represents the largest multicenter cohort for laryngeal lesion classification, spanning diverse populations across Europe and Asia. Nevertheless, the more challenging seven-class histological classification task demonstrated that, even with a large and heterogeneous dataset, fine-grained diagnosis remains inherently difficult. The ViTb16 model achieved robust performance overall, yet it produced a substantial number of errors on visually ambiguous or clinically subtle lesions, particularly those that are under-represented or exhibit significant overlap in appearance. These findings highlight the need for continued research, larger and more balanced multicenter datasets, and standardization of both data acquisition protocols and annotation guidelines to support the development of more reliable fine-grained diagnostic models.

In the literature (Table 2.3), for multiclass lesion classification, (Yin et al., 2021) reported an overall Accuracy of 73% on the *Laryngoscope8* WL dataset, while (Cho et al., 2021) achieved an F1-score of 88% using an EfficientNet-based architecture on WL images only. In the context of multicentric studies for binary risk assessment, (Yan et al., 2023) reported an Accuracy of 78%, whereas (Wei et al., 2024) and (Kang et al., 2025) achieved F1-scores of up to 89% and 87%, respectively. The classification framework proposed in this thesis, trained and validated on large, heterogeneous, multicenter datasets acquired under different acquisition conditions, achieved performance metrics within the range of reported state-of-the-art results for both binary risk stratification and multiclass lesion classification.

Importantly, the model outperforms general practitioners, otolaryngology residents, indicating that carefully trained DL models can effectively support clinical workflows, especially when operators lack extensive experience, for instance, in low-resource settings.

Chapter 7

Tumor Margin Definition

7.1 Overview

Accurate delineation of tumor margins is a fundamental requirement in both diagnostic and intraoperative laryngology. In a diagnostic context, precise segmentation enables the quantification of lesion extent and vascularization patterns, which are features that play a key role in early identification of malignancy. Intraoperatively, especially during TOLMS, the ability to identify the full spatial extent of a lesion is crucial for achieving complete resection while preserving as much healthy tissue as possible, thereby minimizing postoperative impairment of voice, swallowing, and breathing. However, margin estimation during endoscopy is extremely challenging due to illumination artifacts and variability, device heterogeneity, and subtle transitions between healthy and abnormal tissue. To comprehensively explore the capabilities of current and emerging segmentation methodologies for laryngeal tumor margin delineation, we systematically evaluated three families of methods: prompt-based foundation models, prompt-less SAM-based models, and prompt-less domain-specific CNNs.

The chapter is organized as follows: Section 7.2 describes the materials and methods used in the study, including dataset details and the architectures compared. Section 7.3 outlines the experimental protocol, including data pre-processing, training settings, and performance metrics. Section 7.4 presents the results of the segmentation experiments, while Section 7.5 discusses the findings and their implications for clinical practice.

The work described in this chapter has been published in the following peer-reviewed journal article:

- Sampieri, C., Azam, M. A., Ioppi, A., Baldini, C., Moccia, S., Kim, D., Tirrito, A., Paderno, A., Piazza, C., Mattos, L.S. & Peretti, G. (2024). Real-time laryngeal cancer

boundaries delineation on white light and narrow-band imaging laryngoscopy with deep learning. The Laryngoscope, 134(6), 2826-2834.

7.2 Materials and Methods

7.2.1 Dataset

As described in Section 3.4 of Chapter 3, the segmentation experiments were performed using three clinically annotated laryngeal datasets, including a large *Genova-Larynx* training/internal test set, an independent *Seoul-Larynx* dataset for external validation, and a prospective *H&NANCE-Larynx* cohort for real-world performance evaluation.

It is worth noting that these datasets exclusively include SCC images obtained from various clinical settings during several in-office and intraoperative procedures.

7.2.2 Architectures Comparison

We systematically evaluated three families of segmentation methodologies, two of which are based on the widely diffuse SAM foundation model and one using conventional CNN architectures. The goal was to identify the most effective approach for accurate and autonomous tumor margin delineation in laryngeal endoscopy.

(i) Prompt-based foundation modules, including SAM as well as its medical adaptations and fine-tuned variants;

(ii) Prompt-less foundation models, leveraging automatic prompt generation to guide segmentation without manual input;

(iii) Conventional CNN baselines specifically trained for laryngeal carcinoma segmentation.

In Table 7.1, a summary of the investigated segmentation models is provided.

Conventional CNN baselines

To contextualize the performance of foundation models in the specific field of laryngeal carcinoma segmentation, two CNN-based architectures were included as baseline comparators: *SegMENT-Plus* and *nnU-Net*.

As explained in Section 2.5 of Chapter 2, *SegMENT-Plus* (Sampieri et al., 2024) was specifically designed for laryngeal carcinoma segmentation and employs an encoder-decoder

structure based on EfficientNet. Among the published approaches relative to laryngology, it represents the most robust state-of-the-art CNN baseline, as it was trained and validated on relatively large and heterogeneous datasets, showing strong performance and high computational efficiency.

In addition to *SegMENT-Plus*, the segmentation performance of *nnU-Net* (Isensee et al., 2021) was analyzed. It represents a widely adopted self-configuring segmentation framework that automatically adapts its architecture, preprocessing, and training settings to the characteristics of the dataset. This methodology has become a de facto baseline in general medical image segmentation due to its ability to produce competitive results across modalities without extensive manual tuning. It was trained from scratch using laryngeal data, following the standard nnU-Net training protocol.

Prompt-based foundation modules

Recent advances in AI have led to the development of foundation models, i.e., large-scale architectures trained on broad and heterogeneous datasets, typically in a self-supervised or weakly supervised manner, to acquire general-purpose representations that can be transferred across a wide range of downstream tasks. Unlike traditional models, which are optimized for a single and narrow application, foundation models learn semantic, structural, and contextual features, enabling strong generalization across domains. Once pretrained, these models can be efficiently adapted to new tasks through lightweight fine-tuning or prompting mechanisms, making them particularly attractive for medical imaging, where data scarcity and annotation costs are significant concerns. In this context, prompt-based foundation models require external inputs, such as points, scribbles, or bounding boxes, to guide the model toward the region or object of interest.

Segment Anything Model (*SAM*) is one of the most noteworthy foundation models for segmentation among the available models. Originally introduced by Kirillov et al. (Kirillov et al., 2023), its architecture integrates a ViT-based image encoder, a prompt encoder, and a mask decoder, that jointly support flexible segmentation. The image encoder is responsible for input processing and the extraction of the image embeddings. The prompt encoder processes the input prompt, while the mask decoder takes the output of the image and prompt encoders and produces the final segmentation masks. Although SAM was trained on natural images, its strong generalization capability suggests promising potential for medical imaging tasks. In this work, *SAM* was evaluated both in zero-shot mode and after domain-specific fine-tuning, enabling the assessment of its impact on laryngeal lesion margin delineation. A schematic representation of the SAM architecture is provided in Figure 7.1.

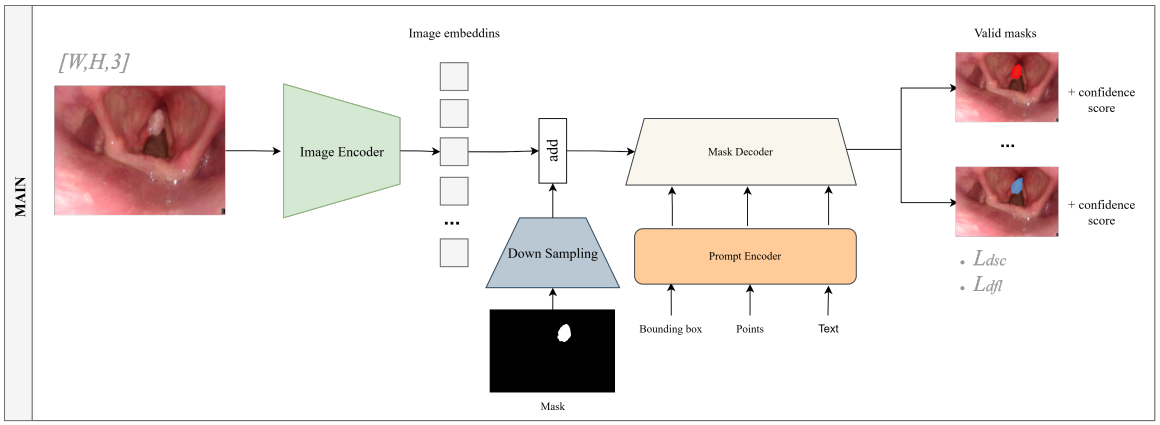


Figure 7.1 Schematic representation of the Segment Anything Model (SAM) architecture, adapted from (Kirillov et al., 2023). The model consists of three main components: an image encoder based on Vision Transformer (ViT), a prompt encoder that processes user-provided prompts (points, boxes, etc.), and a mask decoder that generates segmentation masks based on the combined information from the image and prompt encoders.

MedSAM (Ma et al., 2024) is a variant of *SAM*, where the same architecture was pretrained on a large collection of medical datasets. Unlike SAM, it incorporates knowledge suited to multimodal and biomedical imagery. *MedSAM* represents a promising intermediary between generic foundation models and domain-specific architectures, especially for medical applications where obtaining diverse and representative data is challenging. Similar to its general-purpose counterpart, *MedSAM* was evaluated both in zero-shot mode and after a soft fine-tuning on laryngeal data.

Authors in (Lin et al., 2024) introduced *SAMUS*, an adaptation of *SAM* originally developed for ultrasound imaging, and characterized by specific architectural adjustments intended to handle noise and irregular tissue boundaries. The key structural adjustments involve input size reduction from 1,024×1,024 to 256×256 pixels to lower the computational burden, the addition of a parallel CNN branch to extract global features and integrate them to the ViT output through cross-attention, and the use of adaptive modules inserted into the pre-trained model for domain adaptation. Overall, these modifications reduce memory consumption during fine-tuning of the pre-trained model and improve image feature modeling. Furthermore, the authors developed a fully autonomous extension, *AutoSAMUS*, which is detailed in the subsequent Paragraph 7.2.2. *SAMUS* model was trained starting from *SAM* weights and evaluated on the laryngeal data, both with input prompts and in its fully autonomous configuration.

Prompt-based models provide high accuracy and flexibility, but require human interaction or prior definition of input prompts (bounding box or point), which limits their autonomy in real-time clinical workflows.

Prompt-less foundation modules

Prompt-less or autonomous approaches aim to eliminate user input requirements, enabling the model to infer the target structure directly from the image and thus operate in a fully automatic manner.

As previously mentioned, authors in (Lin et al., 2024) extended *SAMUS* with a prompt-generation module. This allowed the model to infer what and where to segment without external guidance. In this fully autonomous variant, referred to as *AutoSAMUS*, the prompt generator returns latent prompts, which correspond to points, to direct the SAM-based decoder output. This approach is particularly valuable for endoscopic applications where manual prompting is impractical, as the segmentation algorithm must operate in real time alongside the incoming endoscopic video stream.

Another investigated approach was *SAMed* (Zhang and Liu, 2023). It introduces Low-Rank Adaptation (LoRA) layers into the image encoder of *SAM*, empowering efficient fine-tuning by updating only a small set of parameters. During training, *SAMed* learns how to encode the prompt from a default embedding, which remains unchanged during inference, thereby allowing prompt-free usage. Overall, *SAMed* maintains the core structure of *SAM* but enhances it with lightweight, trainable modules that adapt the model specifically for medical image segmentation tasks. In this work, it was tested on the laryngeal dataset to evaluate its performance in prompt-less tumor margin delineation.

Following the pipeline of the clinical workflow and the availability of the detection module proposed in Chapter 5, an alternative approach to face the prompt-less segmentation task is to leverage predicted bounding boxes as pseudo-prompts for guiding a SAM-based model. In the *YOLO+MedSAM* approach, the YOLO-based detection network detects the lesion and generates bounding boxes, which are then passed as pseudo-prompts to the fine-tuned version of *MedSAM*, enabling segmentation without human intervention. This strategy bridges detection and segmentation in an automated cascade, replicating the natural flow of the clinical practice, and serves as a practical intermediary between fully prompt-based and prompt-less systems.

Lastly, further tests were conducted to evaluate a hybrid approach, referred to as *SegMENT-Plus+MedSAM*, that concatenates the outputs of the domain-specific *SegMENT-Plus* model to the fine-tuned *MedSAM* model. In this two-stage framework, the SegMENT-Plus network

provides an initial coarse delineation of the lesion. After that, the coarse mask generated by *SegMENT-Plus* is converted into a pseudo-prompt for *MedSAM* by extracting a bounding box that encloses the segmented lesion. This pseudo-prompt is then passed to *MedSAM*'s prompt encoder to refine the initial segmentation. This strategy aims to combine the local feature extraction ability of CNNs with the global contextual reasoning of transformer-based foundation models.

Prompt-less models offer the advantage of full autonomy and represent an important research direction toward real-time surgical assistance systems.

