



Università di Genova

PHD PROGRAM

Biotechnologie in medicina traslazionale

Curriculum:

Bioimaging

Dipartimento di Medicina Sperimentale (DIMES)

DEVELOPMENT OF ARTIFICIAL INTELLIGENCE- ENHANCED OPTICAL BIOPSY TOOLS FOR ENDOSCOPIC EVALUATION OF LARYNGEAL TUMORS

Candidate:

Claudio Sampieri

Supervisors:

Prof. Giorgio Peretti

Prof. Francesco Mora

TABLE OF CONTENTS

TABLE OF CONTENTS.....	3
INTRODUCTION.....	7
THE HEAD AND NECK AND LARYNGEAL TUMORS	8
THE ENDOSCOPY OF THE LARYNX	11
History.....	11
Endoscopy techniques.....	14
Imaging techniques	19
ARTIFICIAL INTELLIGENCE APPLIED TO IMAGES: THE COMPUTER VISION	22
Artificial intelligence	22
Computer vision	26
Machine Learning Methods	29
Training, Validation, and Test Datasets.....	31
Artificial intelligence and laryngeal endoscopy.....	32
SYSTEMATIC REVIEW OF THE LITERATURE	34
PURPOSE OF THE THESIS.....	45
MATERIALS AND METHODS	47
STUDY DESIGN AND OVERVIEW.....	48
Dataset collection and framework.....	48
INFORMATIVE FRAME SELECTION.....	50
Purpose.....	50
Dataset.....	50
Model development.....	51
Evaluation metrics.....	51
Video testing	53
Statistical analysis	53
LARYNGEAL TUMORS DETECTION	54
Purpose.....	54
Dataset.....	54
Model development.....	55
Evaluation metrics.....	56
Statistical analysis	57
LARYNGEAL TUMORS DIAGNOSIS	58
Purpose.....	58
Dataset.....	58
Model development.....	60
Evaluation metrics.....	60

Comparison with physicians and large language models	63
Statistical analysis	63
LARYNGEAL TUMORS SEGMENTATION	65
Purpose.....	65
Dataset.....	65
Evaluation metrics.....	67
Comparison with human physicians	69
Statistical analysis	69
LARYNGEAL MOTILITY ASSESSMENT	70
Purpose.....	70
Dataset.....	70
Model development.....	71
Evaluation metrics.....	75
Statistical analysis	76
RESULTS	77
INFORMATIVE FRAME SELECTION.....	78
Dataset overview	78
Model performance on internal dataset.....	78
External validation	79
Model performance on the video dataset	80
Study interpretation.....	81
LARYNGEAL TUMORS DETECTION	83
Dataset overview	83
Model performance on internal dataset.....	83
External validation	84
Real-time feasibility	84
Study interpretation.....	86
LARYNGEAL TUMORS DIAGNOSIS	87
Dataset overview	87
CADx model performances on internal and external datasets	88
CADx model performance comparison with clinical raters and Large Language Models	89
Study interpretation.....	90
LARYNGEAL TUMORS SEGMENTATION	91
Dataset overview	91
Internal test performance	91
External validation	92
Comparison with human physicians	93

Table of contents

Real-time video performance	95
Study interpretation.....	97
LARYNGEAL MOTILITY ASSESSMENT	98
Dataset overview.....	98
Binary classification performance	98
Multi-class classification performance	99
Ablation study and feature relevance.....	101
Study interpretation.....	102
DISCUSSION	103
General considerations	104
Informative frames selection.....	108
Laryngeal tumors detection.....	110
Laryngeal tumors diagnosis	113
Laryngeal tumors segmentation	116
Laryngeal motility assessment	119
Conclusions.....	122
LIST OF ABBREVIATIONS	125
REFERENCES.....	127

INTRODUCTION

THE HEAD AND NECK AND LARYNGEAL TUMORS

Head and neck mucosal squamous cell carcinoma (HNSCC) includes a variety of malignancies arising from the oral cavity, pharynx, and larynx. These cancers are notably influenced by lifestyle factors such as tobacco and alcohol use, and the epidemiology of these cancers continues to evolve. Globally, HNSCC accounts for almost 900,000 new cases annually¹. The incidence of HNSCC varies significantly by geographic region, largely due to differences in risk factors such as smoking rates, alcohol consumption, and prevalence of human papillomavirus (HPV) infections.

Tobacco and alcohol are the primary risk factors for head and neck cancers. Tobacco use alone is responsible for a substantial proportion of cases, especially for laryngeal tumors, and its combined use with alcohol significantly increases the risk. HPV is an increasingly recognized cause, particularly for oropharyngeal cancers, which have seen a rise in incidence even as tobacco-related cancers decline². Other risk factors include Epstein-Barr virus infection, particularly for nasopharyngeal cancers, and occupational exposure to substances such as chromium, nickel, and asbestos.

Over the past two decades, there has been a general decline in the incidence and mortality of HNSCC in high-income countries, attributed mainly to reduced tobacco use, higher prevalence of HPV-driven carcinomas (which carry a better prognosis), and earlier diagnosis³. However, disparities persist, with Asian, African, and Latin American populations exhibiting higher incidence and mortality trends compared to Europe and North America. Socioeconomic factors also play a significant role, with lower socioeconomic status and educational attainment linked to higher risks of HNSCCs⁴.

The International Agency for Research on Cancer (IARC) projects the incidence of HNSCC to rise significantly from 892,128 new cases in 2022 to 1,402,113 by 2050 (**Figure 1**). Similarly, mortality rates are projected to increase from 458,365 deaths per year in 2022 to 832,659 by 2050 with higher trends for low-income populations⁵ (**Figure 2**).

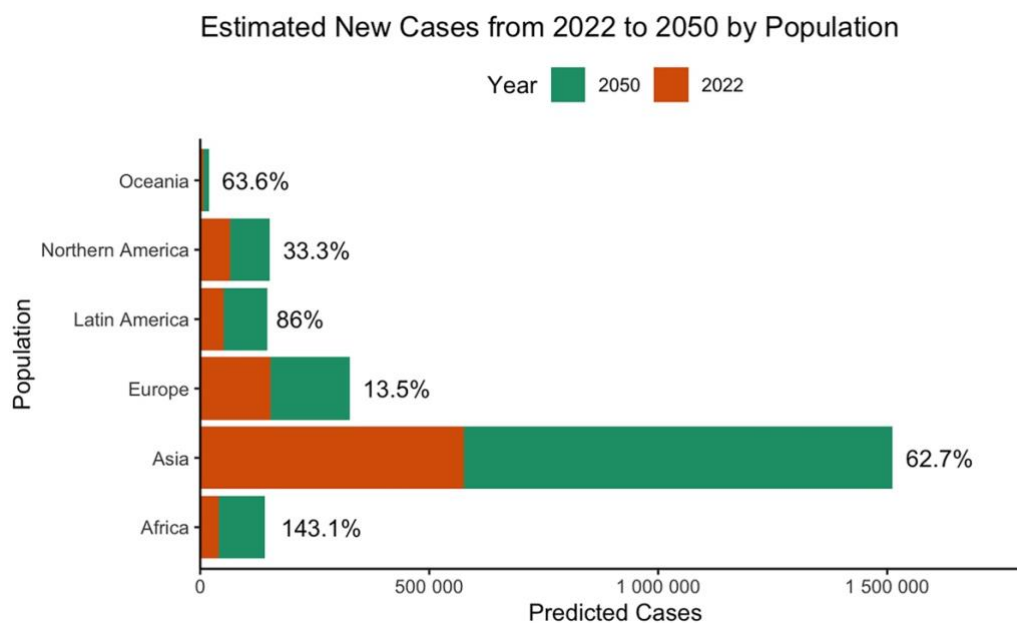


Figure 1: Estimated new cases of oral cavity, pharyngeal, and laryngeal cancer from 2022 to 2050. For every continent, the increment (percentage) is reported. The figure was made with tabular data available at <https://gco.iarc.fr/tomorrow/en>

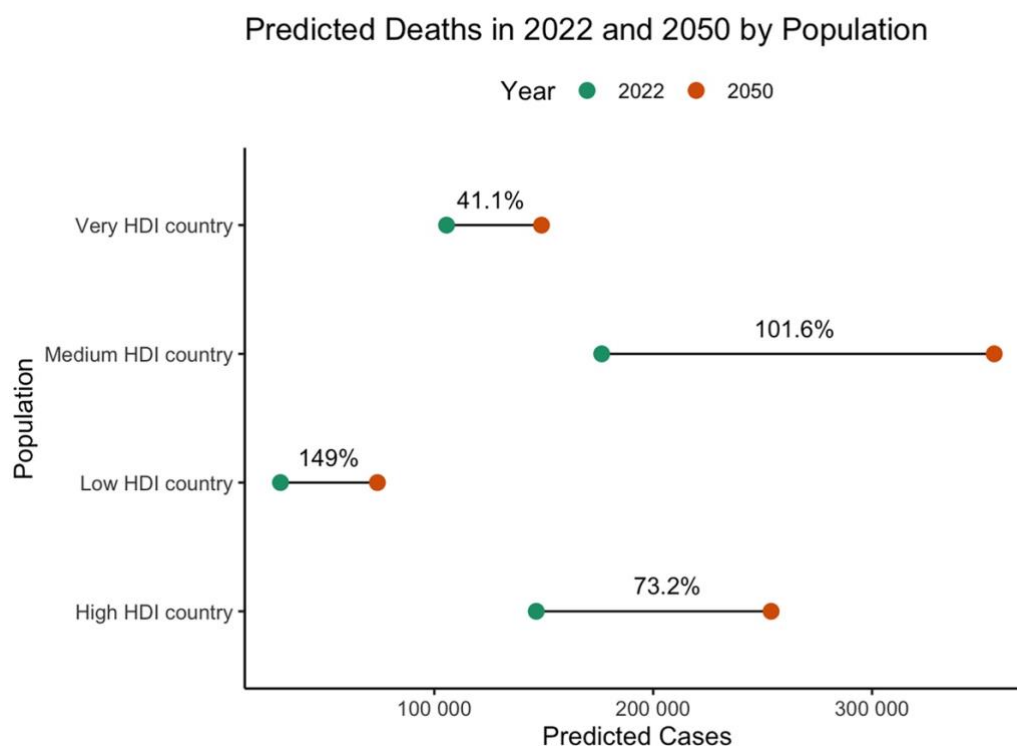


Figure 2: Estimated deaths due to oral cavity, pharyngeal, and laryngeal cancer from 2022 to 2050. For every group of countries, the increment (percentage) is reported. HDI: Human Development Index. The figure was made with tabular data available at <https://gco.iarc.fr/tomorrow/en>

These observed disparities, with a general decrease in the mortality rates for HNSCC in high-income countries, are largely due to reductions in smoking and improvements in early detection methods in the developed countries. In fact, the overall 5-year survival rate for head and neck cancers remains less than 50%, and this is strictly correlated to the stage at diagnosis.