7.3 Experimental Protocol

7.3.1 Data Pre-processing

The *Genova-Larynx* dataset was split into training, validation, and internal test sets using a patient-wise split to avoid data leakage. This resulted in a training set of 3,146 frames, a validation set of 393 frames, and 100 images used for internal testing. The *Seoul-Larynx* and *H&NANCE-Larynx* datasets were reserved exclusively for external validation and real-world performance assessment, respectively. All images were resized to network specific sizes, ranging from 256×256 pixels for *SegMENT-Plus* and *SAMUS*, 512×512 for *SAMed* and 1,024×1,024 for *SAM* and *MedSAM*, while *nnU-Net* automatically determined the image input size during the pre-processing phase. All data were also normalized to clip pixel intensity values in the range [0, 1]. Moreover, data augmentation techniques, including random rotations, flips, and brightness or contrast adjustments, were applied during the training of CNN baselines to enhance model generalization. Foundation models were trained without data augmentation, following the protocols established in the original works, as they already possess strong generalization capabilities due to the pretraining on large and diverse datasets.

7.3.2 Training Settings

The training process followed different approaches adapted to each model's specific requirements and the computational constraints of the experiment. Detailed training settings for each model are summarized in Table 7.1.

The *SegMENT-Plus* model, which was used as the state-of-the-art baseline for carcinoma segmentation, was trained using the configuration described in the original paper (Sampieri et al., 2024), replicated for consistency and comparison. The model utilized a batch size of

Table 7.1 Summary of training settings for the investigated segmentation models: *SegMENT-Plus, nnU-Net, SAM, MedSAM, SAMUS, AutoSAMUS, and SAMed.*

Model	**Epochs**	**Batch Size**	**Optimizer**	**Initial Learning Rate (lr_0)**	**Image Size**	**Inference time (s)**
SegMENT-Plus (Sampieri et al., 2024)	200	8	AdamW	1×10^{-4}	256×256	≈ 0.55
nnU-Net (Isensee et al., 2021)	1000	2	SGD	1×10^{-2}	1280×768	≈ 0.37
SAM (Kirillov et al., 2023)	10	8	AdamW	1×10^{-4}	1024×1024	≈ 0.20
MedSAM (Ma et al., 2024)	10	8	AdamW	1×10^{-4}	1024×1024	≈ 0.20
SAMUS (Lin et al., 2024)	400	8	AdamW	5×10^{-3}	256×256	≈ 0.03
AutoSAMUS (Lin et al., 2024)	400	8	AdamW	5×10^{-3}	256×256	≈ 0.03
SAMed (Zhang and Liu, 2023)	200	8	AdamW	5×10^{-3}	512×512	≈ 0.26

8 images resized to 256×256 pixels and an initial learning rate of 1e-4, optimized with the AdamW optimizer. The training was conducted for a maximum of 200 epochs, with early stopping based on validation loss to prevent overfitting.

The *nnU-Net* model was trained from scratch using its self-configuring framework, which automatically determines optimal training parameters based on the dataset characteristics. Given the *Genova-Larynx* dataset, the model set a batch size of 2 and a patch size of 1,280×768 pixels. The chosen optimizer was SGD, configured with a learning rate of 1e-2, a momentum of 0.99, and a weight decay of 3e-05, for a maximum of 1000 epochs. Additionally, data augmentation techniques were applied, including scaling within a range of 0.7-1.4, rotations between -15° and 15°, gamma correction with values between 0.7 and 1.5, and mirroring on both axes.

For *SAM* and *MedSAM*, the pre-trained models obtained from the official repositories were first tested without modification, i.e., "zero-shot" experiments, and subsequently fine-tuned on the *Genova-Larynx* training dataset. The fine-tuning process involved the following hyperparameters: 10 epochs, a batch size of 8, and the AdamW optimizer with an initial learning rate of 1e-4. During fine-tuning, the prompt encoder was frozen, and only the image encoder and mask decoder weights were considered as learnable components. This approach

allowed the models to adapt to the appearance of laryngeal tissues without overfitting. Input images were resized to 1,024×1,024 pixels to match the original training configuration. Bounding boxes extracted from the segmentation masks provided by clinicians were used as input prompts during both training and inference.

SAMUS was fine-tuned starting from the *SAM* weights on the *Genova-Larynx* dataset. The training was conducted for 400 epochs with a batch size of 8, using the AdamW optimizer with an initial learning rate of 5e-3. The input images were resized to 256×256 pixels, as per the architectural design. Both the image encoder and mask decoder were fine-tuned, while the prompt generator was trained from scratch. Ground-truth segmented masks were exploited by the model as input prompts during training and inference. The fully autonomous *AutoSAMUS* model was trained with the same parameters. Its weights were initialized with those of the *SAMUS* counterpart pre-trained on laryngeal data, except for the prompt-generator module, which was trained from scratch.

Lastly, for what concerns the implementation of *SAMed*, the training settings were 200 epochs, a batch size of 8, and the AdamW optimizer with an initial learning rate of 5e-3. Input images were resized to 512×512 pixels. The model utilized a low-rank (4-rank) adaptation for the image encoder, which increased computational efficiency while maintaining high performance. During training, the prompt encoder was fine-tuned using a default embedding, which remained fixed during inference.

For the ensemble model *YOLO+MedSAM*, no additional training steps were applied. Indeed, the YOLO-based detection network, previously trained as described in Chapter 5, was used to generate bounding boxes. These boxes served as pseudo-prompts for the fine-tuned *MedSAM* model, which was trained as explained above.

Likewise, for the hybrid model *SegMENT-Plus+MedSAM*, the *SegMENT-Plus* model was first trained replicating the official implementation by (Sampieri et al., 2024). The produced coarse masks were then converted into bounding box pseudo-prompts and used to guide the fine-tuned *MedSAM* model during inference.

7.3.3 Performance Metrics

Segmentation performance on the three different datasets was quantitatively assessed using the Intersection over Union (IoU) (Eq. 2.9) and the Dice Similarity Coefficient (DSC) (Eq. 2.8) metrics. These metrics are standard measures for evaluating the overlap between predicted segmentation masks and ground truth annotations. Mean and standard deviation

values were reported for each metric to summarize the overall performance across the datasets.

7.4 Results

This section presents the performance results of the segmentation models evaluated for laryngeal carcinoma margin delineation. The models tested include those listed in Table 7.1. These models were evaluated on both *internal (Genova-Larynx)* and *external datasets (Seoul-Larynx, H&NANCE-Larynx)* to assess their generalization across centers, modalities, and patients.

The results demonstrated a clear distinction in performance between prompt-based and prompt-less foundation models, with *MedSAM* outperforming all other models across the datasets (Table 7.2). Notable observations include:

- On the *Genova-Larynx* set, the fine-tuned *SAM* achieved the highest IoU of 0.84 and DSC of 0.91, outperforming other models, including the baselines *nnU-Net* and *SegMENT-Plus*. However, *MedSAM* closely followed with identical scores but a slightly higher standard deviation, indicating their strong capability in segmenting laryngeal carcinomas even with a very lightweight adaptive training.
- The *Seoul-Larynx* set revealed consistent high performance for *MedSAM*, with an IoU of 0.86 and DSC of 0.92, confirming its effectiveness across different clinical centers and data acquisition conditions.
- On the more challenging *H&NANCE-Larynx* dataset, where early and difficult-to-segment lesions were present together with limited-quality acquisitions, *MedSAM* still led with IoU = 0.82 and DSC = 0.90, demonstrating its ability to generalize across diverse lesion types, while other models like *SAMUS* and *nnUNet* showed a significant drop in performance.

When compared to conventional CNN baselines like *SegMENT-Plus* and *nnU-Net*, all foundation models clearly outperformed in terms of both IoU and DSC across all test sets, confirming their potential for easy adaptation to laryngeal carcinoma segmentation. In particular, *MedSAM* demonstrated superior generalization, making it suitable for precise tumor margin delineation in clinical settings thanks to the guidance of user-defined prompts.

While *AutoSAMUS* achieved reasonable results, it did not reach the effectiveness of *MedSAM* or *SAM*, especially on challenging data from the *H&NANCE-Laryx dataset*. This

indicates that, while prompt-less models offer autonomy, they still lack segmentation accuracy compared to prompt-based models.

Qualitative results illustrating the segmentation outputs of the different models on representative cases from the *H&NANCE-Larynx* dataset are presented in Figure 7.2.

From the illustrated samples, it is evident that *MedSAM* achieved the most accurate delineation of tumor margins, consistently predicting contours that closely match the ground-truth annotations, even in complex cases (row 3). In contrast, discrepancies were observed with *SegMENT-Plus* for lesions in areas affected by specular reflections or reduced illumination. *AutoSAMUS*, although capable of producing reasonable outputs and correctly avoiding the specular reflection artifact in the first row, still struggled to capture the small, poorly illuminated lesion in row 3 due to the absence of prompt-based attention guidance. In the last row, both *SegMENT-Plus* and *AutoSAMUS* failed to accurately delineate the lesion, leading to substantial over-segmentation relative to the annotated ground truth, whereas *MedSAM* maintained a high level of precision thanks to the prompt.

7.5 Discussion

The experimental results presented in this chapter highlight the significant advancements that foundation models, particularly prompt-based ones like *MedSAM*, can bring to the task of laryngeal carcinoma margin delineation. The superior performance of *MedSAM* across all test sets, including challenging external data, underscores its robustness and generalization capabilities. This is particularly important in medical imaging, where variability in acquisition protocols and patient demographics can significantly impact model performance. In contrast, AutoSAM shows competitive inference efficiency, supporting practical real-time applications, which is further facilitated by the reduced input resolution required by the model. However, the significant performance gap observed between prompt-based models and prompt-less models such as *AutoSAMUS* and *SAMed* emphasizes the benefits of controlling model attention toward relevant anatomical structures or RoI. Indeed, although fully autonomous models achieved promising results, prompt-based models still offer higher accuracy and stability in such a complex task.

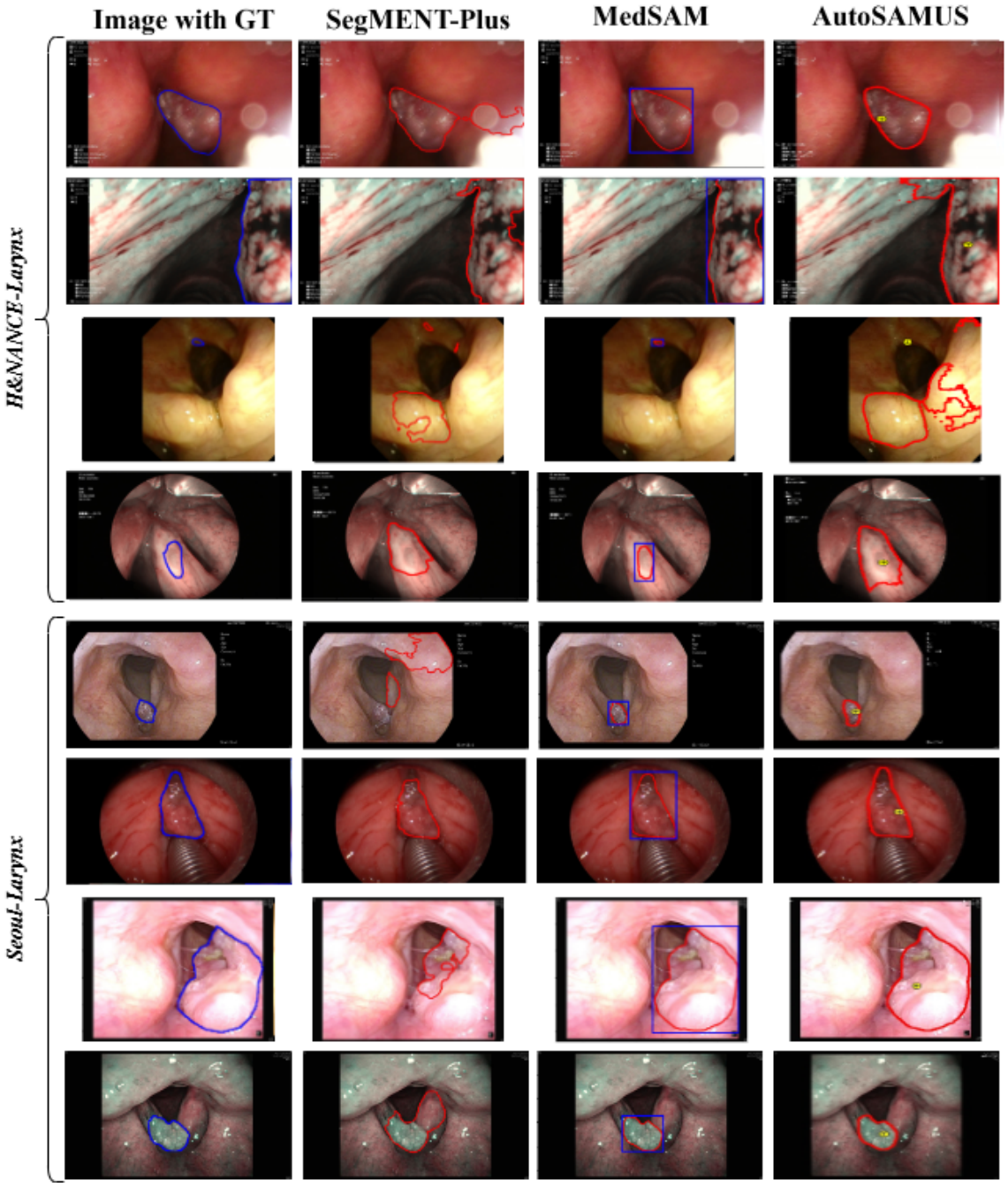


Figure 7.2 Qualitative segmentation results on representative cases from the *H&NANCE-Larynx* and *Seoul-Larynx* test sets. For each category of conventional, prompt-based, and prompt-less foundation models, segmentation outputs of the best-performing tested methodologies are shown. From left to right: original image with ground truth mask, *SegMENT-Plus* from (Sampieri et al., 2024), *MedSAM* (fine-tuned), and *AutoSAMUS*.

Table 7.2 Segmentation performance across all test sets: *internal (Genova-Larynx), external (Seoul-Larynx, H&NANCE-Larynx)*.

Model	*Genova-Larynx*	*Seoul-Larynx*	*H&NANCE-Larynx*
Conventional CNN baselines			
SegMENT-Plus	**IoU**: **0.68** ± **0.18**, **DSC**: **0.79** ± **0.16**	**IoU**: 0.51 ± 0.31, **DSC**: 0.61 ± 0.32	**IoU**: **0.56** ± **0.31**, **DSC**: **0.66** ± **0.31**
nnU-Net	**IoU**: 0.64 ± 0.23, **DSC**: 0.75 ± 0.22	**IoU**: **0.63** ± **0.23**, **DSC**: **0.75** ± **0.21**	**IoU**: 0.54 ± 0.31, **DSC**: 0.64 ± 0.30
Prompt-based foundation modules			
$SAM_{zero\text{-}shot}$	**IoU**: 0.74 ± 0.16, **DSC**: 0.84 ± 0.12	**IoU**: 0.78 ± 0.14, **DSC**: 0.87 ± 0.11	**IoU**: 0.83 ± 0.22, **DSC**: 0.88 ± 0.18
$MedSAM_{zero\text{-}shot}$	**IoU**: 0.79 ± 0.12, **DSC**: 0.88 ± 0.09	**IoU**: 0.80 ± 0.13, **DSC**: 0.88 ± 0.09	**IoU**: 0.81 ± 0.10, **DSC**: 0.89 ± 0.06
SAM	**IoU**: **0.84** ± **0.07**, **DSC**: **0.91** ± **0.04**	**IoU**: 0.83 ± 0.09, **DSC**: 0.90 ± 0.06	**IoU**: 0.81 ± 0.11, **DSC**: 0.89 ± 0.08
MedSAM	**IoU**: 0.84 ± 0.08, **DSC**: 0.91 ± 0.05	**IoU**: **0.86** ± **0.08**, **DSC**: **0.92** ± **0.05**	**IoU**: **0.82** ± **0.10**, **DSC**: **0.90** ± **0.06**
SAMUS	**IoU**: 0.74 ± 0.15, **DSC**: 0.84 ± 0.11	**IoU**: 0.69 ± 0.20, **DSC**: 0.80 ± 0.17	**IoU**: 0.63 ± 0.25, **DSC**: 0.74 ± 0.23
Prompt-less foundation modules			
AutoSAMUS	**IoU**: **0.73** ± **0.16**, **DSC**: **0.83** ± **0.12**	**IoU**: **0.63** ± **0.25**, **DSC**: **0.74** ± **0.24**	**IoU**: **0.60** ± **0.28**, **DSC**: **0.70** ± **0.28**
SAMed	**IoU**: 0.65 ± 0.22, **DSC**: 0.76 ± 0.20	**IoU**: 0.50 ± 0.30, **DSC**: 0.60 ± 0.32	**IoU**: 0.52 ± 0.30, **DSC**: 0.63 ± 0.30
YOLO + MedSAM	**IoU**: 0.50 ± 0.31, **DSC**: 0.61 ± 0.31	**IoU**: 0.53 ± 0.38, **DSC**: 0.60 ± 0.39	**IoU**: 0.50 ± 0.34, **DSC**: 0.59 ± 0.34
SegMENT-Plus + MedSAM	**IoU**: 0.68 ± 0.19, **DSC**: 0.79 ± 0.16	**IoU**: 0.52 ± 0.32, **DSC**: 0.62 ± 0.33	**IoU**: 0.57 ± 0.31, **DSC**: 0.67 ± 0.32

Chapter 8

Synthetic Data Generation

8.1 Overview

In the previous chapters of this thesis, results indicated that DL models can effectively support key steps of the laryngeal endoscopic workflow, including informative frame selection, lesion detection, and risk assessment. Despite the observed promising results, the routine use of computer-assisted systems in laryngology remains limited. The lack of large, heterogeneous, and consistently annotated datasets hampers the real-world deployment of such tools. Actually, laryngology is still characterized by fragmented data resources and task-specific datasets, which are often mono-centric, focused on narrow clinical objectives, and not representative of real-world variability. As a consequence, AI models trained on such data frequently struggle to generalize to unseen context, institutions, acquisition settings, and underrepresented lesion types (Sampieri et al., 2023).

Synthetic data generation has recently emerged as a promising strategy to mitigate data scarcity. By creating artificial yet realistic samples, generative models can augment existing datasets, enhance model generalization. In this chapter, a clinically-guided synthetic data generation framework based on Latent Diffusion Models (LDMs) and ControlNet is proposed. The method is designed to generate realistic laryngeal endoscopic images together with their annotations, under the guidance of lesion location and clinically meaningful textual descriptions.

The chapter is organized as follows: Section 8.2 describes the datasets, the proposed generative framework, validation through conventional metrics and downstream tasks, and the human-observer realism study. Section 8.4 presents qualitative and quantitative results on synthetic data generation, downstream tasks performance, and human-observer assessment. Finally, Section 8.5 discusses the findings and limitations.

The work described in this chapter has been published in the following peer-reviewed journal article:

- Baldini, C., Kushibar, K., Osuala, R., Balocco, S., Diaz, O., Lekadir, K., & Mattos, L. S. (2025, September). Clinically-Guided Data Synthesis for Laryngeal Lesion Detection. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 54-63). Cham: Springer Nature Switzerland.

The code used for the LDM + controlNet implementation was also released, and it is now available online at https://github.com/ChiaraBaldini/endoLDMC.git.

8.2 Materials and Methods

8.2.1 Datasets

To train the generative model, the good-quality and well-annotated in-office *Athens-Larynx* dataset of 929 WL and NBI images collected at the Otolaryngology–Head and Neck Unit of the Saint Savvas Oncology Hospital (Greece) was used. This dataset is referred to as the *internal generation set*, as it was used to train and validate the proposed generative model. It should be noted that such data, described in detail in Chapter 3, were originally collected for lesion detection and classification tasks. Hence, they included bounding box annotations with lesion histological type labels.

Furthermore, to evaluate the utility of synthetic data for downstream tasks, two additional external sets of data were employed. For the first downstream task of lesion detection, the detection model was trained on the *internal generation set*, and its performance was then assessed externally on 88 randomly selected data from the *Genova-Larynx* collection (*external detection test set*), maintaining the same size of the internal test set. On the contrary, the histological classification task was assessed by training the classifier on the dataset described in Subsection 6.2.1 of Chapter 6, after restricting the data to samples acquired during in-office laryngeal examinations. This filtering produced a subset of 6,475 data, referred to as the *In-office dataset*. The model was then tested on the *Barcelona-Larynx* dataset introduced in Chapter 6, consisting of 150 images from as many different patients acquired at the Hospital Clinic of Barcelona (Spain) during routine in-office examinations. Healthy images were excluded as not belonging to the categories considered in the classification task. The resulting set of 137 images is referred to as the (*external classification test set*).

8.2.2 Clinically-guided Latent Diffusion Framework

An overview of the proposed generative framework is provided in Fig. 8.1. The method is based on a Latent Diffusion Model (LDM) conditioned via ControlNet on two complementary sources of clinical information: (i) a textual caption describing the optical modality and lesion type, and (ii) a mask encoding the location and shape of the lesion bounding box.

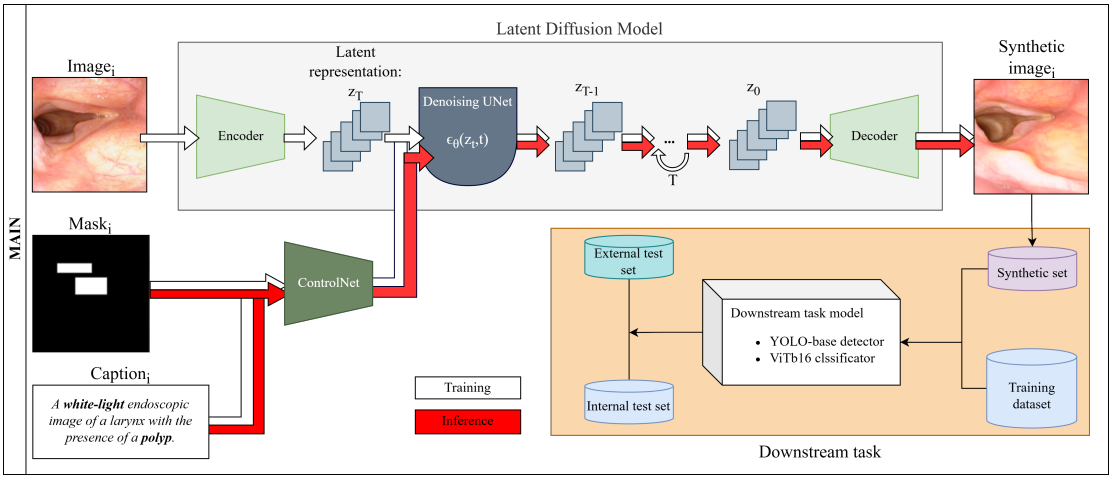


Figure 8.1 Overview of the proposed clinically-guided generative framework. During training (white arrows), captions containing optical modality and lesion type, together with bounding box masks, are used to condition a Latent Diffusion Model (LDM) via ControlNet. During inference (red arrows), random noise, captions, and masks are combined to generate synthetic laryngeal endoscopic images with a realistic appearance and controllable lesion location.

Latent Diffusion Model

Diffusion Models (DMs) learn to approximate the data distribution by progressively adding Gaussian noise to the input data and then learning a reverse denoising process. Given an observed sample x_0 and Gaussian noise $\varepsilon \sim \mathcal{N}(0,1)$, a forward diffusion process constructs a sequence of increasingly noisy samples x_t, while a neural network ε_θ is trained to predict the noise at each step. The DM loss can be written as:

$$L_{\text{DM}}(\theta) = \mathbb{E}_{x_0,\varepsilon\sim\mathcal{N}(0,1),t\sim U[1,T]}\left[\|\varepsilon - \varepsilon_\theta(x_t,t)\|_2^2\right]. \tag{8.1}$$

Latent Diffusion Models (LDMs) (Rombach et al., 2022) perform this process in the latent space of an autoencoder instead of pixel space, reducing computational cost and memory usage. An image x_0 is first encoded into a latent representation z_0; noise is iteratively added

to obtain z_t, and a UNet-based denoiser estimates the noise to recover z_0. A decoder then projects the denoised latent back to image space. The LDM loss is:

$$L_{\text{LDM}}(\theta) = \mathbb{E}_{z_0,\varepsilon\sim\mathcal{N}(0,1),t\sim U[1,T]}\left[\|\varepsilon - \varepsilon_\theta(z_t,t)\|_2^2\right]. \tag{8.2}$$

In this work, the pre-trained autoencoder and CLIP-based text encoder components from a publicly available LDM (Rombach et al., 2022) were employed, and the LDM was fine-tuned on laryngeal data using a transfer learning strategy.

ControlNet Conditioning

ControlNet (Zhang et al., 2023b) extends diffusion models by introducing an additional network branch that processes conditioning inputs (e.g., edge maps, segmentation masks) and injects their features into the denoising network. This allows the generative process to be guided by structural or semantic constraints while preserving the flexibility and diversity of the base LDM.

In the medical imaging context, precise control over lesion location and appearance is critical. Therefore, the methodology proposed here conditioned the LDM with two clinically motivated inputs:

- Textual caption: a short sentence describing the optical modality (WL or NBI) and the lesion class, e.g., "A white-light endoscopic image of a larynx with the presence of a polyp."
- Bounding box mask: a 3-channel binary mask derived from expert-annotated bounding boxes, encoding the lesion location in the image space.

During training, the ControlNet branch was fed with the bounding box mask and the text embedding, and its outputs were fused with the LDM features to jointly learn the relationship between anatomical structures, lesion type, and spatial extent. During inference, the model received random latent noise, together with a caption and a desired bounding box mask, and generated a new synthetic image that respects both the clinical context and the spatial constraints.

8.3 Experimental Protocol

8.3.1 Data preparation

For training the generative model, the *internal generation set* was split into training, validation, and test subsets of 727, 111, and 91 images, respectively. All images were resized to 640×640 pixels as this was the input resolution set for the pre-trained LDM. Lesion bounding box coordinates were converted from the YOLO format into 3-channel binary masks, and textual captions were automatically defined based on lesion type annotation and optical modality available information in the dataset.