Cancers detected at stage I-II generally have better outcomes. The five-year survival rate for patients diagnosed with stage I or II can be as high as 70-90%⁶. Treatment at these stages often involves limited surgery and/or localized radiation, which can be very effective in local control of the disease while providing optimal functional outcomes with consequent high quality of life and social functioning. Conversely, at higher stages, extended mutilating surgery or high-dose chemoradiation treatments are required with heavy implications regarding many social aspects of these patients and their families, including financial toxicity repercussions.

Endoscopic examination plays a fundamental role in evaluating patients with suspected head and neck cancer. It provides direct access to the mucosal lining of the upper aerodigestive tract (UADT), allowing clinicians to recognize lesions or abnormal tissue changes that may escape detection by other diagnostic approaches. Detecting disease at an early stage through endoscopy facilitates timely diagnosis and management, which can substantially improve prognosis. Identifying tumors sooner often reduces the need for highly invasive therapies and increases the chances of curative treatment. Nevertheless, endoscopy is not without shortcomings. Its effectiveness depends on the expertise of the physician performing the procedure, which introduces variability in both recognition and interpretation of findings. In addition, subtle or early lesions may remain undetected, particularly in regions with limited visibility.

THE ENDOSCOPY OF THE LARYNX

Almost all laryngeal diseases, or their manifestations, are detectable by analyzing the inner aspects of the larynx and its motion. Especially neoplastic diseases, either benign or malignant, arise in the vast majority of cases from the surface of the mucosa lining the larynx and therefore it is paramount to visualize these structures in vivo.

History

Endoscopic visualization of the larynx has undergone remarkable evolution over the past two centuries. Before the 1800s, anatomical understanding of the larynx was limited to post-mortem studies. The first in vivo visualization occurred in 1827, when the Swiss L. Senn used a small mirror to examine a child's larynx. Benjamin Guy Babington, an English physician, followed with a more structured tool (the glottoscope) in 1829, integrating a laryngeal mirror and an epiglottic retractor, marking the origin of indirect laryngoscopy.

A major breakthrough occurred in 1854 when the Spanish singer, voice teacher, and scientist Manuel García, residing in London, successfully performed the first auto-laryngoscopy (**Figure 3**). He first visualized his own larynx using a dental mirror and a second handheld mirror to reflect sunlight. Garcia transferred the method to study other humans by a system of mirrors for reflecting the sunlight to illuminate the dental mirror inserted through the mouth, so as to directly analyze the phonation. He presented his invention to the Royal Society and, in 1862, was granted an honor of a medical degree followed by many international distinctions.

Though not a physician, his contribution was later adopted clinically by Johann Czermak in 1858 in Budapest and Ludwig Türck (Vienna), who introduced artificial light and head mirrors for indirect visualization—a technique that remained a diagnostic mainstay for over a century⁷.

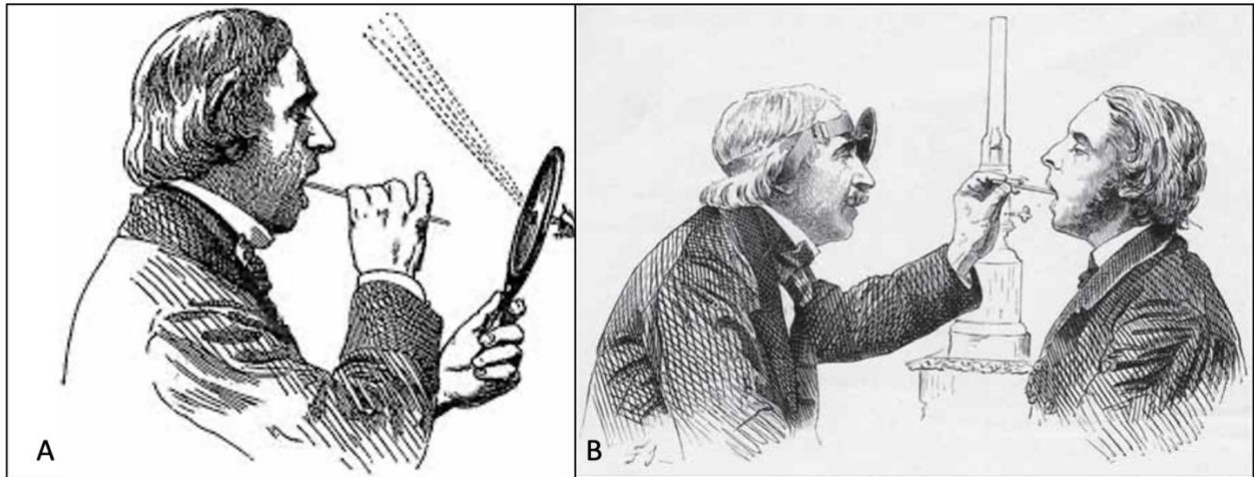


Figure 3. A) Garcia's autolaryngoscopy. B) Czermark's laryngoscopy technique

The transition from indirect to direct laryngoscopy in the late 19th century was facilitated by advancements in anesthesia and optical engineering. Direct laryngoscopy allows the direct visualization of the larynx without the use of mirrors, achieved through tube-shaped instruments coupled with a light source, known as laryngoscopes. Advances in anesthesia created the right conditions for achieving the full relaxation required for implementing this technique. In 1895 in Germany, Alfred Kirstein used a modified esophagoscope to achieve direct inspection of the vocal cords. His work laid the foundation for Gustav Killian, who, in 1910, introduced a shorter version of the tube permitting the direct visualization that he called the suspension laryngoscope. This hands-free device enabled bimanual surgery under direct vision (**Figure 4**). This innovation revolutionized laryngeal surgery and remains foundational in microlaryngeal procedures today.⁸

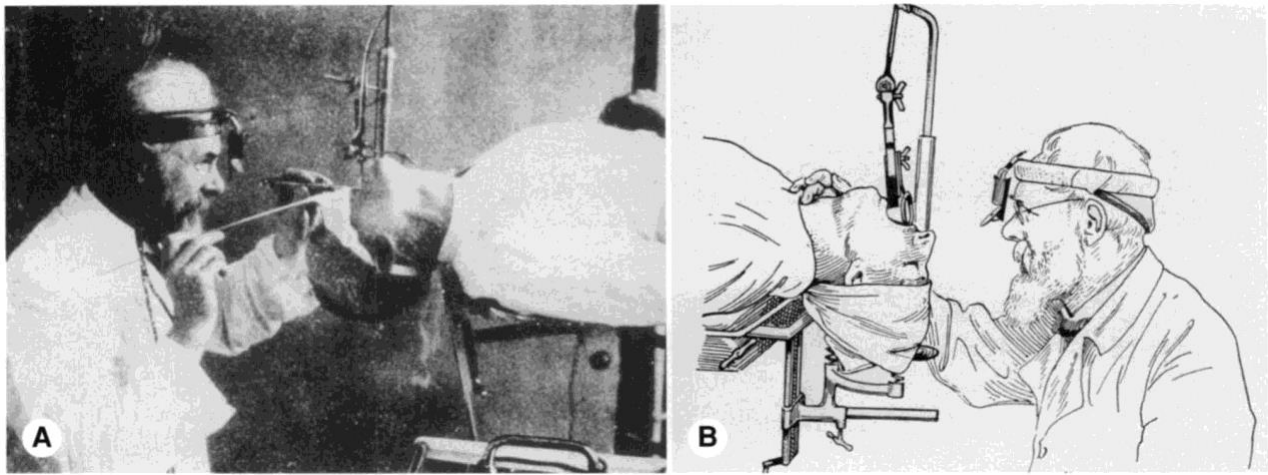


Figure 4. Gustave Killian performing suspension laryngoscopy A) real photo B) schematic representation

Initial illumination for suspension laryngoscopy was a headlight. Subsequently, many devices exploiting electricity were developed: first, an electrified handle with a proximal incandescent bulb and a prism to direct the light distally was designed. Later, a distal illumination by means of detachable electrified light-carriers was introduced. Afterwards, Broyles first demonstrated the advantage of fiberoptic light coupled with laryngoscopes in 1962. About the same time, the additional light provided by the surgical-microscope began to be applied to laryngoscopy thus enhancing the self-contained illumination of the direct laryngoscopes. There has not been dramatic change in suspension laryngoscopy since the present days with the recent advent of exoscopy as a substitute of the traditional operative microscope⁹.

Aside from the progress of direct laryngoscopy, the laryngeal examination outside the operative theater made a substantial qualitative leap with the introduction of the fibreglass-optic telescopes which made possible previously unachievable magnification and resolution of laryngeal images both intraoperatively and in the office. This breakthrough was achieved by Harald H. Hopkins (Great Britain) in the 1960s in collaboration with the Karl Storz Ltd. Company (Germany)¹⁰: they first released the first cystoscopy telescope in 1967, but similar rigid fiber endoscopes have since been used throughout the body. In turn, from there on, the Hopkins optics have been a key platform technology for laryngoscopic imaging. Finally, the advances in optics led to the introduction of the

flexible fiberoptic laryngoscope in 1976, which allowed a definitive visualization of the larynx in its natural state and a full exploration of the pharynx¹¹. The advent of digital imaging in the last decades has brought laryngoscopy to the highest standards of precision, magnification, and definition, progressively replacing analogic devices and providing innovative imaging-enhancing techniques **(Figure 5)**.