To assess the generalization capability of the model and to generate a larger pool of synthetic samples, 727 additional bounding box masks were created by applying random rotations and scale variations to the original masks of the *internal generation set*. These augmented masks, together with the corresponding captions, were used to generate diversified synthetic images and analyze their quality.

8.3.2 Training settings

The pre-trained LDM backbone was first fine-tuned for 100 epochs with a batch size of 4, an initial learning rate of 5e-5, and an image resolution of 640×640 pixels. Subsequently, the LDM weights were frozen, and a ControlNet adapter was integrated and fine-tuned with the same hyperparameters, enabling controlled generation based on masks and captions. All experiments were conducted on an NVIDIA A100 GPU (80 GB of memory).

8.3.3 Evaluation Metrics

The objective of the proposed framework is not only to generate images that resemble real data but also to produce clinically relevant diversity that can improve the performance of downstream tasks. For this reason, a combination of distributional and task-driven experimental metrics was considered. The Fréchet Inception Distance (FID) (Heusel et al., 2017) is a commonly used metric to evaluate the similarity between real and synthetic image distributions. However, this does not fully align with this specific task as it might undervalue the contribution of clinically diverse synthetic data. To better account for intrinsic data variability, we compute the $\text{FID}_{\text{ratio}}$ (Osuala et al., 2023), defined as:

$$FID_{\text{ratio}} = 1 - \frac{FID_{rs} - FID_{rr}}{FID_{rs}},$$

where FID_{rs} is the FID between real and synthetic sets, and FID_{rr} is the FID computed between two disjoint real subsets.

We also compute the Inception Score (IS) (Salimans et al., 2016) for both real and synthetic images to quantify the variability and distinctiveness of generated samples.

To evaluate the utility of synthetic images for CAD systems development, two downstream tasks were further implemented: general lesion detection and histological classification. For both tasks, the impact of augmenting the real training set with synthetic samples on model performance was analyzed as described in the following paragraphs.

Downstream Lesion Detection Task

The YOLOv8n network was employed as lesion detector, following the implementation previously introduced in Chapter 5 for lesion detection in laryngeal endoscopy. The detector was first trained exclusively on real images from the *internal generation set* training dataset. Then, it was progressively augmented with synthetic samples generated by the proposed LDM + ControlNet framework, adding 5%, 10%, 20%, 40%, and 80% synthetic images with respect to the number of real samples. Detection performance was evaluated on the test samples of the *internal generation set* and on the *external detection test set* dataset (Subsection 8.2.1). After noticing that the addition of only 10% synthetic images improved performance optimally, as explained in the following results section (Section 8.4.2), a 3-fold cross-validation experiment was carried out using different random subsets of synthetic images corresponding to 10% of the real training size and the same real data. This permitted the analysis of the effect of specific choices of synthetic samples. Furthermore, to better exploit the synthetic pool, an Uncertainty Estimation (UE) strategy was introduced: the three YOLO models trained in the cross-validation stage were used to obtain detection predictions on all synthetic images generated, and the variance of the predicted confidence scores was computed for each image. Synthetic samples were then sorted in descending order of variance, and the top 10% were selected as "challenging" instances for detection training. This UE-selected subset was used to retrain the detector and compared the results with those obtained with the randomly selected synthetic subsets. Unless otherwise specified, the YOLOv8 detector was trained for 100 epochs using AdamW, an initial learning rate of 1e-3, early stopping with a patience of 50 epochs, a batch size of 32, and the default data augmentation settings. Its performance was evaluated using Precision (Eq. 2.1), Recall (Eq. 2.2), and mean Average Precision at IoU 0.5 ($mAP_{@IoU=0.5}$) (Eq. 2.6).

Downstream Histological Classification Task

For the histological classification task, the ViTb16 architecture was trained to classify lesions into seven histological categories, following the approach described in Chapter 6. The classifier was first trained on real images from the *In-office set*, divided into 4,888 images for training, 777 for validation, and 810 for internal testing. Then, the training set was augmented with approximately 20% (1,466) samples synthesized by the proposed generative framework. During synthetic image selection, data generated by using a caption containing the "SCC" word were excluded in order to augment only the underrepresented lesions. Moreover, as the model was structured to accept lesion RoI inputs rather than the full images, synthetic data were cropped around the lesion based on the bounding box masks used to generate them. The same training hyperparameters described in Chapter 6 were used. Classification performance was evaluated via Precision (Eq. 2.1), Recall (Eq. 2.2), F1-score (Eq. 2.3), and Accuracy (Eq. 2.7) on test data of the *In-office set* and on the *external classification test set* (Subsection 8.2.1).

Expert-Observer Realism Study

To assess the perceptual realism of the generated images from a clinical perspective, we conducted an expert-observer study involving five otorhinolaryngologists with 6-11 years of experience. A set of 20 images was created, consisting of 10 real (*internal generation set*) and 10 synthetic images. Each clinician was asked to assess each image using a six-point Likert scale in response to the statement "*This is a REAL image*": [Strongly disagree, Disagree, Slightly disagree, Slightly agree, Agree, Strongly agree]. Following (Alyafi et al., 2020; Garrucho et al., 2023), the Likert responses were mapped to probabilities [0.05, 0.23, 0.41, 0.50, 0.77, 0.95], which were then used to compute an ROC curve and the corresponding AUC for each participant. Accuracy was computed by thresholding the probability at 0.5.

8.4 Results

8.4.1 Qualitative Generation Results

Figure 8.2 shows representative examples of synthetic laryngeal images generated by the proposed framework. The model was able to reproduce diverse anatomical configurations of the vocal folds, different lesion types (e.g., squamous cell carcinoma, cysts, leukoplakia), and both WL and NBI appearances. For a fixed bounding box mask, varying the textual caption

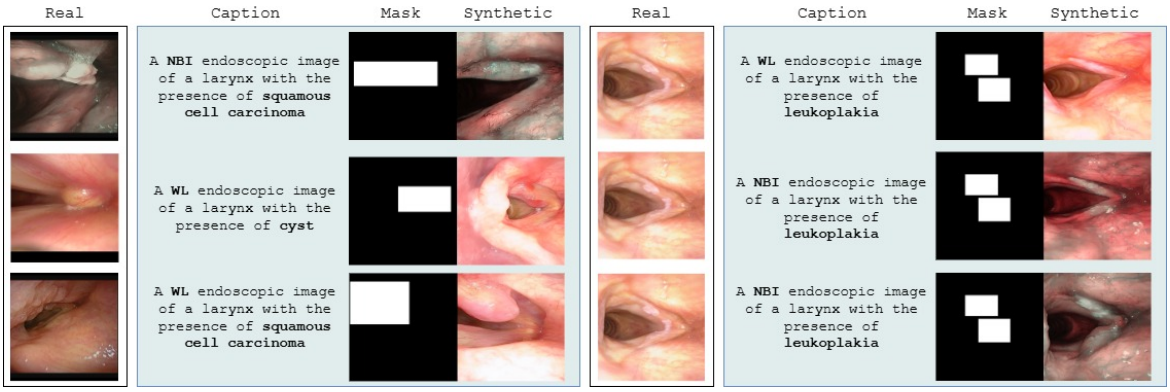


Figure 8.2 Qualitative examples of synthetic data generation. Starting from real images (first column), the corresponding captions (second column) and bounding box masks (third column) were used to generate synthetic laryngeal images (fourth column). The model reproduced different lesion types (e.g., squamous cell carcinoma, cyst, leukoplakia) and optical modalities (WL and NBI), while preserving anatomical plausibility.

led to distinct synthetic images (e.g., three cases on the right), highlighting the capacity of the model to deal with spatial and semantic conditioning.

When trained with the *internal generation set*, the $\text{FID}_{\text{ratio}}$ between real and synthetic distributions was equal to 0.836, indicating that the divergence between real and synthetic data is comparable to the natural variability observed within real data splits. The Inception Score was 3.840 ± 0.270 for real samples and 4.240 ± 0.270 for synthetic images, confirming that the generated data exhibit a high degree of diversity while remaining consistent with clinically meaningful patterns of real data.

8.4.2 Impact of Synthetic Data on Lesion Detection

Figure 8.3a shows the evolution of $\text{mAP}_{@\text{IoU=0.5}}$ for the downstream detection task on both *internal generation* and *external detection test sets* when increasing the proportion of synthetic images in the training data. On the *internal generation test set*, the best performance ($\text{mAP}_{@\text{IoU=0.5}} = 0.89$) was obtained when augmenting the real dataset with 10% synthetic samples; for higher proportions, the performance saturated or slightly decreased, suggesting that excessive synthetic data may alter the balance of the training distribution. A similar trend was observed with external data, where 10% synthetic data also resulted in the highest $\text{mAP}_{@\text{IoU=0.5}}$.

Table 8.1 reports detailed results for the baseline (real-only), and the inclusion of both the three random 10% synthetic sample sets, and the UE-selected subset. On average, adding 10% synthetic images improved $\text{mAP}_{@\text{IoU=0.5}}$ by +8% on the *internal generation test set* and

+16% on the *external detection test set* with respect to the real-only baseline. When using the UE-based selection strategy, the improvement reached +9% and +22%, respectively. These findings indicate that expanding the training set with "challenging" synthetic samples can significantly enhance the ability of the detector to handle hard cases and to generalize to unseen clinical data.

8.4.3 Impact of Synthetic Data on Histological Classification

Synthetic data augmentation led to a measurable performance improvement also for the histological classification task. When internally tested on 810 data from the *In-office* test set, the overall weighted F1-score increased from 73% to 76%. A similar trend was observed on the *external classification test set*: +6% in F1-score. Class-wise analysis revealed that performance increased across all lesion categories, with higher gains for underrepresented (*papilloma*) or challenging (*leucoplakia*) conditions, except for cysts and polyps.

8.4.4 Expert-Observer Realism Study

The results of the expert-observer study are summarized in Fig. 8.3c. Each point represents one clinician's judgment on one image, with the y-axis encoding the confidence level and the marker color indicating whether the image was correctly (green) classified as real or synthetic. Overall, clinicians frequently misclassified synthetic images as real, especially when their confidence was moderate to high. Across all five users, the real-versus-synthetic classification task yielded a mean AUC of 0.51 ± 0.09, and the mean Accuracy in correctly predicting them of $67 \pm 13\%$ (Table 8.3). These values support the empirical observation that the generated images are highly realistic and difficult to distinguish from real endoscopic images, even for experienced otorhinolaryngologists.

8.5 Discussion

In this chapter, a clinically-guided synthetic data generation framework for laryngeal endoscopy based on LDMs and ControlNet was introduced. This is the first work exploring realistic image-annotation synthesis for laryngeal lesions. By conditioning the generative process on textual descriptions and lesion bounding box masks, the method can produce anatomically plausible, clinically diverse images that cover different lesion types and optical modalities.

The distributional evaluation (FID_{ratio} and IS) indicated that the synthetic data were diverse and consistent with the real data distribution. More importantly, the downstream experiments demonstrated that adding a relatively small fraction (10-20%) of synthetic images to the real training set substantially improved lesion detection and classification performances, particularly on external data from a different clinical center. The proposed UE selection strategy further enhanced these gains by focusing on the most informative synthetic samples, while classification improvements were most notable for classes that originally suffered from data scarcity (e.g., *papilloma)*.

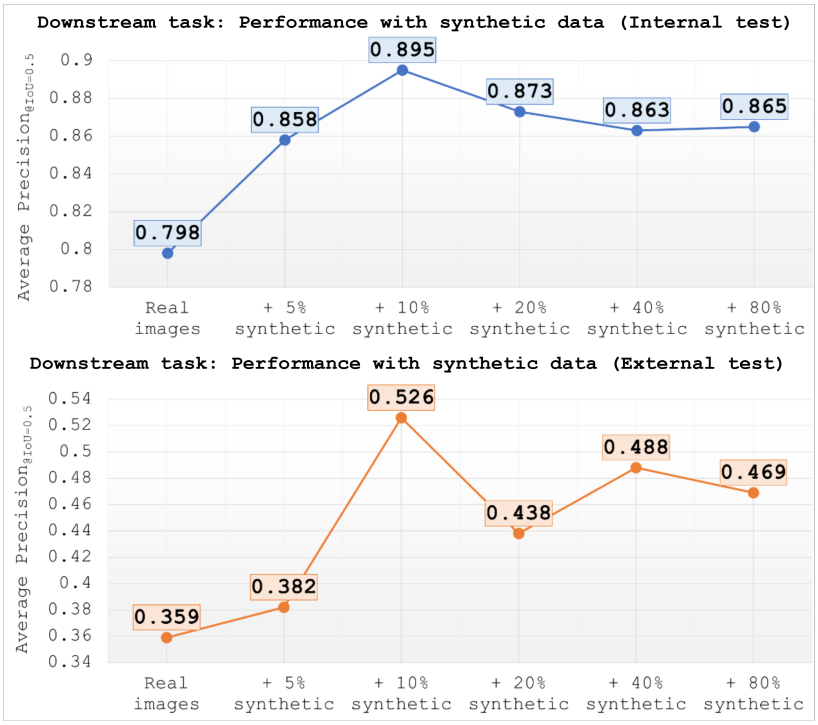
The human-observer study confirmed the high realism of the generated images, as experienced clinicians struggled to reliably distinguish synthetic from real images. This outcome supports the use of such data not only for developing AI models but also for education, simulation, and decision-support research in laryngology.

Likewise, the only closely related study by (Khazrak et al., 2025) adopted a diffusion-based approach to generate synthetic stroboscopic laryngeal images, primarily targeting visual realism and class balancing for classification purposes. However, the evaluation was restricted to a single acquisition setting and did not incorporate explicit spatial conditioning. In contrast, the clinically guided LDM proposed in this thesis enables controlled generation of synthetic laryngeal endoscopic images conditioned on lesion type, imaging modality, and spatial localization.

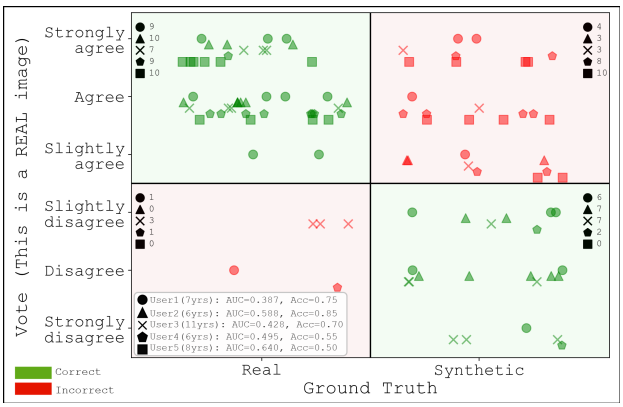
Nevertheless, some limitations remain. The current generative framework is trained on a limited number of centers and lesion types, and the conditioning is restricted to bounding boxes and coarse textual descriptions. Future work will focus on enriching the conditioning signals (e.g., pixel-wise masks, multi-label descriptions), extending the method to additional lesion categories and acquisition settings, and conducting privacy-preserving analysis.

Table 8.1 Impact of adding 10% synthetic images to the real training dataset on YOLOv8-based lesion detection. Results are reported for the *internal generation* and *external detection test* sets.

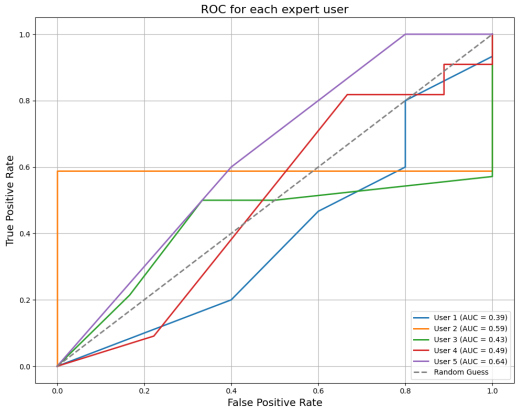
Train dataset	*Internal test*			*External test*		
	Precision	**Recall**	**mAP$_{@IoU=0.5}$**	**Precision**	**Recall**	**mAP$_{@IoU=0.5}$**
Real only	0.761	0.763	0.798	0.498	0.376	0.359
Real + 10% synthetic (fold 1)	0.848	0.851	0.895	0.614	0.526	0.526
Real + 10% synthetic (fold 2)	0.828	0.889	0.900	0.660	0.463	0.518
Real + 10% synthetic (fold 3)	0.809	0.789	0.846	0.631	0.468	0.515
Mean$_{\pm\ std}$ (random 10%)	$0.828_{\pm 0.016}$	$\mathbf{0.843}_{\pm 0.041}$	$0.880_{\pm 0.024}$	$0.635_{\pm 0.019}$	$\mathbf{0.485}_{\pm 0.028}$	$0.519_{\pm 0.004}$
Real + 10% synthetic (UE-selected)	**0.905**	0.833	**0.888**	**0.800**	0.464	**0.580**



(a)



(b)



(c)

Figure 8.3 (a) Evolution of $mAP_{@IoU=0.5}$ on the *internal generation* and *external detection test* sets as a function of the percentage of synthetic images added to the real training set (random selection, fold 1). Performance peaks at +10% synthetic data and then reaches a plateau. (b) Human-observer study: responses from otorhinolaryngologists assessing realism. The marker location denotes whether the image was correctly (green) classified as real or synthetic, while (c) illustrates the ROC curves achieved by individual expert users.

Table 8.2 Histological classification tests (*In-office dataset* and external data from *Barcelona-Larynx*): comparison between the classifier trained only on real data and the classifier trained on real data plus approximately 20% synthetic samples of minority classes.

Class	*Real only*			*Real + Synthetic*		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Genova-Larynx in-office test set						
0 = *Cyst*	0.73	0.49	0.59	**0.80**	**0.53**	**0.64**
1 = *Granuloma*	**0.29**	**0.07**	**0.11**	0.20	**0.07**	**0.11**
2 = *Leukoplakia*	0.36	0.49	0.41	**0.45**	**0.80**	**0.58**
3 = *Papilloma*	0.52	**0.63**	0.57	**0.67**	0.57	**0.61**
4 = *Polyp*	0.83	0.75	0.79	**0.84**	**0.80**	**0.82**
5 = *Reinke's edema*	**0.75**	0.85	**0.80**	0.73	**0.86**	0.79
6 = *SCC*	0.76	0.84	0.80	**0.83**	**0.85**	**0.84**
Accuracy	73%			**77%**		
Weighted Avg	0.73	0.73	0.73	**0.76**	**0.77**	**0.76**
External test classification						
0 = *Cyst*	1.00	0.33	0.50	1.00	0.33	0.50
1 = *Granuloma*	**0.92**	0.71	0.80	0.87	**0.76**	**0.81**
2 = *Leukoplakia*	**0.58**	0.50	0.54	0.56	**0.71**	**0.62**
3 = *Papilloma*	**1.00**	0.06	0.11	0.75	**0.18**	**0.29**
4 = *Polyp*	**0.55**	**1.00**	**0.71**	0.54	0.95	0.69
5 = *Reinke's edema*	0.69	0.69	0.69	**1.00**	**0.77**	**0.87**
6 = *SCC*	0.62	**0.85**	0.72	**0.71**	**0.85**	**0.77**
Accuracy	65%			**69%**		
Weighted Avg	0.74	0.68	0.61	**0.75**	**0.69**	**0.67**

Table 8.3 Quantitative analysis of the expert-observer realism study results across the five participants.

User ID	**Years of experience**	**AUC**	**Accuracy**
User 1	7	0.39	75%
User 2	6	0.59	85%
User 3	11	0.43	70%
User 4	6	0.49	55%
User 5	8	0.64	50%
Mean $\pm$ std	-	0.51 ± 0.09	$67 \pm 13\%$

Chapter 9

Preliminary Clinical Validation

9.1 Overview

AI modules developed in Chapters 4, 5, 6, and 7 have shown promising quantitative and qualitative results on retrospective datasets. However, the effective translation of these tools into clinical practice requires their evaluation on prospective clinical data, under real-world acquisition conditions, and in collaboration with otolaryngologists.

The goal of this chapter is to describe the ongoing clinical study conducted at the Unit of Otolaryngology and Head and Neck Surgery of IRCCS San Martino Hospital in Genova (Italy), to assess the feasibility, usability, and preliminary diagnostic performance of the proposed AI-assisted laryngoscopy prototype. This evaluation focuses on the system's ability to support clinicians during in-office examinations, early diagnosis and surgical planning.

The chapter is organized as follows: Section 9.2 describes the design of the AI-assisted laryngoscopy prototype, including hardware and software components. Section 9.3 outlines the clinical study protocol, including patient selection criteria, data collection procedures, and ethical considerations. Section 9.4 details the data analysis methods used to evaluate the system's performance. Finally, Section 9.5 presents preliminary results from the clinical study, discussing the implications for future research and clinical practice.

This chapter's work has been published in the conference proceedings article that follows:

- Baldini, C., Azam, M. A., Thorniley, M., Sampieri, C., Ioppi, A., Peretti, G., & Mattos, L. S. (2024, October). AI-Assisted Laryngeal Examination System. In MICCAI Workshop on Cancer Prevention through Early Detection (pp. 133-143). Cham: Springer Nature Switzerland.

9.2 Prototyping and GUI Development

The design of the AI-assisted prototype enabled the practical integration of all AI modules described in the previous chapters into a fully functional physical system for online endoscopic assistance. The prototyping phase involved iterative refinement, real-time optimization, and constant collaboration with clinical partners.

From a hardware perspective, the system relies on a Dell Precision 7820 workstation equipped with a standard monitor and an NVIDIA RTX A6000 GPU (48 GB of memory), selected to support real-time inference of endoscopic frames across multiple DL models operating in parallel. It was mounted on a medical cart (Pro-cart 35E[1]) and complemented with additional accessories to ensure compatibility with standard endoscopic systems. The Graphical User Interface (GUI) was developed in PyQt5, with particular emphasis on responsiveness and an intuitive layout. The underlying software architecture was designed to support the modular integration of the various AI-enhanced tasks, allowing for data flow from video acquisition to real-time analysis and visualization. Significant effort was dedicated to the conversion of AI models into the ONNX format for deployment, enabling lightweight and efficient inference. Furthermore, the system was configured to function with both offline video recordings and live endoscopic data streams. Although the prototype was designed to operate without specific requirements regarding the endoscopic equipment, during this initial clinical validation phase, the online input was acquired using an Olympus ENF-VH endoscope paired with a CV-170 video processor[2] and transferred to the workstation via an SDI video cable. Figure 9.1 exemplifies the clinical integration of the prototype: the upper monitor displays the live endoscopic feed, while real-time lesion detection results are shown on a secondary monitor positioned outside the clinician's direct field of view. A foot-pedal mechanism allows the operator to save frames considered relevant for diagnosis or treatment planning.

In collaboration with clinicians, several functionalities were implemented to support the full endoscopic examination pipeline:

- Optional real-time display of AI outputs, including informativeness predictions as green or red frames, bounding boxes of the detected lesions, risk stratification scores, and tumor margin overlays;
- A foot-pedal mechanism to allow hands-free saving of informative frames;

[1] iTD GmbH, Germany.
[2] Olympus Medical System Corporation, Tokyo, Japan

Figure 9.1 Illustration of the AI-assisted laryngoscopy prototype integrated into the clinical workflow. The interface displays a live endoscopic video feed on the top monitor and real-time AI outputs (bounding boxes of the detected lesions) on a secondary monitor not directly visible to the main operator. There is also a foot-pedal mechanism used by the operator for saving relevant frames.

- An additional annotation tool to manually draw labels on previously saved frames, supported by a 27-inch ELO 2703LM LCD touchscreen[3] and stylus;
- An annotation validation system based on inter-clinician agreement, designed to reduce annotation variability and improve consistency across users.
- An automated report generation that stores AI predictions together with patient metadata and procedural details.

Overall, the proposed prototype, hereafter denoted as VERA, provides a modular assistance platform capable of subsequently processing laryngeal endoscopy data for informative frame selection (See Chapter 4), lesion detection (Section 5.3.3 of Chapter 5), *Low-risk* versus *High-risk* lesion classification (Section 6.2.2 of Chapter 6) and tumor margin delineation (*SegMENT-Plus* implementation in Chapter 7), supporting continuous software updates, and meeting the evolving needs of clinicians.

9.3 Clinical Study Protocol

The clinical study is designed as a prospective observational study aimed at assessing the real-time performance of the AI modules, their integration into the clinical workflow, and their impact on diagnostic accuracy. The study protocol was developed in collaboration with otolaryngologists and approved by the local Ethics Committee[4] to ensure compliance with ethical standards and patient safety. Patients undergoing flexible laryngoscopy during routine outpatient examinations were enrolled in the study after providing informed consent. Inclusion criteria encompassed adults (age > 18 years) with suspected laryngeal lesions, while exclusion criteria included patients with prior laryngeal surgery.

Data collection involved recording laryngoscopic videos using the VERA prototype, with clinicians performing standard examinations while the system operated in the background, avoiding bias. Informative frames were saved both manually (via foot-pedal) and automatically (based on the AI frame selection module). Saved frames were then reviewed and annotated for the presence of lesions by clinicians, with bounding boxes drawn around detected abnormalities, and the histological type of the detected lesion, including also the possibility to label a lesion as *Uncertain* and *Healthy*. Cases annotated as *Uncertain* were considered as *High-risk* cases during the evaluations in order to avoid overlooking them.

[3] Elo Touch Solutions, Inc., United States.
[4] CET Liguria: 508/2023

Additionally, for malignant lesions (*SCC*), the tumor margins were manually delineated. Such annotations by experienced otolaryngologists were used as a reference for AI-performance evaluation.

9.4 Data Analysis

For the clinical study analysis, the performance of VERA was evaluated at the patient level, aggregating predictions across all frames recorded during each examination. This approach reflects real clinical decision-making, where the final diagnosis is based on the overall video endoscopic assessment rather than individual frames. Two core components were assessed: lesion localization performance (annotated and predicted bounding boxes overlap) and risk classification accuracy (*Low-risk* or *High-risk* labels versus predictions).