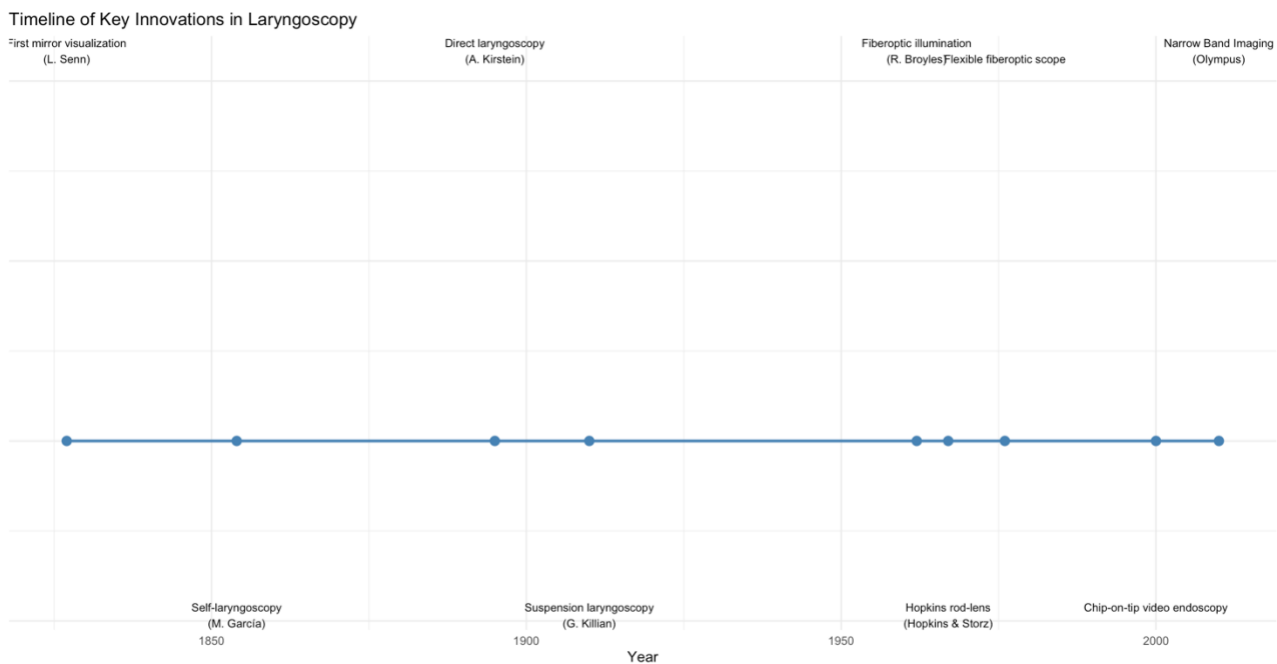


Figure 5. Summary and timeline of key Innovations in laryngoscopy

Endoscopy techniques

Compared to the traditional method of indirect laryngoscopy using a mirror, endoscopic approaches represent a major technological advance in laryngeal evaluation. They allow brighter illumination and magnification of the structures, enabling a detailed exploration of all laryngeal subsites. Endoscopy is also more adaptable for interventional procedures, which cannot be performed with mirror laryngoscopy. One of the most relevant advantages of endoscopic systems is the possibility to record both still images and video sequences. This capability has become essential not only for

clinical documentation and patient follow-up, but also for educational purposes, telemedicine, and research applications, including quantitative image analysis and computer vision-based methods.

Transoral rigid laryngoscopy

Rigid laryngeal endoscopy is performed with a 70° or 90° rigid telescope introduced transorally to visualize the pharynx and larynx. This technique provides high-quality visualization, with excellent illumination and magnification of the vocal folds, far superior to the indirect mirror technique. However, the procedure is carried out in a non-physiological setting: the patient must protrude the tongue, which can be uncomfortable, and topical anesthesia may be needed to suppress the gag reflex. Despite this, a subset of patients cannot be adequately examined due to anatomical restrictions of the soft palate or tongue base, or because of an exaggerated gag reflex. Another limitation is functional assessment: disorders such as muscle tension dysphonia or swallowing impairments cannot be fully appreciated given the artificial examination position (**Figure 6**).

Nonetheless, rigid endoscopy remains a key technique because it offers superior image resolution, especially when paired with high-definition cameras. With a stroboscopic light source, it enables detailed analysis of vocal fold vibration. Rigid telescopes also serve an important role intraoperatively, during suspension laryngoscopy, where different angled optics can be used to examine the entire larynx, including subglottic and tracheal regions.

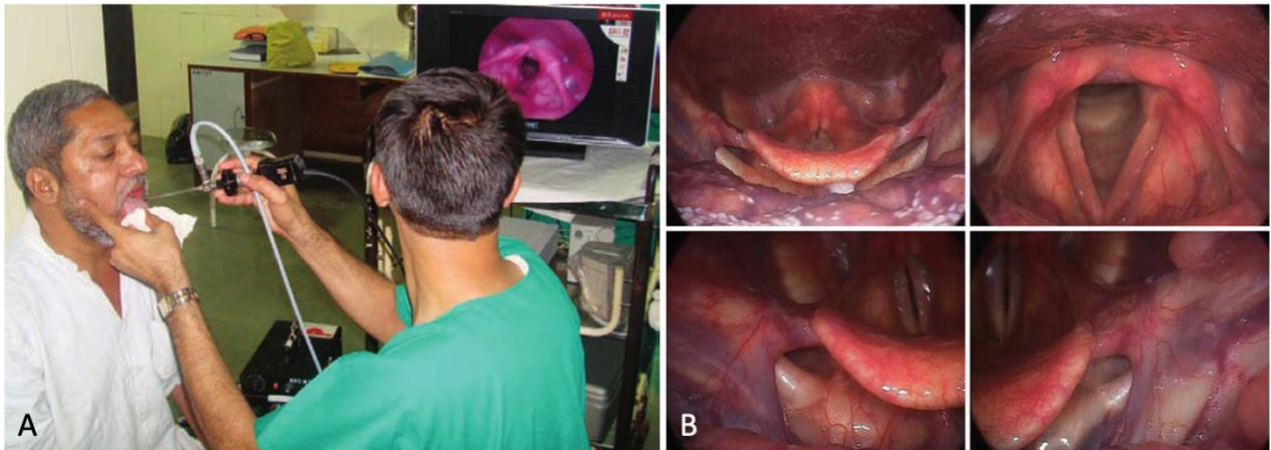


Figure 6. A) In-office rigid laryngoscopy technique. B) In-office laryngeal images obtainable when coupling the rigid telescopes with an HD camera

Transnasal flexible laryngoscopy

Flexible transnasal laryngoscopy offers a unique advantage: it is the only technique that permits examination of the UADT in a natural, physiological position. The endoscope is introduced through the nasal cavity, allowing visualization of the nasopharynx, oropharynx, larynx, and hypopharynx without altering the patient's posture. Importantly, this enables dynamic assessment during breathing, phonation, and swallowing, providing functional information that rigid laryngoscopy cannot deliver. This approach is particularly advantageous in patients with strong gag reflexes, those with anatomical limitations to oral examination, and in pediatric populations. Moreover, it is the only office-based method that allows exploration of the subglottis. The setup generally requires a flexible fiberoptic endoscope, a light source, and—when available—an endoscopic camera to enable video recording. A stroboscopic unit can also be coupled, although classical fiberoptic scopes are limited by reduced image resolution and brightness. This is due to light transmission through optical fibers, which inevitably leads to degradation of image quality over the length of the scope (**Figure 7**).

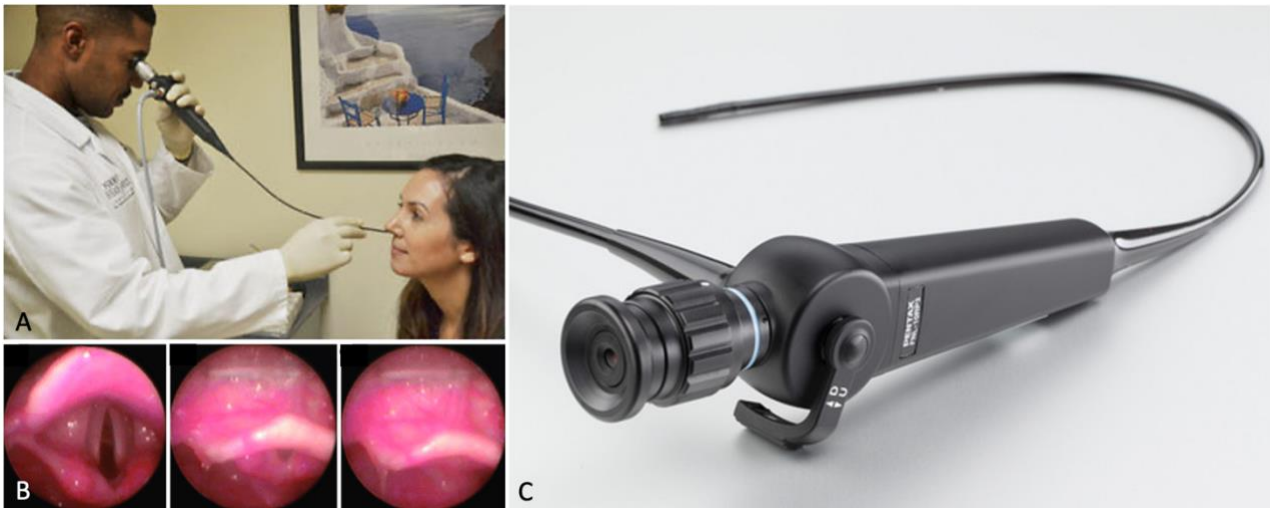


Figure 7. A) transnasal flexible laryngoscopy technique. B) laryngeal images obtainable with standard analogic fiberoptic scopes. C) transnasal flexible scope

Transnasal flexible high-definition video-laryngoscopy

Recent technological advances have transformed flexible laryngoscopy through the miniaturization of digital sensors placed directly at the distal tip of the endoscope. By eliminating fiber bundles, these systems avoid the optical losses inherent in older devices and provide images of outstanding clarity. Current high-definition video-laryngoscopes deliver image quality approaching that of rigid telescopes, while retaining the advantages of dynamic, physiological examination.

In addition to image quality, many modern scopes incorporate an operative channel that permits office-based interventions such as biopsies, laser applications, or local injections. Stroboscopic evaluation is also feasible with these devices, though it remains less accurate for very fine vibratory details due to the inherent optical limitations and the wide-angle lenses used. Such lenses are necessary to maintain a wide field of view despite the small diameter of the instrument, but they produce a characteristic geometric distortion known as “barreling,” which bends straight structures. While modern digital systems have reduced this effect compared to earlier models, it remains largely absent in rigid endoscopes (**Figure 8**).