9.4.1 Lesion Localization Performance

For each patient, all detected bounding boxes produced by the AI detection module were matched with expert-annotated bounding boxes in corresponding frames, and only detections with a non-zero overlap were considered for the final localization evaluation. The IoU (Eq. 2.4) was computed for each matched pair, and the per-patient mean IoU was obtained by averaging across all frames. The overall VERA localization performance was then summarized by reporting the IoU mean and standard deviation of the per-patient mean IoUs across the entire cohort.

9.4.2 *Low-risk* vs *High-risk* Classification Performance

For each patient, the lesion risk category was inferred by aggregating all N lesion-level predictions produced by the AI system across the examination. Each detection i provided a bounding box $\{bb_i\}$, a predicted class $\mathrm{c}_i \in \{\mathrm{LR}, \mathrm{HR}\}$ and an associated confidence $P_c(i)$. For each class $\mathrm{C} \in \{\mathrm{LR}, \mathrm{HR}\}$, the following quantities were computed:

The number of occurrences of class C was computed as:

$$\mathrm{Occ}_C = \sum_{i=1}^{N} \mathbf{1}(\mathrm{c}_i = C) \tag{9.1}$$

Hence, the mean probability (MeanProb_C) of class C across all detections of that class was:

$$\text{MeanProb}_c = \frac{\sum_{i=1}^{N} P_c(i)\,\mathbf{1}(\text{c}_i = C)}{\text{Occ}_C} \tag{9.2}$$

At this point, a weighted probability (S(p)_C) was further obtained by multiplying the MeanProb_C per the percentage of detections for such class (PercDet_C):

$$\text{PercDet}_C = \frac{\text{Occ}_C}{N} \tag{9.3}$$

$$S_C(p) = \text{MeanProb}_C \times \text{PercDet}_C \tag{9.4}$$

For each patient, the final predicted class corresponds to $\arg\max_C S_C(p)$. Ground truth was defined as *High-risk* if any lesion annotated in the patient's video was labeled as *SCC*, *leucoplakia*, or *Uncertain*. Precision (Eq. 2.1), Recall (Eq. 2.2), and F1-score (Eq. 2.3) were computed to evaluate VERA classification performance.

9.5 Results

As of the reporting date, the study included 110 patients. In-office and eventually intraoperative examinations were conducted for each patient, and approximately 700 endoscopic frames were manually saved and annotated by clinicians. Data represented both WL and NBI modalities, a full spectrum of benign and malignant conditions, and a huge variety of acquisition configurations (e.g., lighting, angle, image quality).

The distribution of devices used for the acquisition confirmed a predominance of flexible endoscopy, consistent with standard outpatient practice: 74% flexible endoscopes, 26% rigid endoscopes, and $< 1\%$ exoscopes.

Table 9.1 summarizes the distribution of labeled lesion types in the clinical study cohort. The dataset is heavily imbalanced toward *High-risk* conditions, with *SCC* and *leucoplakia* representing 51% and 25% of all cases, respectively.

Lesion Localization Performance

For the 110 enrolled patients, the per-patient mean IoU was calculated as the average of IoU values from all frames in which AI and clinician annotations overlapped. The overall mean IoU for the cohort was 0.74 ± 0.09. Moreover, performance was stable across different endoscopic devices, with mean IoU values of 0.74 ± 0.08 for flexible scopes, 0.73 ± 0.09

Table 9.1 Distribution of lesion types in the clinical study cohort.

Lesion Type	**Percentage**
SCC	51%
Leukoplakia	25%
Granuloma	9%
Cyst	0%
Papilloma	3%
Polyp	8%
Reinke's edema	2%
Uncertain	2%

for rigid scopes, and 0.71 ± 0.09 for exoscopes. This consistency indicates robustness to variations in optical modality, equipment, and examination conditions. Nonetheless, a substantial proportion of frames (approximately 40%) showed no overlapping detections between the AI system and clinicians. This can be attributed either to frames in which the larynx was not clearly visible and thus classified by the system as *Uninformative (U)*, or to cases in which the AI failed to localize the lesion. These observations highlight the challenges inherent to real-world deployment and emphasize the importance of incorporating temporal information to further strengthen system performance.

Low-risk vs. *High-risk* Classification Performance

The patient-level predictions demonstrated strong concordance with clinician-provided classification labels, addressing the limitations of frame-level predictions mentioned earlier with temporal aggregation. In fact, the AI system was able to generate a final risk assessment for 106 out of 110 patients. When aggregating frame-level AI outputs and grouping lesions into *Low-risk* and *High-risk* categories, the VERA system achieved a Precision of 93%, a Recall of 79%, and an F1-score equal to 85% in identifying *High-risk* lesions (Table 9.2). These results indicate that the system provides reliable patient-level risk stratification, helping clinicians prioritize cases according to the urgency of required treatment.

To better understand the VERA failure cases, all misclassified patients were inspected to identify recurring patterns in the incorrect predictions. Representative cases are illustrated in Figure 9.2. Generally, visualization limitations contributed directly to the AI's errors. Most *High-risk* lesions wrongly classified as *Low-risk* instances occurred for patients with small, poorly illuminated, or partially occluded malignant lesions. As shown in Figure 9.2(A)-(C), 9

Table 9.2 Patient-level performance of the VERA system for *Low-risk* versus *High-risk* classification. Metrics were computed from the aggregated confusion matrix of 110-patient analyses.

Class	**Precision**	**Recall**	**F1-score**
Low-risk	0.80	0.54	0.64
High-risk	0.93	0.79	0.85
Weighted Avg	**0.89**	**0.75**	**0.82**

errors out of 12 involved frames affected by motion blur, or underexposure. *Leucoplakia* was also particularly error-prone, as its flat, whitish appearance can resemble benign inflammatory changes as well as specular reflections or saliva. In case (E) of Figure 9.2, lesions belonging to the *leucoplakia* category were annotated by clinicians; however, the AI detection module was not applied as the frame was evaluated as *Uninformative (U)*. Example (F) represents a following frame from the same video recording in which the quality improved, and the detection module identified a lesion. However, the predicted bounding box included a specular reflection, and the risk classification resulted in a wrong prediction. Furthermore, *leucoplakia* can often develop in parallel to other benign lesions at advanced stages (5 cases out of 12). This can force the model to focus on the largest or most visually salient benign lesions, overlooking the smaller malignant abnormalities. Conversely, the number of *Low-risk* lesions categorized as *High-risk* cases was limited. An example is illustrated in Figure 9.2 (G): here, a granuloma was incorrectly classified as a malignant lesion, likely due to the presence of abnormal vascular patterns. Case (D) can be interpreted as a false negative; however, it represents a borderline and ambiguous scenario, where Reinke's edema is mild and lacks clear morphological cues. In contrast, case (H) corresponds to a more evident failure, as the leukoplakia is extensive and clearly visible even to a non-expert observer.

9.6 Discussion

The predominance of *High-risk* lesions (76% of cases) reflects the typical patient population, while the high percentage of flexible endoscopies (74%) confirms the widespread adoption of this modality for laryngeal examination in outpatient settings. The integration and preliminary clinical validation of the AI-assisted laryngoscopy, i.e., the VERA prototype, demonstrated promising results in terms of lesion localization and risk stratification across multiple endoscopic devices and settings. Data analysis exhibited good concordance on

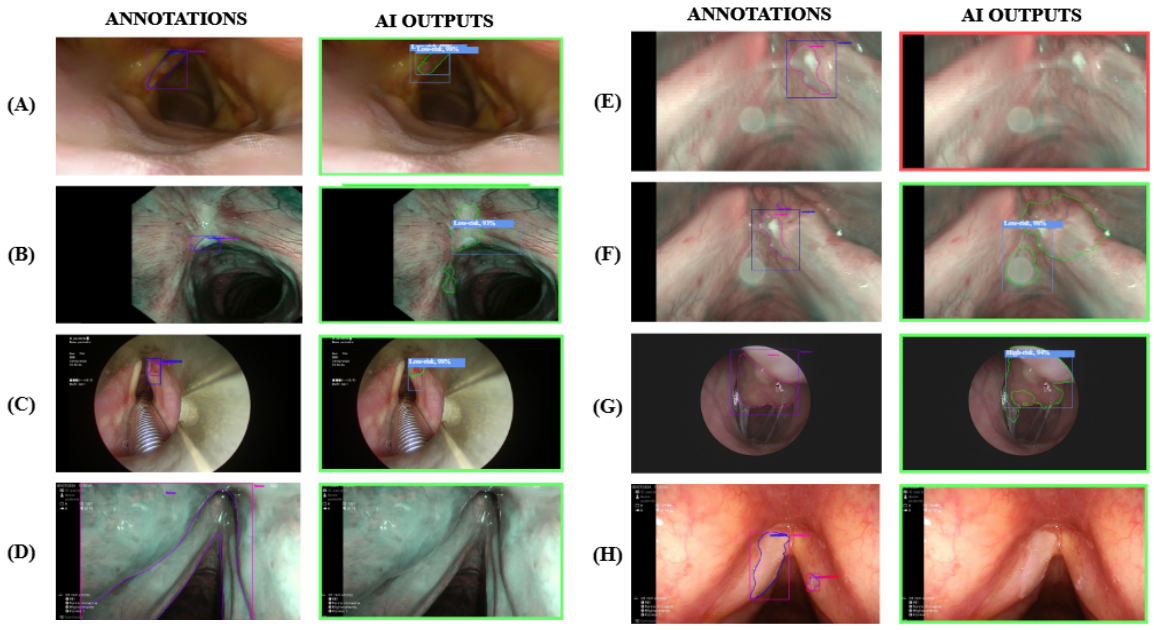


Figure 9.2 Representative examples of misclassification of failure cases from the clinical study. (A-C) Examples of *High-risk* lesions misclassified as *Low-risk* due to motion blur, underexposure, or lesion ambiguity. (G) Example of a benign lesion (granuloma) misclassified as *High-risk* due to vascular enhancement. Frames (D) and (H) correspond to false-negative cases.

detected lesion bounding boxes with manual annotations by clinicians (mean IoU of 0.74), as well as a robust *High-risk* classification performance (F1-score = 85%).

Importantly, the proposed pipeline is designed to operate in real time. Informative frame selection continuously runs in the background, enabling efficient processing by restricting lesion detection and binary risk assessment to informative frames only. More computationally demanding tasks, such as tumor margin delineation, can be selectively activated by the clinician on demand, ensuring both responsiveness and practical usability during routine examinations.

However, it is important to discuss some issues related to the current study and its outcomes. Firstly, the large number of frames without overlapping detections underscores the dependence of the analysis on the subjective selection and annotation of frames performed by clinicians. Moreover, detection failures highlight the challenges of applying automated systems in real clinical settings and the critical role of temporal information integration to mitigate errors related to image quality, motion artifacts, and heterogeneous acquisition. To date, analysis of the VERA results has been limited to lesion detection and risk classification. The performance of tumor margin delineation will be taken into account in future work as more *SCC* cases are collected and annotated. Future research should also prioritize architectures that integrate temporal dynamics and employ more diverse datasets with well-balanced class representation to enhance generalization across lesion types and imaging conditions. To conclude, results confirmed the potential of AI-assisted laryngoscopic assessment, while emphasizing the need for further optimization and prospective validation on larger datasets.

Chapter 10

Conclusions and Open Challenges

10.1 Thesis Contributions

10.1.1 Summary of Key Results

This thesis presented a comprehensive framework for AI-assisted laryngoscopy. It provides support for several tasks: informative frame selection, lesion detection, risk stratification, fine-grained classification, and tumor margin delineation. The results obtained across these tasks, as well as those from synthetic data generation and preliminary clinical validation, demonstrated that AI systems can approach expert-level performance, reduce diagnostic variability, and function seamlessly alongside routine clinical workflows. Key findings include:

- The informative frame selection model achieved high accuracy in filtering relevant frames. This can enable efficient downstream analysis and accelerate clinical decision-making.
- The YOLO-based detection model demonstrated robust lesion localization across diverse datasets, with $\text{mAP}_{@\text{IoU=0.5}}$ scores (Eq. 2.6) reaching 0.74. The proposed SRE-YOLO variant can help in detecting small and early lesions, improving the baseline performance by 5%.
- The *Low-risk* versus *High-risk* lesion stratification module provided overall F1-scores (Eq. 2.3) ranging from 85% to 89% on three test datasets from different clinical centers. Comparing the AI results with labels assigned by clinicians at several levels of expertise, the AI predictions correlated well with those of laryngologists, outperforming less

experienced residents or practitioners. Thereby, the AI system could serve as a valuable second opinion tool, particularly in settings with limited expertise availability.

- The *MedSAM* segmentation approach was identified as the most effective method to delineate tumor margins, facilitating surgical planning and intraoperative guidance. It achieved DSC scores (Eq. 2.8) > 0.90 on different test sets, outperforming domain-specific CNNs.

- Synthetic data generation pipeline produced high-quality images that enhanced models' training and generalization. Synthetic data can be useful to mitigate annotated data scarcity and improve metrics for underrepresented lesion types (+3-6% in fine-grained classification F1-score).

- The preliminary clinical study analysis validated the feasibility and usability of the integrated AI prototype (VERA) into the real clinical workflow. The tool could assist clinicians in real-time procedures, enhancing diagnostic confidence and potentially reducing procedure times. The analysis revealed positive feedback from clinicians regarding the system's benefits within the diagnostic process and confirmed the results obtained in the experimental evaluation (85% F1-score for risk classification).

10.1.2 Comparison with the State of the Art

Compared with previously published approaches, the most significant contribution of this thesis is the collection of the largest and most heterogeneous dataset for AI-assisted laryngoscopy reported to date. Existing studies typically rely on monocentric cohorts, often restricted to a single optical modality or lesion type. In contrast, our dataset aggregated data from multiple international centers, covering both WL and NBI modalities, from in-office and intraoperative examinations. In addition, it represents a broad spectrum of benign and malignant lesions, including rare subtypes. Such diversity, unprecedented in the field, enabled an extensive multicenter validation strategy across retrospective datasets, which demonstrated consistent and robust performance across populations, devices, and clinical environments.

Furthermore, through the integration of the developed VERA prototype into routine clinical workflows, the system underwent an initial prospective evaluation, advancing the clinical translation of AI models in laryngology.

From a more technical point of view, each AI component was compared against existing methods in the literature. They consistently provided competitive or superior performance across multiple tasks:

- The informative frame selection model outperformed previous results by achieving a higher F1-score (93%) on a more diverse and comprehensive dataset, while allowing the real-time processing of both WL and NBI frames.

- The lesion detection model surpassed or matched the $mAP_{@IoU=0.5}$ of existing YOLO-based methods, such as (Azam et al., 2022a), (Wellenstein et al., 2023), (Kim et al., 2023), and (Nie et al., 2025), and employed a more exhaustive validation setup. The proposed SRE-YOLO model enriched the detection capability, especially in addressing small and early-stage lesions, which are often missed by conventional algorithms.

- The classification module obtained improved F1-scores (Eq. 2.3) compared to prior works, particularly in multi-center evaluations. This highlights its generalizability. In particular, the risk stratification approach reached metrics close to those in (Wei et al., 2024) and (Srivastava et al., 2025), in which both WL and NBI data were included. The former study reported an F1-score of 89% in binary classification on data from 2 hospitals, while the latter used a similar RoI-based cropping strategy to improve normal, benign, premalignant, and malignant discrimination in WL and NBI data (F1-score=87%). The multi-class task achieved Accuracy values aligned with those in (Yin et al., 2021) (77% versus 73%), in which the same set of 7 lesion categories was applied to a smaller single-center dataset of WL in-office images.

- The segmentation capacity of the SAM-based models, including variants pre-trained on large medical datasets, was investigated for the task of tumor margin definition. *MedSAM* outperformed traditional segmentation models and prompt-free counterparts, achieving higher DSC values thanks to the guidance of prompts and robustness across datasets.

- A synthetic data generation pipeline was introduced as an innovative approach to produce realistic and diverse synthetic laryngeal images. By conditioning the generation of lesions with specific clinical information about location and histological type, synthetic data can serve for data augmentation and dataset imbalance mitigation.

10.2 Impact of the Proposed Research

10.2.1 Scientific Impact

This thesis advances the scientific understanding of AI applications in laryngology by demonstrating the effectiveness of DL models for multiple diagnostic tasks. The comprehensive dataset, collected from diverse international centers and curated to ensure annotation consistency and standardization, sets a new benchmark for robust model development and validation. The integration of foundation models like *SAM* and the translation of super-resolution techniques, initially proposed in different domains, into laryngology highlights the potential of transfer learning and tailored architectural design in enhancing AI diagnostic performance. Additionally, the synthetic data generation approach offers new prospects for addressing data scarcity and class imbalance in medical imaging datasets. Through clear and transparent reporting of methodologies and results, this work contributes to the reproducibility of AI research in the field.

10.2.2 Clinical Impact

The developed framework has the potential to significantly impact clinical practice by enhancing diagnostic and surgical accuracy, reducing variability, and improving workflow efficiency. The system can streamline the review process by automatically extracting informative frames from videos, allowing clinicians to focus on clinically relevant information. The lesion detection and risk stratification modules provide real-time decision support, aiding in early diagnosis, urgent cases discrimination, and appropriate treatment planning. The tumor margin delineation tool can assist surgeons in preoperative planning and intraoperative navigation, potentially improving surgical outcomes. The preliminary clinical validation indicates that the VERA system can be easily integrated into existing workflows to support operators and streamline the process. Moreover, its modular design allows for adaptation and scalability across different clinical settings, making it a versatile tool for otolaryngologists worldwide.

10.2.3 Economical Impact

AI implementation in laryngoscopy has the potential to yield significant economic benefits for healthcare systems. The autonomous nature of AI has the potential to accelerate routine examinations and reduce clinicians' workload. A key consequence is improved resource allocation for healthcare providers. Early, accurate diagnosis supports timely intervention and

may reduce reliance on more invasive and costly treatments needed in advanced stages. The improved diagnostic accuracy can also decrease the rates of misdiagnosis and unnecessary procedures, leading to cost savings for both healthcare providers and patients. Over time, these enhancements can produce meaningful cost savings, making AI-assisted laryngoscopy a cost-effective option for improving patient care, particularly in settings with limited specialists or advanced equipment.

10.2.4 Translational and Social Impact

Lastly, the adoption of AI-assisted laryngoscopy can have a positive social impact. It can allow general practitioners and less experienced clinicians to deliver high-quality care on a global scale, reducing disparities in healthcare access and optimizing patient outcomes and quality of life. In this perspective, beyond the technical development and evaluation of the system, this work also included investigations and efforts toward its potential clinical translation. Comprehensive system documentation and device specifications were produced to support future regulatory and validation processes, and a competitor analysis was carried out to contextualize the proposed AI solution within the current landscape of clinical and commercial solutions. Furthermore, through the start-up mentorship program, an initial business plan was developed, and further aspects concerning innovation, usability, and potential commercialization were explored.

10.3 Limitations and Future Work

This study is constrained by some limitations. Although the dataset is the largest and most heterogeneous to date, it still exhibits class imbalance, with *High-risk* lesions represented as the predominant type. This imbalance could potentially bias model training and affect performance on underrepresented classes. Despite efforts to mitigate this limitation through synthetic data generation, additional studies are required to assess both the privacy preservation aspects and the clinical reliability of the synthetic samples. Another limitation related to the AI modules implementation is the fact that they may still be sensitive to variations in image quality, acquisition protocols, and patient demographics not fully captured in the training data. Hence, while initial clinical integration was successfully achieved with encouraging results, larger-scale prospective studies across multiple populations and healthcare environments are necessary to fully validate the system's effectiveness. Possible future developments include:

- Further refinement of the AI pipeline, including the exploration of advanced generative and foundation models and expansion of the dataset, focusing on underrepresented lesion types and diverse patient populations.
- Integration of temporal information or multimodal data, combining endoscopic images with clinical metadata, voice analysis, or other imaging modalities to provide a more comprehensive assessment and face the task of personalized treatment recommendation.
- Advancing the system toward commercial deployment, first within Europe and subsequently in broader international markets, to make the technology accessible in numerous clinical centers. This process may be supported through partnerships with major distributors in the medical device sector and dedicated dissemination activities to increase visibility and adoption.
- Refining evaluation protocol and proposing task-specific metrics that more accurately reflect clinical needs. For instance, in the context of lesion detection for screening purposes, a high IoU is not strictly necessary; what matters clinically is that the lesion is identified, even if the predicted bounding box does not precisely match the clinician's annotation.

References

Alyafi, B., Diaz, O., and Marti, R. (2020). Dcgans for realistic breast mass augmentation in x-ray mammography. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, pages 473–480. SPIE.

Araujo, T., Santos, C. P., De Momi, E., and Moccia, S. (2019). Learned and handcrafted features for early-stage laryngeal scc diagnosis. *Medical & Biological Engineering & Computing*, 57(12):2683–2692.

Azam, M. A., Sampieri, C., Ioppi, A., Africano, S., Vallin, A., Mocellin, D., Fragale, M., Guastini, L., Moccia, S., Piazza, C., et al. (2022a). Deep learning applied to white light and narrow band imaging videolaryngoscopy: toward real-time laryngeal cancer detection. *The Laryngoscope*, 132(9):1798–1806.

Azam, M. A., Sampieri, C., Ioppi, A., Benzi, P., Giordano, G. G., De Vecchi, M., Campagnari, V., Li, S., Guastini, L., Paderno, A., et al. (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*, 12:900451.

Bajhaiya, D., Unni, S. N., and Koushik, A. (2023). Deep learning–powered generation of artificial endoscopic images of gi tract ulcers. *iGIE*, 2(4):452–463.

Berardini, D., Migliorelli, L., Galdelli, A., and Marín-Jiménez, M. J. (2025). Edge artificial intelligence and super-resolution for enhanced weapon detection in video surveillance. *Engineering Applications of Artificial Intelligence*, 140:109684.

Bhattacharjee, R., Devi, K. S., and Vijaykanth, S. (2023). Detecting laryngeal cancer lesions from endoscopy images using deep ensemble model. In *2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT)*, pages 1–6. IEEE.

Borgli, H., Thambawita, V., Smedsrud, P. H., Hicks, S., Jha, D., Eskeland, S. L., Randel, K. R., Pogorelov, K., Lux, M., Nguyen, D. T. D., et al. (2020). Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data*, 7(1):283.

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., and Jemal, A. (2024). Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 74(3):229–263.

Bur, A. M., Zhang, T., Chen, X., Kavookjian, H., Kraft, S., Karadaghy, O., Farrokhian, N., Mussatto, C., Penn, J., and Wang, G. (2023). Interpretable computer vision to detect and classify structural laryngeal lesions in digital flexible laryngoscopic images. *Otolaryngology–Head and Neck Surgery*, 169(6):1564–1572.

Cen, Q., Pan, Z., Li, Y., and Ding, H. (2019). Laryngeal tumor detection in endoscopic images based on convolutional neural network. In *2019 IEEE 2nd International conference on electronic information and communication technology (ICEICT)*, pages 604–608. IEEE.

Cho, W. K., Lee, Y. J., Joo, H. A., Jeong, I. S., Choi, Y., Nam, S. Y., Kim, S. Y., and Choi, S.-H. (2021). Diagnostic accuracies of laryngeal diseases using a convolutional neural network-based image classification system. *The Laryngoscope*, 131(11):2558–2566.

Dao, T. T. P., Huynh, T.-L., Pham, M.-K., Le, T.-N., Nguyen, T.-C., Nguyen, Q.-T., Tran, B. A., Van, B. N., Ha, C. C., and Tran, M.-T. (2024). Improving laryngoscopy image analysis through integration of global information and local features in vofocd dataset. *Journal of Imaging Informatics in Medicine*, 37(6):2794–2809.

Darvish, M. and Kist, A. M. (2024). A generative method for a laryngeal biosignal. *Journal of Voice*.

Davaris, N., Voigt-Zimmermann, S., Kropf, S., and Arens, C. (2019). Flexible transnasal endoscopy with white light or narrow band imaging for the diagnosis of laryngeal malignancy: diagnostic value, observer variability and influence of previous laryngeal surgery. *European Archives of Oto-Rhino-Laryngology*, 276(2):459–466.

Dhull, A. K., Atri, R., Dhankhar, R., Chauhan, A. K., and Kaushal, V. (2018). Major risk factors in head and neck cancer: a retrospective analysis of 12-year experiences. *World journal of oncology*, 9(3):80.

Diamantis, D. E., Gatoula, P., and Iakovidis, D. K. (2022). Endovae: Generating endoscopic images with a variational autoencoder. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5. IEEE.

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, S., Guo, J., Huang, D., Liu, Y., Zhang, X., and Lu, S. (2025). Diagnostic accuracy of deep learning-based algorithms in laryngoscopy: a systematic review and meta-analysis. *European Archives of Oto-Rhino-Laryngology*, 282(1):351–360.

Dunham, M. E., Kong, K. A., McWhorter, A. J., and Adkins, L. K. (2022). Optical biopsy: automated classification of airway endoscopic findings using a convolutional neural network. *The Laryngoscope*, 132:S1–S8.

Esmaeili, N., Davaris, N., Boese, A., Illanes, A., Friebe, M., and Arens, C. (2022). Contact endoscopy – narrow band imaging (CE-NBI) data set for laryngeal lesion assessment.

Esmaeili, N., Sharaf, E., Gomes Ataide, E. J., Illanes, A., Boese, A., Davaris, N., Arens, C., Navab, N., and Friebe, M. (2021). Deep convolution neural network for laryngeal cancer classification on contact endoscopy-narrow band imaging. *Sensors*, 21(23):8157.

Fiz, I., Koelmel, J. C., and Sittel, C. (2018). Nature and role of surgical margins in transoral laser microsurgery for early and intermediate glottic cancer. *Current Opinion in Otolaryngology & Head and Neck Surgery*, 26(2):78–83.

Galdran, A., Costa, P., and Campilho, A. (2019). Real-time informative laryngoscopic frame classification with pre-trained convolutional neural networks. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 87–90. IEEE.