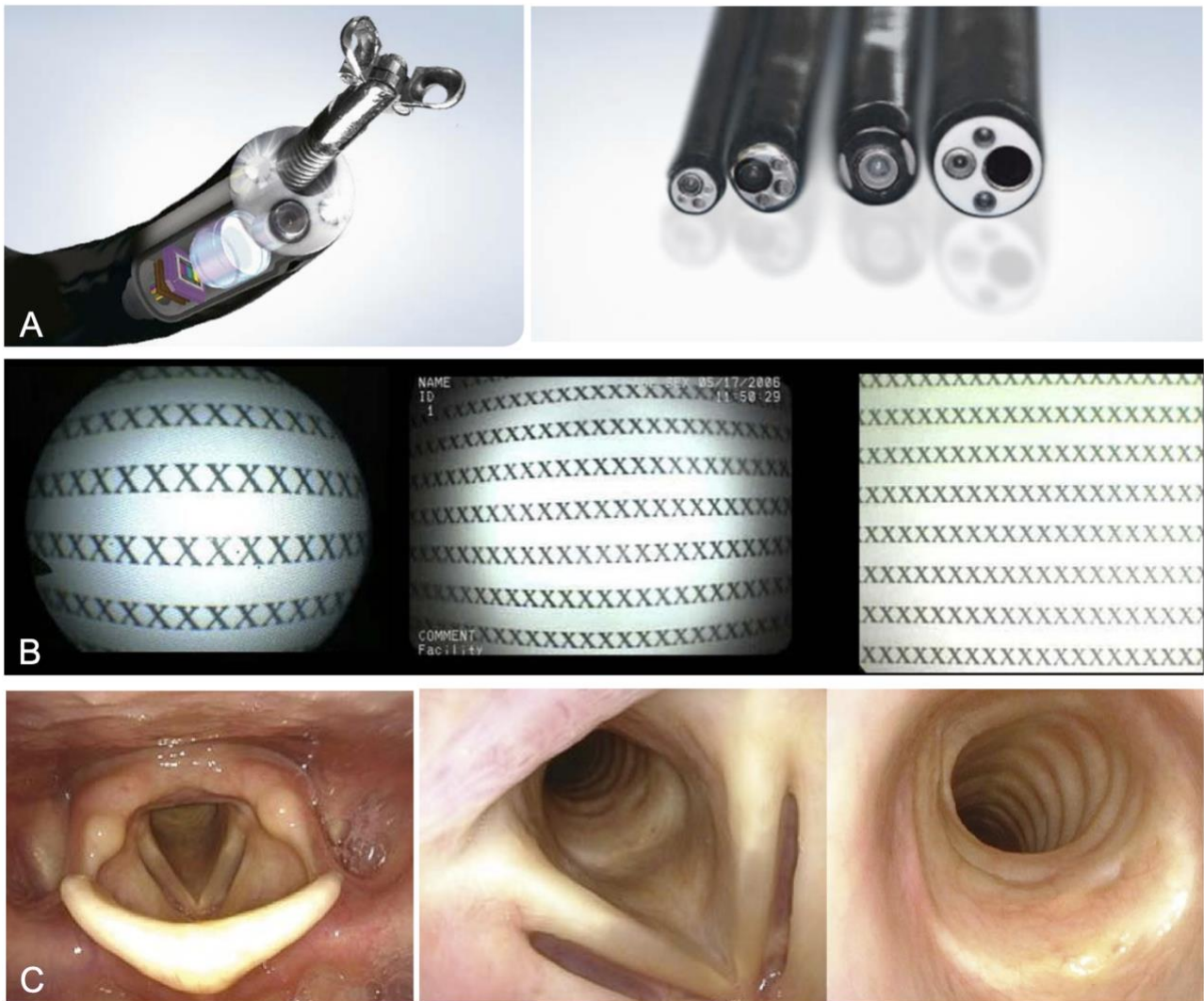


Figure 8. A) digital chip and lenses apparatus: note the operative channel and the scopes available in different diameters. B) "barreling" phenomenon with standard flexible endoscope, with digital sensor flexible endoscope and with rigid endoscope. C) images of the larynx achievable with high-definition flexible video-laryngoscope.

In-office procedures with transnasal flexible endoscopy

The advent of modern flexible transnasal endoscopes has expanded the role of laryngoscopy beyond diagnostic evaluation to include a wide range of in-office procedures. The physiological position and excellent tolerance of the transnasal approach make it particularly suitable for interventions performed under local anesthesia without the need for sedation or general anesthesia. Through the operative channel of contemporary videoendoscopes, clinicians can carry out biopsies of suspicious lesions, local steroid or botulinum toxin injections, application of topical treatments, and even selected fiber-laser procedures. These minimally invasive interventions provide several advantages:

they reduce the need for operating room resources, shorten patient recovery time, and allow for repeated procedures when necessary. Importantly, the ability to combine dynamic functional assessment with targeted intervention in the same session enhances the diagnostic and therapeutic yield of the examination, especially in the management of laryngeal and UADT disorders.

Imaging techniques

Videostroboscopy

Videostroboscopy is a specialized technique for the dynamic evaluation of vocal fold vibration. It uses a stroboscopic light synchronized with phonation to create the illusion of slow-motion movement, allowing clinicians to study the mucosal wave and vibratory characteristics of the vocal folds. Synchronization is achieved either with a neck-mounted microphone or with electroglottography. Although not a recent invention, videostroboscopy has become widely adopted due to its essential role in the diagnostic work-up of dysphonia. When combined with video recording, it allows detailed post-examination review, objective documentation, and direct comparison across follow-up visits. Key diagnostic parameters include glottic closure pattern, symmetry and periodicity of vibration, and the amplitude and regularity of the mucosal wave. Standard protocols recommend assessing vocal fold vibration at different pitches (modal, high, and low) and at different sampling rates, including ultra-slow motion (one glottal cycle every two seconds). For most structural pathologies, stroboscopy provides sufficient temporal resolution to reach an accurate diagnosis.

Narrow Band Imaging

Conventional white light contains the full visible spectrum, which often results in limited contrast between vascular structures and surrounding mucosa. Narrow Band Imaging (NBI) selectively uses blue and green wavelengths that are strongly absorbed by hemoglobin, thereby highlighting the vascular architecture of the mucosa and submucosa. Blue light is preferentially absorbed by

superficial capillaries, while green light penetrates deeper, enhancing visualization of subepithelial vessels. This selective illumination generates high-contrast images in which vascular changes—such as those associated with neoangiogenesis—are more easily identified. In malignant or pre-malignant lesions, vascular patterns shift from regular, parallel courses to irregular, perpendicular orientations enriched by intrapapillary capillary loops, which appear as dark punctate spots under NBI¹² (**Figure 9**). These vascular signatures have been shown to improve early detection of laryngeal cancer, guide the delineation of surgical margins, and support the identification of unknown primary tumors of the UADT. Multiple studies have demonstrated the superior diagnostic accuracy of NBI over conventional white light endoscopy in both preoperative and follow-up settings^{13–15}.

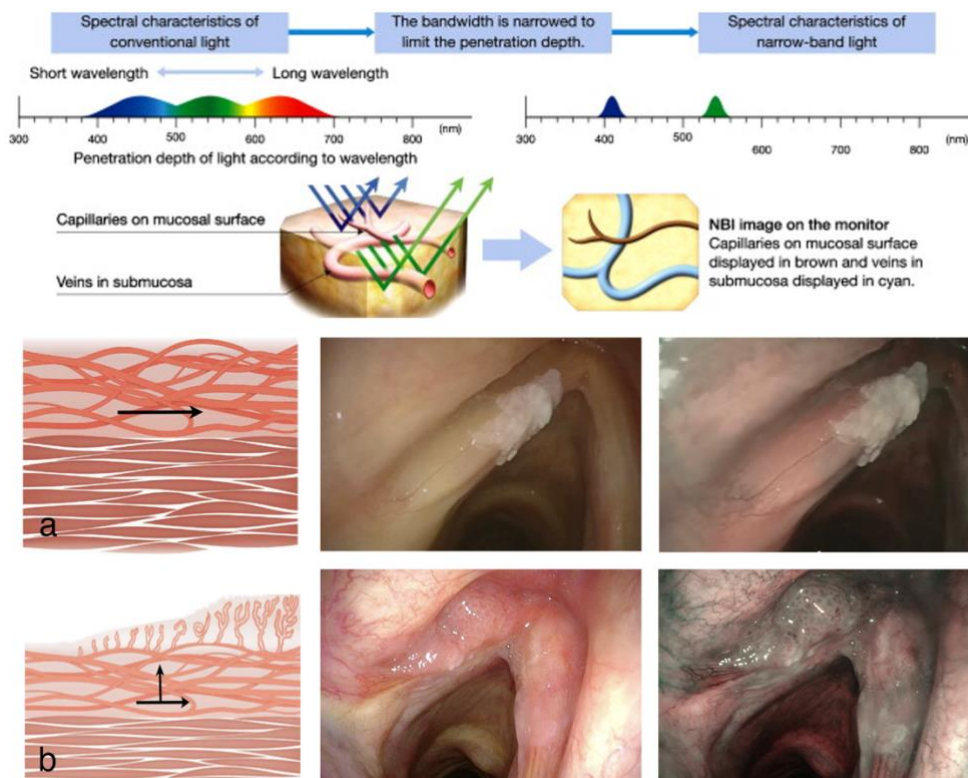


Figure 9 A) parallels vascular abnormalities nonsuspicious for malignancy: schematic representation, white light image, and NBI image. B) perpendicular vascular changes suspicious for malignancy: schematic representation, white light image, NBI image

Nowadays, laryngoscopy remains the cornerstone for evaluating lesions of UADT, offering direct visualization of mucosal abnormalities and playing a crucial role in the early detection of head and neck cancer. Nonetheless, its diagnostic accuracy is closely tied to the operator's expertise, leading to variability in the recognition of subtle or early-stage lesions. Additional constraints, such as

suboptimal illumination, restricted fields of view, and patient-related factors, may further compromise lesion detection. Moreover, laryngoscopy requires the training of highly specialized otolaryngologists or physicians, which restricts the availability of the technique and contributes to increased healthcare costs. These challenges highlight the need for complementary technologies that can improve objectivity, reproducibility, and accessibility. In this context, artificial intelligence has emerged as a promising solution, with the potential to augment conventional endoscopy by providing automated, consistent, and sensitive tools for lesion recognition, thereby supporting earlier diagnosis and more effective clinical decision-making.

ARTIFICIAL INTELLIGENCE APPLIED TO IMAGES: THE COMPUTER VISION

Artificial intelligence

Artificial intelligence (AI) is broadly defined as the ability of machines to perform tasks that traditionally require human intelligence, such as reasoning, learning, and problem solving. The concept has evolved from early rule-based expert systems to contemporary machine learning and deep learning methods, which can autonomously identify patterns in large datasets. The last decade has witnessed a dramatic expansion of AI across multiple domains of society, from finance to autonomous vehicles, with healthcare representing one of its most promising frontiers.