Garrucho, L., Kushibar, K., Osuala, R., Diaz, O., Catanese, A., Del Riego, J., Bobowicz, M., Strand, F., Igual, L., and Lekadir, K. (2023). High-resolution synthesis of high-density breast mammograms: Application to improved fairness in deep learning based mass detection. *Frontiers in oncology*, 12:1044496.

Gidaris, S., Singh, P., and Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

Gómez, P., Kist, A. M., Schlegel, P., Berry, D. A., Chhetri, D. K., Dürr, S., Echternach, M., Johnson, A. M., Kniesburges, S., Kunduk, M., et al. (2020). Bagls, a multihospital benchmark for automatic glottis segmentation. *Scientific data*, 7(1):186.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Inaba, A., Hori, K., Yoda, Y., Ikematsu, H., Takano, H., Matsuzaki, H., Watanabe, Y., Takeshita, N., Tomioka, T., Ishii, G., et al. (2020). Artificial intelligence system for detecting superficial laryngopharyngeal cancer with high efficiency of deep learning. *Head & Neck*, 42(9):2581–2592.

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. (2021). nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211.

Jaworek, A. J., Earasi, K., Lyons, K. M., Daggumati, S., Hu, A., and Sataloff, R. T. (2018). Acute infectious laryngitis: A case series. *Ear, Nose & Throat Journal*, 97(9):306–313.

Ji, B., Ren, J., Zheng, X., Tan, C., Ji, R., Zhao, Y., and Liu, K. (2020). A multi-scale recurrent fully convolution neural network for laryngeal leukoplakia segmentation. *Biomedical Signal Processing and Control*, 59:101913.

Joseph, J. S. and Vidyarthi, A. (2024). Dual deep learning and feature-based models for classification of laryngeal squamous cell carcinoma using narrow band imaging. *Traitement du Signal*, 41(1):237.

Kang, Y.-F., Yang, L., Hu, Y.-F., Xu, K., Cai, L.-J., Hu, B.-B., and Lu, X. (2025). Self-attention mechanisms-based laryngoscopy image classification technique for laryngeal cancer detection. *Head & Neck*, 47(3):944–955.

Khazrak, I., Zainaee, S., M. Rezaee, M., Ghasemi, M., and C. Green, R. (2025). Feasibility of improving vocal fold pathology image classification with synthetic images generated by ddpm-based genai: a pilot study. *European Archives of Oto-Rhino-Laryngology*, pages 1–15.

Kim, G. H., Hwang, Y. J., Lee, H., Sung, E.-S., and Nam, K. W. (2023). Convolutional neural network-based vocal cord tumor classification technique for home-based self-prescreening purpose. *BioMedical Engineering OnLine*, 22(1):81.

Kim, G. H., Sung, E.-S., and 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):51.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026.

Kumazawa, Y., Ikenoyama, Y., Takamatsu, M., Kido, K., Namikawa, K., Tokai, Y., Yoshimizu, S., Horiuchi, Y., Ishiyama, A., Yoshio, T., et al. (2025). Differences in clinical characteristics between missed and detected laryngopharyngeal cancers. *Journal of Gastroenterology and Hepatology*.

Kuznetsov, K., Lambert, R., and Rey, J.-F. (2006). Narrow-band imaging: potential and limitations. *Endoscopy*, 38(01):76–81.

Kwon, I., Wang, S.-G., Shin, S.-C., Cheon, Y.-I., Lee, B.-J., Lee, J.-C., Lim, D.-W., Jo, C., Cho, Y., and Shin, B.-J. (2025). Diagnosis of early glottic cancer using laryngeal image and voice based on ensemble learning of convolutional neural network classifiers. *Journal of Voice*, 39(1):245–257.

Laves, M.-H., Bicker, J., Kahrs, L. A., and 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.

Li, C., Liu, H., Liu, Y., Feng, B. Y., Li, W., Liu, X., Chen, Z., Shao, J., and Yuan, Y. (2024). Endora: Video generation models as endoscopy simulators. In *International conference on medical image computing and computer-assisted intervention*, pages 230–240. Springer.

Lim, B., Son, S., Kim, H., Nah, S., and Mu Lee, K. (2017). Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Lin, X., Xiang, Y., Yu, L., and Yan, Z. (2024). Beyond adapting sam: Towards end-to-end ultrasound image segmentation via auto prompting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 24–34. Springer.

Liu, S., Chen, Z., Yang, Q., Yu, W., Dong, D., Hu, J., and Yuan, Y. (2025). Polyp-gen: Realistic and diverse polyp image generation for endoscopic dataset expansion. *arXiv preprint arXiv:2501.16679*.

Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J. (2018). Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768.

Luan, B., Sun, Y., Tong, C., Liu, Y., and Liu, H. (2019). R-fcn based laryngeal lesion detection. In *2019 12th International Symposium on Computational Intelligence and Design (ISCID)*, volume 2, pages 128–131. IEEE.

Ma, J., He, Y., Li, F., Han, L., You, C., and Wang, B. (2024). Segment anything in medical images. *Nature Communications*, 15:654.

Matava, C., Pankiv, E., Raisbeck, S., Caldeira, M., and 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):44.

Moccia, S., De Momi, E., and Mattos, L. S. (2017). Laryngeal dataset.

Moccia, S., Vanone, G. O., De Momi, E., Laborai, A., Guastini, L., Peretti, G., and Mattos, L. S. (2018a). Learning-based classification of informative laryngoscopic frames. *Computer methods and programs in biomedicine*, 158:21–30.

Moccia, S., Vanone, G. O., De Momi, E., and Mattos, L. S. (2018b). Nbi-Infframes.

National Cancer Institute, S. R. P. (2025). Seer cancer stat facts. https://seer.cancer.gov/statfacts/. Accessed: 2025-11-05.

Nie, X., Zhang, X., Wang, D., Liu, Y., Xing, L., and Liu, W. (2025). Laryngeal cancer diagnosis based on improved yolov8 algorithm. *Machine Learning: Science and Technology*, 6(1):015011.

Osuala, R., Skorupko, G., Lazrak, N., Garrucho, L., García, E., Joshi, S., Jouide, S., Rutherford, M., Prior, F., Kushibar, K., et al. (2023). medigan: a python library of pretrained generative models for medical image synthesis. *Journal of Medical Imaging*, 10(6):061403–061403.

Pan, X., Bai, W., Ma, M., and Zhang, S. (2022). Rant: A cascade reverse attention segmentation framework with hybrid transformer for laryngeal endoscope images. *Biomedical Signal Processing and Control*, 78:103890.

Patrini, I., Ruperti, M., Moccia, S., Mattos, L. S., Frontoni, E., and De Momi, E. (2020). Transfer learning for informative-frame selection in laryngoscopic videos through learned features. *Medical & biological engineering & computing*, 58(6):1225–1238.

Qiu, M., Li, Y., Huang, W., Zhang, H., Zheng, W., Fan, X., and Lei, W. (2024). 3d-lsptm: An automatic framework with 3d-large-scale pretrained model for laryngeal cancer detection using laryngoscopic videos. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–4. IEEE.

Ren, J., Jing, X., Wang, J., Ren, X., Xu, Y., Yang, Q., Ma, L., Sun, Y., Xu, W., Yang, N., et al. (2020). Automatic recognition of laryngoscopic images using a deep-learning technique. *The Laryngoscope*, 130(11):E686–E693.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Rosen, C. A., Amin, M. R., Sulica, L., Simpson, C. B., Merati, A. L., Courey, M. S., Johns III, M. M., and Postma, G. N. (2009). Advances in office-based diagnosis and treatment in laryngology. *The Laryngoscope*, 119(S2):S185–S212.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Sampieri, C., Azam, M. A., Ioppi, A., Baldini, C., Moccia, S., Kim, D., Tirrito, A., Paderno, A., Piazza, C., Mattos, L. S., et al. (2024). Real-time laryngeal cancer boundaries delineation on white light and narrow-band imaging laryngoscopy with deep learning. *The Laryngoscope*, 134(6):2826–2834.

Sampieri, C., Baldini, C., Azam, M. A., Moccia, S., Mattos, L. S., Vilaseca, I., Peretti, G., and Ioppi, A. (2023). Artificial intelligence for upper aerodigestive tract endoscopy and laryngoscopy: a guide for physicians and state-of-the-art review. *Otolaryngology–Head and Neck Surgery*, 169(4):811–829.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Srivastava, R., Kumar, N., and Sandhan, T. (2025). Binary classification of laryngeal images utilising resnet-50 cnn architecture. *Indian Journal of Otolaryngology and Head & Neck Surgery*, 77(2):644–651.

Suárez-Quintanilla, J., Cabrera, A. F., and Sharma, S. (2023). Anatomy, head and neck: larynx. In *StatPearls [internet]*. StatPearls Publishing.

Tan, M. and Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR.

Vilaseca, I., Valls-Mateus, M., Nogués, A., Lehrer, E., López-Chacón, M., Avilés-Jurado, F. X., Blanch, J. L., and Bernal-Sprekelsen, M. (2017). 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.

Walton, C., Conway, E., Blackshaw, H., and Carding, P. (2017). Unilateral vocal fold paralysis: a systematic review of speech-language pathology management. *Journal of Voice*, 31(4):509–e7.

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2021). Scaled-yolov4: Scaling cross stage partial network. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 13029–13038.

Wang, C.-Y., Liao, H.-Y. M., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., and Yeh, I.-H. (2020). Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 390–391.

Wei, J., Li, Y., Qiu, M., Chen, H., Fan, X., and Lei, W. (2024). Sam-fnet: Sam-guided fusion network for laryngo-pharyngeal tumor detection. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2667–2672. IEEE.

Wellenstein, D. J., Woodburn, J., Marres, H. A., and van den Broek, G. B. (2023). Detection of laryngeal carcinoma during endoscopy using artificial intelligence. *Head & Neck*, 45(9):2217–2226.

Wilmes, C. M., BSc, A. G., Marres, H. A., Wellenstein, D. J., and van den Broek, G. B. (2025). A systematic review of the clinical impact of implementing artificial intelligence in upper aerodigestive tract endoscopy. *Head & Neck*.

Xiong, H., Lin, P., Yu, J.-G., Ye, J., Xiao, L., Tao, Y., Jiang, Z., Lin, W., Liu, M., Xu, J., et al. (2019). Computer-aided diagnosis of laryngeal cancer via deep learning based on laryngoscopic images. *EBioMedicine*, 48:92–99.

Yan, P., Li, S., Zhou, Z., Liu, Q., Wu, J., Ren, Q., Chen, Q., Chen, Z., Chen, Z., Chen, S., et al. (2023). Automated detection of glottic laryngeal carcinoma in laryngoscopic images from a multicentre database using a convolutional neural network. *Clinical Otolaryngology*, 48(3):436–441.

Yao, P., Witte, D., Gimonet, H., German, A., Andreadis, K., Cheng, M., Sulica, L., Elemento, O., Barnes, J., and Rameau, A. (2022). Automatic classification of informative laryngoscopic images using deep learning. *Laryngoscope investigative otolaryngology*, 7(2):460–466.

Yin, L., Liu, Y., Pei, M., Li, J., Wu, M., and Jia, Y. (2021). Laryngoscope8: Laryngeal image dataset and classification of laryngeal disease based on attention mechanism. *Pattern Recognition Letters*, 150:207–213.

Zhang, J., Lei, J., Xie, W., Fang, Z., Li, Y., and Du, Q. (2023a). Superyolo: Super resolution assisted object detection in multimodal remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15.

Zhang, K. and Liu, D. (2023). Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*.

Zhang, L., Rao, A., and Agrawala, M. (2023b). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847.

Zhang, L., Wu, L., Wei, L., Wu, H., and Lin, Y. (2023c). A novel framework of manifold learning cascade-clustering for the informative frame selection. *Diagnostics*, 13(6):1151.

Zhang, X., Zhao, J., Zong, D., Ren, H., and Gao, C. (2025). Taming vision transformers for clinical laryngoscopy assessment. *Journal of Biomedical Informatics*, 162:104766.

Zhao, Q., He, Y., Wu, Y., Huang, D., Wang, Y., Sun, C., Ju, J., Wang, J., and Mahr, J. J.-l. (2022). Vocal cord lesions classification based on deep convolutional neural network and transfer learning. *Medical physics*, 49(1):432–442.

Zhou, L., Jiang, H., Li, G., Ding, J., Lv, C., Duan, M., Wang, W., Chen, K., Shen, N., and Huang, X. (2023). Point-wise spatial network for identifying carcinoma at the upper digestive and respiratory tract. *BMC Medical Imaging*, 23(1):140.

Żurek, M., Jasak, K., Niemczyk, K., and Rzepakowska, A. (2022). Artificial intelligence in laryngeal endoscopy: systematic review and meta-analysis. *Journal of Clinical Medicine*, 11(10):2752.

Żurek, M., Rzepakowska, A., Osuch-Wójcikiewicz, E., and Niemczyk, K. (2019). Learning curve for endoscopic evaluation of vocal folds lesions with narrow band imaging. *Brazilian Journal of Otorhinolaryngology*, 85(6):753–759.