Classical AI was historically based on predefined rules and logical inference. In these systems, knowledge was encoded manually by experts, and the program followed deterministic rules to arrive at a conclusion. While effective in well-defined and narrow domains, these systems lacked adaptability: they could not generalize beyond the scenarios for which they were explicitly programmed. This rigidity limited their usefulness in fields such as medicine, where biological processes are complex, heterogeneous, and often unpredictable.

Machine learning (ML) represented a paradigm shift. Instead of relying solely on hand-crafted rules, ML algorithms learn from data. By analyzing large datasets, they identify statistical patterns and correlations that can then be used to make predictions on new, unseen data. Classical ML techniques include decision trees, support vector machines, and random forests, each of which can perform classification or regression tasks with varying degrees of success. In medicine, ML approaches have been used for outcome prediction, risk stratification, and early attempts at automated image classification. However, these models typically rely on engineered features: experts must decide

which variables or image characteristics are most relevant. As human experts determine the set of features for understanding the differences between data inputs, usually traditional ML models require more structured data to learn, and this limits them in their ability to process natural data in their raw form. Therefore, the effectiveness of the ML models often depends on the domain expertise of the developers and the empirical image analysis techniques that are designed to translate the image characteristics into numerical values. This exposes ML models to inherent limitations in that the human developer may not be able to translate the complex disease patterns into a finite number of feature descriptors, even if they have seen a large number of cases from the patient population. The hand-engineered features may also have difficulty being robust against the large variations of normal and abnormal patterns in the population. The performance of the traditional ML system is often limited in its discriminative power or generalizability, often resulting in overfitting. This term is used when a ML model learns the training data *too well*, including noise, random fluctuations, or irrelevant details, instead of capturing the general patterns. As a result, the model performs very accurately on the training dataset but fails to generalize when applied to new, unseen data. In practical terms, overfitting reflects a discrepancy between memorization and learning: the model “remembers” specific examples rather than identifying generalizable patterns. A ML model that overfits is like a student who memorizes all the answers to past exam questions without truly understanding the subject. That student will do well on the same practice questions but will struggle when faced with new ones that test the same concepts differently.

Deep learning (DL), a subset of machine learning, has become a transformative technology due to its superior performance in solving complex problems related to information analysis and interpretation. DL uses artificial neural networks with multiple layers that progressively extract more abstract and complex features from the input data. Unlike traditional ML models, DL does not require manual feature engineering; instead, the network learns the most relevant features during training

automatically. In imaging, DL leverages the fact that many natural entities are hierarchical, with higher-level features formed by combining lower-level ones. This is especially useful in images, which are arrays of pixel values where contrasts can form shape combinations. For example, local edge arrangements form motifs, motifs assemble into parts, and parts form complete objects. A typical DL model's first layer detects the presence or absence of edges at specific orientations and locations. The second layer recognizes motifs by identifying arrangements of edges. The third layer groups motifs into larger parts, corresponding to components of familiar objects, and subsequent layers assemble these parts into complete objects (**Figure 10**).

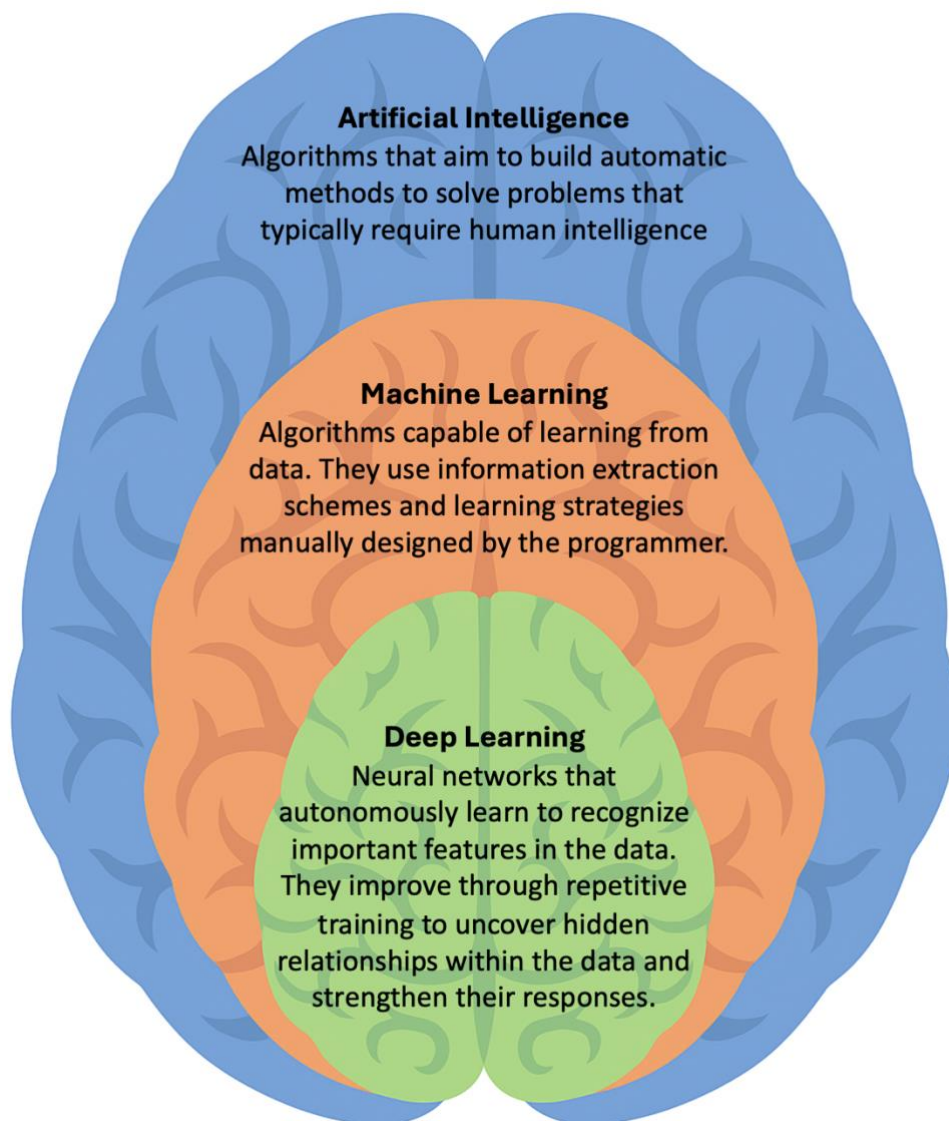


Figure 10. Schematic representation of artificial intelligence strategies

Among the different architectures of deep learning, convolutional neural networks (CNNs) are currently the most widely applied models, particularly in the field of image analysis. CNNs are specifically designed to recognize spatial hierarchies of patterns, making them highly suitable for tasks such as image classification, segmentation, and object detection. Unlike traditional machine learning models, which require manual feature engineering, CNNs can autonomously learn relevant features directly from raw image data. Through an iterative process of forward propagation and backpropagation, the network adjusts its internal weights to minimize prediction error, progressively discovering increasingly abstract and discriminative representations of the input data. This data-driven feature learning represents a key advantage over conventional approaches. If trained on a sufficiently large and diverse dataset, CNNs are able to identify subtle differences in texture, shape, or color that may escape even experienced human observers, often outperforming models based on hand-crafted features. In medical imaging, this capacity translates into the potential to detect early or atypical disease presentations that might otherwise go unrecognized. Moreover, since the learning process is automated, CNNs can process hundreds of thousands of images, far exceeding the lifelong visual exposure of a single specialist. This scalability makes them particularly robust in handling the high inter- and intra-patient variability that characterizes clinical data. However, the remarkable performance of CNNs comes at a cost. Training deep networks requires access to large, well-curated datasets that accurately represent the target population. Without such data, also DL models risk overfitting or developing biases that limit their generalizability. Furthermore, the computational demands of deep learning are substantial: high-performance graphics processing units (GPUs) are often required to handle the intensive parallel computations involved in training. Nonetheless, ongoing advances in hardware and in the design of novel architectures and training algorithms continue to expand the feasibility and efficiency of CNN-based approaches. More recently, the field of artificial intelligence has moved beyond task-specific architectures towards the

development of foundation models. These are extremely large-scale neural networks trained on vast and heterogeneous datasets, often using self-supervised learning. Once trained, they can be adapted (fine-tuned) to a wide variety of downstream tasks with relatively little additional data. Foundation models — such as large language models or vision–language models — have demonstrated unprecedented versatility and performance across domains. In the context of medical imaging, their ability to integrate multimodal inputs (images, text, clinical metadata) and transfer knowledge across tasks represents a paradigm shift. Rather than developing a new model for each diagnostic problem, foundation models may provide a general-purpose backbone that can be specialized for specific clinical applications, including laryngeal endoscopy. Although their deployment in healthcare is still in its early stages, they hold promise for accelerating progress toward more robust, scalable, and generalizable AI systems.

Computer vision

Computer vision (CV) is a branch of artificial intelligence that focuses on enabling computers to interpret and analyze visual information, such as images, videos, or other visual inputs, and to derive meaningful insights from them. In simple terms, if AI enables machines to “think,” CV allows them to “see,” observe, and ultimately understand visual data. The typical workflow of a CV application in medical imaging involves several key stages. First, feature extraction is performed on large databases of medical images. In earlier ML approaches, this step relied heavily on handcrafted descriptors designed by human experts (e.g., texture, shape, intensity) instead, for CNNs this is done automatically (**Figure 11**).

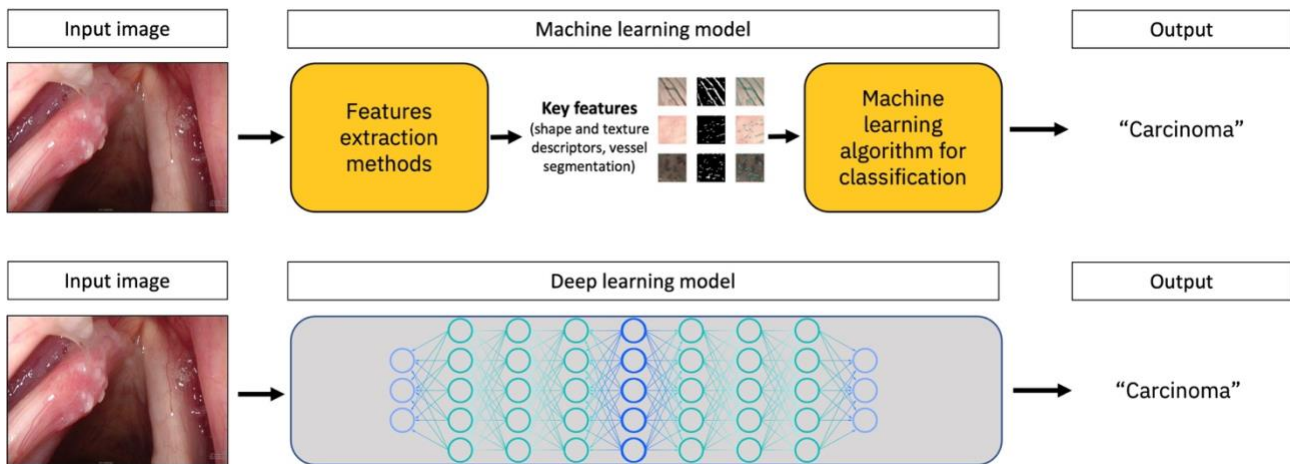


Figure 11. Machine learning and deep learning methods applied to laryngeal endoscopy

The extracted features are then used to train a model for a given task such as classification, detection, or segmentation. Once trained, the model is evaluated on an independent testing dataset, undergoing the same feature extraction process. Performance is quantified using established metrics, such as accuracy, sensitivity, specificity, etc... If the performance is inadequate, the cycle of feature engineering and model optimization is repeated until satisfactory results are achieved (**Figure 12**).

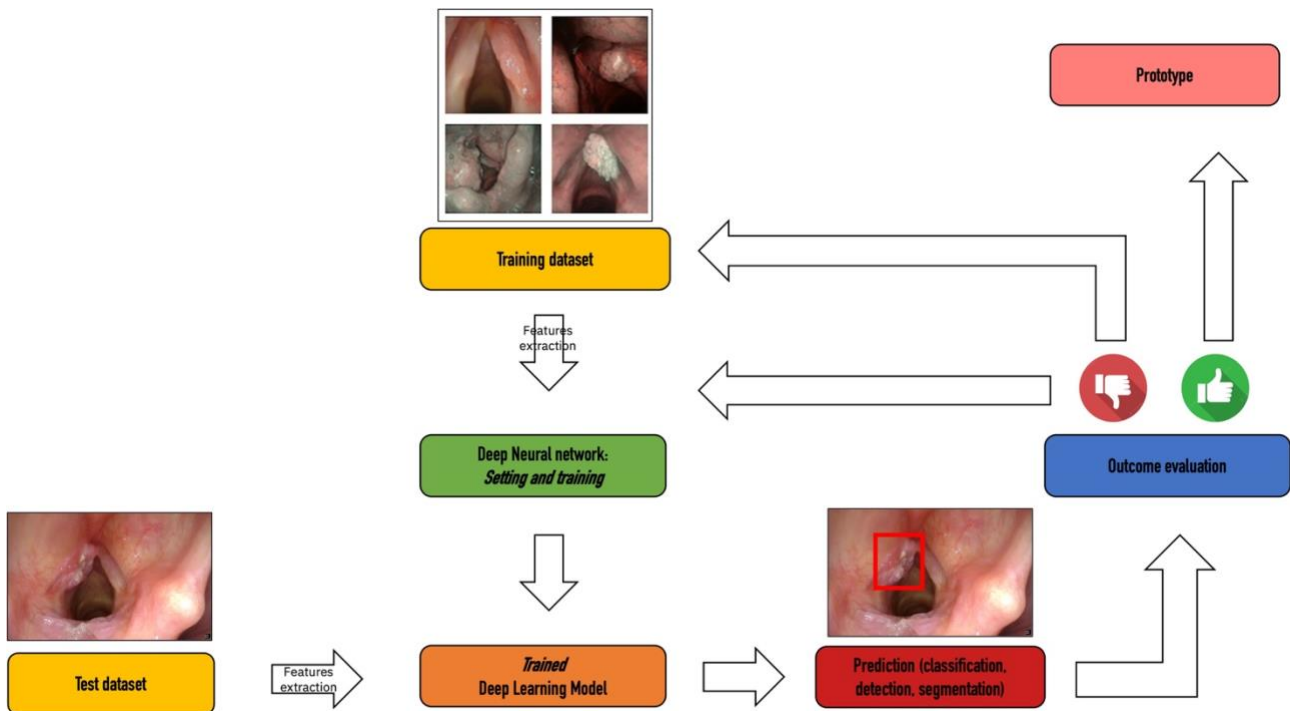


Figure 12. Workflow of a computer vision application in medical imaging

Object recognition is one of the main fields of CV and it refers to the ability of an AI to identify objects on a digital image. A ML model can be instructed to perform multiple different tasks within object recognition (**Figure 13**):

- Classification involves predicting the class of a single object in an image. The ML model output is a data label that assigns a determined class to that image.
- Localization refers to identifying the location of one or more objects in an image. To do so, the ML model outputs results in a graphical representation of the object location by means of a bounding box around the object extent.
- Detection is more challenging and combines these two tasks. The ML model output in this case is a bounding box around each object of interest in the image together with an assigned class label for each bounding box.
- Semantic segmentation recognizes objects by highlighting the specific contour of the object (instead of a coarse bounding box) by a pixel wise mask for each object in the image.

Another interesting field of CV is Motion recognition. The ML algorithm output might be different points of interest of a determined object, which are followed during motion time to track them. The movement may also be classified and explained by the model. This field of CV has been so far mainly investigated for human motions, but it may find a useful role even in medical imaging.

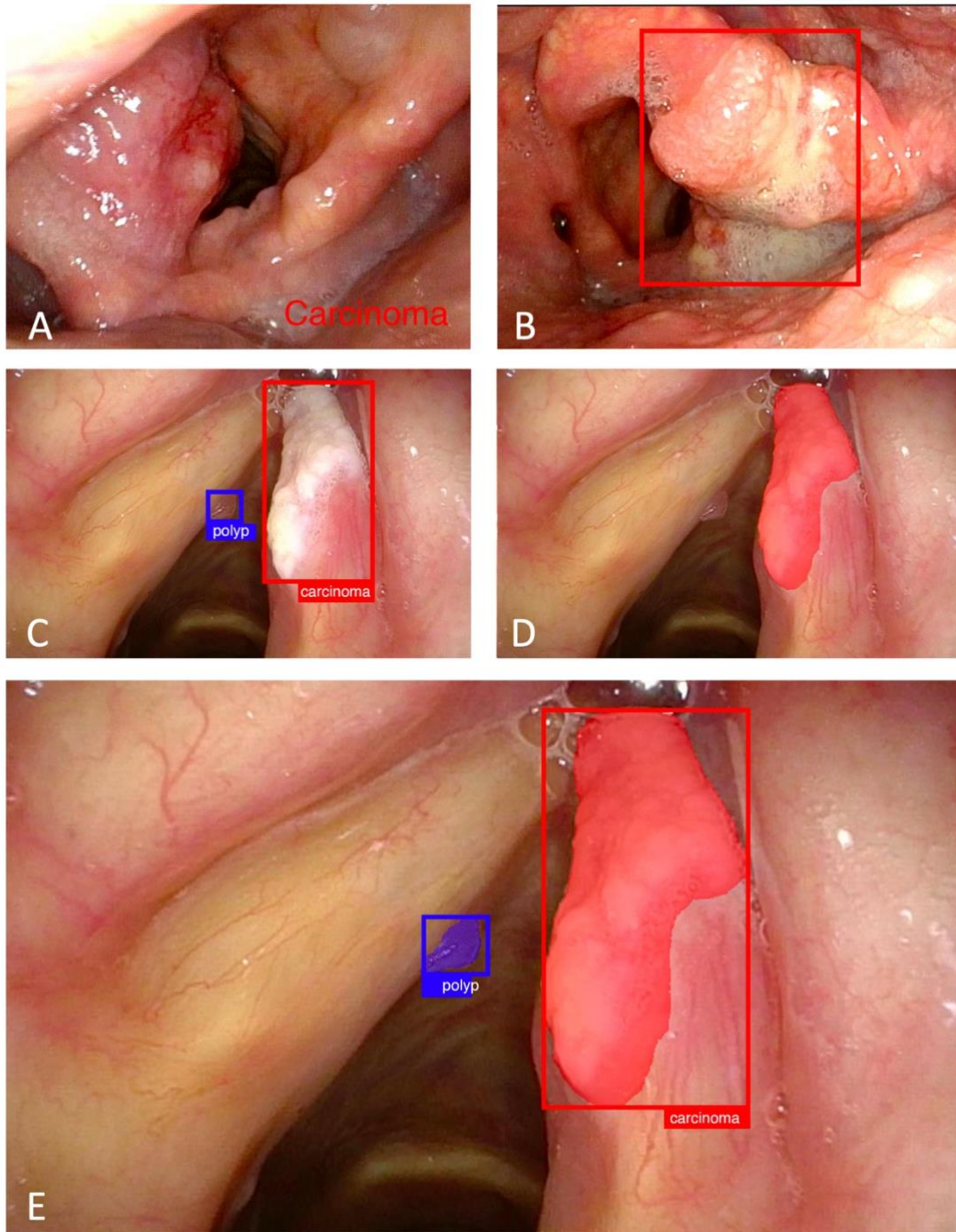


Figure 13. Computer vision tasks applied to upper-aerodigestive tract endoscopy and laryngoscopy: A) Classification; B) Localization; C) object detection; D) semantic segmentation; E) Instance segmentation

Machine Learning Methods

The most common form of ML — including most DL approaches — is supervised learning. In this setting, algorithms rely on datasets where each example has been annotated (or “labeled”) by human

experts. The quality and quantity of these annotations are critical for model performance. This is so far the most common approach in medical imaging, as the quality and correctness of the training information have to be carefully supervised by expert physicians to guarantee a trustworthy model suitable for clinical practice. For instance, to build an algorithm for image classification in laryngeal endoscopy, a large dataset of endoscopic images is collected, and each image is labeled with its diagnostic category (e.g., benign, malignant). In contrast, unsupervised learning does not require labeled data. Instead, the algorithm attempts to uncover hidden structures and patterns directly from the input. In practice, this means grouping images with similar features or learning compact representations of complex data. For laryngeal endoscopy, unsupervised approaches could be used to cluster images of lesions with similar appearance, even if no diagnosis has been assigned by an expert. A related and increasingly important approach is self-supervised learning, where models create their own training signals by predicting missing information (e.g., the next frame in a video or the masked part of an image). This allows them to learn useful features without large labeled datasets¹⁶. In medical imaging, where annotated data is expensive and time-consuming to obtain, these methods are especially promising but still not much explored. In fact, unsupervised models are still less accurate for diagnostic classification tasks compared to supervised ones, and their clinical applicability remains limited.

Of note, a particularly practical technique for medical imaging is transfer learning. This approach takes advantage of models that have already been trained on large, general-purpose image datasets, such as ImageNet, which contains millions of everyday objects¹⁷. The pretrained model learns basic visual features (edges, textures, shapes) that are also useful in medical images.

In a second step, called fine-tuning, the pretrained model is adapted to a specific medical task, for example, distinguishing benign from malignant vocal fold lesions in laryngoscopic images. This strategy significantly reduces the amount of annotated medical data required to achieve high accuracy. Transfer learning has therefore become one of the most widely used techniques in medical

AI research. However, its effectiveness is not always guaranteed: medical images often differ substantially from natural images, and the transfer of knowledge between these domains may not always lead to optimal performance¹⁸.

Training, Validation, and Test Datasets

A fundamental aspect of developing a ML model is the careful division of data into training, validation, and test sets. Each serves a different role in the process of model development and evaluation. The training set is used to teach the model. Ideally, it should be large enough to cover the variability present in the real-world target population. For medical imaging applications, and specifically for laryngeal endoscopy, the training dataset should include the full spectrum of abnormalities and lesion types encountered in clinical practice. This means capturing different lesion sizes, shapes, locations, and appearances, as well as examples across diverse patient populations. As for humans, DL models' performance improves with experience; thus, the number of training images is critical. For this purpose, image augmentation techniques are generally used, which resize, crop, turn, and flip the original images to artificially create new training information. Whether this cannot replace the value of a large and varied dataset in the first place, it can still improve the model's robustness. However, while augmentation introduces helpful variability, it cannot substitute for entirely new independent data. The augmented versions remain highly correlated with the original image and therefore provide limited additional information. For instance, if the original dataset lacks images of certain lesion types (e.g., papillomatosis), augmentation cannot generate them artificially. The validation set is used during training to optimize the model's parameters. It acts as a checkpoint, allowing developers to compare different models or to monitor how performance evolves as training progresses. Importantly, the validation set should not overlap with the final test set. In practice, it is often obtained by splitting part of the training data. Because the validation data influence model

adjustments, performance on this set reflects how well the algorithm has tuned itself but does not yet reveal its true generalizability.

The test set, in contrast, is a completely independent dataset that is not used during training. Its purpose is to provide an unbiased estimate of the model's real-world performance. By evaluating predictions on cases that the algorithm has never encountered, the test set simulates clinical deployment, answering the key question: *will the model work reliably on new patients?* Of note, while DL models can be very good at analyzing images belonging to the same group of the training cohort, they can lack accuracy when tested on different populations. Therefore, external validation datasets, acquired in different institutions, and thus diverse in terms of acquisition technique and population demographics, should be used to test the generalization ability of the model.

Datasets are often divided into training, validation, and test sets using a hold-out strategy, where a fixed proportion of the data is randomly assigned to each subset. Random allocation ensures that every sample has an equal chance of being included in training, validation, or testing, which maximizes statistical representativeness. However, in medical imaging, this approach carries the risk of data leakage: images from the same patient may appear in both training and test sets, leading to overoptimistic performance estimates (the previously mentioned problem of overfitting)¹⁹. To avoid this, many studies employ a patient-wise split, where all images from a single patient are kept in the same subset. This strategy better reflects real clinical scenarios, as the algorithm is evaluated on entirely unseen patients.

Artificial intelligence and laryngeal endoscopy

AI application in otolaryngology—head and neck surgery is a nascent field with an increasing number of publications. One of the newer domains of AI application relevant to otolaryngologists is lesion detection and image analysis on video-endoscopy, a field recently dubbed “videomics”²⁰. Laryngoscopy is one of the most common video-endoscopic procedures performed by

otolaryngologists and provides a rich source of images for AI models in CV. So far only a few publications exist in the literature regarding this topic and no clinical trial has been published²¹.

AI offers numerous advantages in the context of endoscopy for diagnosing HNSCC. AI algorithms can significantly enhance the accuracy of lesion detection by analyzing endoscopic images in real-time, identifying subtle abnormalities that might be missed by human eyes. This capability can standardize and improve diagnostic accuracy, reducing variability due to operator skill and experience. AI can also assist in differentiating between benign and malignant lesions, providing decision support to clinicians and potentially reducing the need for unnecessary biopsies.

This thesis investigates the role of artificial intelligence across the main computer vision tasks relevant to the recognition of laryngeal neoplasms during upper aerodigestive tract endoscopy.

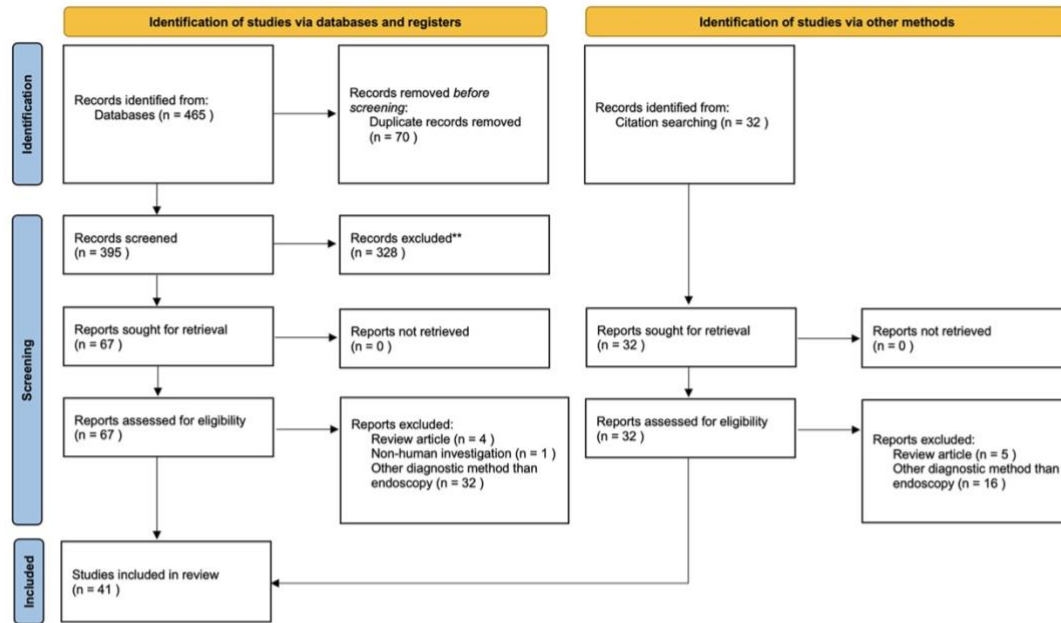
It begins with a systematic review of the literature, mapping the current technological landscape, identifying the deep learning approaches most frequently adopted, and highlighting persistent methodological and clinical gaps. These insights were then used to define realistic and clinically relevant research objectives.

SYSTEMATIC REVIEW OF THE LITERATURE

A review of the literature was carried out in September 2022 following the PRISMA guideline²². Relevant articles were identified by searching through PubMed, Google Scholar, Cochrane, and Embase databases. MeSH terms— including UADT districts, AI techniques, and endoscopic examinations were used. Pertinent articles identified in the references after full-text reading were also included for comprehensiveness. Inclusion criteria: 1) Application of AI-based models (machine learning and deep learning) to endoscopy of the UADT; 2) Reported outcome measures of the algorithm performance. Exclusion criteria were: 1) Use of other diagnostic methods rather than endoscopy; 3) Non-standard endoscopy techniques (i.e. high-speed endoscopy, contact endoscopy, etc.); 2) Review articles, systematic reviews, and letters to the editor; 3) Non-human investigations; 4) Papers published in non-peer reviewed journals; 5) Other languages than English. No time-frame constraint was adopted.

The search results and the selection steps are reported in the flow chart **Figure 14**.

After the selection steps, 41 studies were included in the review. The relevant information of the included articles such as date of publication, first author, AI algorithm, anatomical site of application, AI-task investigated, main features, and outcomes are displayed in **Table 1**.



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

Figure 14. The systematic review flowchart reported according to the PRISMA guidelines. search strings across the selected scientific literature databases

Table 1. Summary of the included studies organized by computer vision task. CNN, convoluted neural network; AUC, area under the curve; ML, machine learning; DL, deep learning; RoI, region of interest; NBI, narrow band imaging; WL, white light; SCC, squamous cell carcinoma.

First author	Title	Task	Anatomical site	AI model	Notes	Outcomes
Hanif (2021) ²³	Upper Airway Classification in Sleep Endoscopy Examinations using Convolutional Recurrent Neural Networks	Classification	Upper airway tract	ResNet18	Classification of upper airway regions using 24 endoscopy videos	Accuracy = 0.82 ; F1-score = 0.79
Cho (2021) ²⁴	Diagnostic Accuracies of Laryngeal Diseases Using a Convolutional Neural Network-Based Image Classification System	Classification	Larynx	EfficientNet, InceptionV3, MobileNetV2, VGG16	Classification of laryngeal benign lesions (cysts, granulomas, nodules, normal cases, palsies, papillomas, and polyps) using 4,106 laryngoscopic images. Performance comparison with junior trainees. Activation maps were implemented to spot the relevant pathologic regions on the images	F1-score = 0.71-0.94; AUC = 0.97-1
Xiong (2019) ²⁵	Computer-aided diagnosis of laryngeal cancer via deep learning based on laryngoscopic images	Classification	Larynx	InceptionV3	Classification of laryngeal lesions (cancer, precancerous, benign, normal) on 14,897 images. External validation cohort. Performance comparison with human experts. Heat maps were implemented to localize the lesion	2-class classification: Sensitivity = 0.72-0.73; Specificity = 0.92-0.94; Accuracy = 0.86-0.89; AUC = 0.92. 4-class classification: Accuracy = 0.74-0.77
Ay (2022) ²⁶	Automated classification of nasal polyps in endoscopy video-frames using handcrafted and CNN features	Classification	Nasal cavity	features extraction + ML/CNN-based classification	Classification of nasal polyps on 4,640 images. Both handcrafted and CNN-based features extraction, as well as ML and DL-based classifiers	Accuracy = 0.80-0.97; Recall = 0.72-0.99; Precision = 0.76-0.99; Inference time = 0.210s
Ren (2020) ²⁷	Automatic Recognition of Laryngoscopic Images Using a Deep-Learning Technique	Classification	Larynx	ResNet101	Classification of laryngeal neoplasms on 24,667 laryngoscopy images. Comparison with 12 otolaryngologists. Activation maps were implemented to spot the relevant pathologic regions on the images	Accuracy = 0.96; Sensitivity = 0.89-0.99; Specificity = 0.99; Inference time = 0.045s
Girdler (2021) ²⁸	Feasibility of a deep learning-based algorithm for automated detection and classification of nasal polyps and inverted papillomas on nasal endoscopic images	Classification	Nasal cavity	ResNet152	Classification of 297 endoscopic images into normal, nasal polyps and inverted papillomas. Performance comparison with 3 experts and 3 residents. Activation maps were implemented to spot the relevant pathologic regions on the images	Accuracy = 0.74; Sensitivity = 0.69-0.78; Specificity = 0.82-0.93; F1-score = 0.82-0.907
Cho (2020) ²⁹	Comparison of Convolutional Neural Network Models for Determination of Vocal Fold Normality in Laryngoscopic Images	Classification	Larynx	CNN with 6 layers, VGG16, InceptionV3, Xception	Classification of normal and abnormal laryngeal images (2,216 vocal fold images (RoI-based cropped)). Comparison with the usefulness of Haar cascade classifiers. Activation maps were implemented to spot the relevant pathologic regions on the images	Sensitivity = 0.82-1.00; Specificity = 0.82-0.99; Accuracy = 0.82-1.00; AUC = 0.90-1.00; Inference time = 0.001-0.022s

Dunham (2022) ³⁰	Optical Biopsy: Automated Classification of Airway Endoscopic Findings Using a Convolutional Neural Network	Classification	Larynx	VGG16	2-class (benign, premalignant, or malignant) and 5-class (normal, nodule, papilloma, polyp, web) classification of vocal folds from 19,353 images. Transfer learning strategy. Structural similarity was used to select different images from the same video	Binary classifier: Accuracy = 0.93; Recall = 0.92-0.94; Precision = 0.92-0.94; 5-class classifier: Accuracy = 0.81; Recall = 0.70-0.88; Precision = 0.72-0.90
Witt (2014) ³¹	Detection of chronic laryngitis due to laryngopharyngeal reflux using color and texture analysis of laryngoscopic images	Classification	Larynx and hypopharynx	Hue and texture feature extraction + Multilayer Perceptron	Classification of laryngopharyngeal reflux from laryngoscopic images of 62 subjects. Results comparison across different regions, between hue-only and texture-only features and between different network depth values.	Accuracy = 0.80; AUC = 0.89
Mascharak (2018) ³²	Detecting oropharyngeal carcinoma using multispectral, narrow-band imaging and machine learning	Classification	Oropharynx	Color and texture feature extraction + naïve bayesian classifier	Identification of oropharyngeal carcinoma on NBI and WL images of 30 patients (20 with biopsy-proven, 10 healthy. Results comparison with color-only features.	Accuracy = 0.52-0.70; Sensitivity = 0.47-0.71; Specificity = 0.58-0.69; AUC = 0.51-0.92
Heo (2022) ³³	Deep learning model for tongue cancer diagnosis using endoscopic images	Classification	Oral cavity	CNN, VGG, DenseNet, MobileNet, ResNet, and EfficientNetB3	Classification of oral cancer using 5,576 malignant and non-malignant endoscopic images. Performance comparison with oncologists and general physicians.	Accuracy = 0.64-0.84; Sensitivity = 0.43-0.79; Specificity = 0.56-0.89; Precision = 0.52-0.79; F1-score = 0.57-0.78; AUC = 0.71-0.89
He (2021) ³⁴	A deep convolutional neural network-based method for laryngeal squamous cell carcinoma diagnosis	Classification	Larynx	InceptionV3	Laryngeal SCC diagnosis on 4,591 NBI scans. Independent test cohort from other institutions. Performance comparison with human experts	Accuracy = 0.91; Sensitivity = 0.90; Specificity = 0.91; AUC = 0.96; Inference time = 0.010s
Zhao (2022) ³⁵	Vocal cord lesions classification based on deep convolutional neural network and transfer learning	Classification	Larynx	MobileNetV2	4-class (normal, polyp, keratinization, and carcinoma) and 2-class (urgent, non-urgent) classification of 5,122 laryngoscopic images. Performance comparison with human experts	4-class: Accuracy = 0.80; F1-score = 0.78; AUC = 0.96; 2-class: Accuracy = 0.98; Sensitivity = 0.89; Specificity = 0.99; AUC = 0.98; Inference time = 0.011s
Xu (2021) ³⁶	Deep Learning for nasopharyngeal Carcinoma Identification Using Both White Light and Narrow-Band Imaging Endoscopy	Classification	Nasopharynx	Siamese DCNN	Nasopharyngeal cancer classification on 4,783 endoscopic images. Pairs of NBI and WL images from the same patient were used as input to 2 different Xception-based encoders. Evaluation both at an image-level and patient-level. Activation maps were implemented to spot the relevant pathologic regions on the images	Image-level: Accuracy = 0.95; Sensitivity = 0.96; Specificity = 0.95; AUC = 0.99; Patient-level: Accuracy = 0.96; Sensitivity = 0.97; Specificity = 0.94; Inference time = 0.039s
Ye (2021) ³⁷	Deep Learning for Laryngopharyngeal Reflux Diagnosis	Classification	Larynx and hypopharynx	ResNet18, ResNet50, ResNet101	Diagnosis of nasopharyngeal reflux disorder from endoscopic images of 114 subjects. Attempts to include attention modules. ImageNet-based transfer learning strategies.	Accuracy = 0.59-0.74; AUC = 0.68-0.76

Yin (2021) ³⁸	Laryngoscope8: Laryngeal image dataset and classification of laryngeal disease based on attention mechanism	Classification	Larynx	DenseNet121 to extract heatmaps, Faster-RCNN to localize the critical area, and DenseNet121 to classify	Attention mechanism to obtain and crop interesting local areas where the classification result is correct. 8-class classification (Reinke's edema, glottic cancer, granuloma, normal, vocal cord cyst, vocal cord nodule, vocal cord polyp) of 3,057 images, collected with 2 different devices.	Accuracy = 0.73; AUC = 0.89
Wang (2022) ³⁹	LARNet-STC: Spatio-temporal orthogonal region selection network for laryngeal closure detection in endoscopy videos	Classification	Larynx	Custom made (LARNet-STC)	Detection of laryngeal adductor reflex events in terms of 9,102 laryngoscopic frames classification in open vocal cords, closed vocal cords, and occlusions. Activation maps were implemented to spot the relevant pathologic regions on the images	Recall = 0.96; Precision = 0.90; F1-score = 0.93
Araujo (2019) ⁴⁰	Learned and handcrafted features for early-stage laryngeal SCC diagnosis	Classification	Larynx	Handcrafted features extraction vs. CNN-learned features obtained with ResNet and Inception + ML classifier SVM	Evaluation if CNN-learned features are better than handcraft features for classification of malignant laryngeal tissue using the dataset of Moccia et. al 2017	Recall = 0.98; Precision = 0.98; F1-score = 0.98; Inference time = 41s
Turkmen (2015) ⁴¹	Classification of laryngeal disorders based on shape and vascular defects of vocal folds	Classification	Larynx	Handcrafted features extraction + ML classifier	A concatenation of tasks: segmentation of the vessels, handcrafted feature extraction, and finally classification for recognizing laryngeal benign lesions (5 classes)	Sensitivity= 0.73-0.94; Specificity= 0.89-0.99; Precision= 0.69-0.94.
Moccia (2017) ⁴²	Confident texture-based laryngeal tissue classification for early stage diagnosis support.	Classification	Larynx	Handcrafted features extraction + ML classifier	Evaluate if ML techniques can classify laryngeal tissues in NBI images by exploiting textural information from extracted tissues' patches (n=1320). Original image dataset N=330	Recall = 0.93; Precision = 0.94; F1-score = 0.92.
Barbalata (2016) ⁴³	Laryngeal Tumor Detection and Classification in Endoscopic Video	Classification	Larynx	Linear Discriminant Analysis and Matched Filtering	A concatenation of tasks /segmentation of the region of interest, segmentation of the vessels, and finally classification for recognizing laryngeal cancer on NBI frames (120 images dataset)	Region of interest segmentation (DSC = 0.69); vessel segmentation (DSC = 0.76); classification (accuracy = 0.84)
Patrini (2020) ⁴⁴	Transfer learning for informative-frame selection in laryngoscopic videos through learned features	Classification (informative)	Larynx	VGG16, InceptionV4, ResNetV, Inception-ResNetV2	CNN models to extract features and classify 720 NBI frames in informative, blurred, saliva, or specular reflections and underexposed. Pre-training on ImageNet dataset.	Recall = 0.89-0.94; Precision = 0.90-0.94; F1-score = 0.89-0.94; AUC=0.98-0.99
Moccia (2018) ⁴⁵	Learning-based classification of informative laryngoscopic frames	Classification (informative)	Larynx	Feature extraction + SVM machine learning classifier	Classification of 720 NBI videoframes in informative, blurred, saliva or specular reflections and underexposed frames.	Recall = 0.84; Precision = 0.77; F1-score = 0.80; AUC= 0.91; Inference time = 0.03s

Yao (2022) ⁴⁶	Automatic classification of informative laryngoscopic images using deep learning	Classification (informative)	Larynx	ResNet18	Classification of 22,132 laryngoscopic frames (healthy and with vocal fold polys patients) into informative or uninformative. Pre-training on ImageNet dataset. Activation maps were implemented to spot the relevant pathologic regions on the images	Recall = 0.76-0.90; Precision = 0.63-0.94; F1-score = 0.69-0.92; Inference time = 0.032s
Galdran (2019) ⁴⁷	Real-time informative laryngoscopic frame classification with pre-trained convolutional neural networks	Classification (informative)	Larynx	SqueezeNet	Classification of SCC NBI videoframes into informative, blurry, saliva/secularities, and underexposed frames. Pre-trained on ImageNet dataset.	Recall = 0.94; Precision = 0.95; F1-score = 0.95; Inference time = 0.004s
Kim (2021) ⁴⁸	Automated laryngeal mass detection algorithm for home-based self-screening test based on convolutional neural network	instance segmentation	Larynx	Mask RCNN	Detection of laryngeal masses and vocal folds (dataset= 1224 images)	For recognizing laryngeal masses: recall = 0.80, precision = 0.73, accuracy=0.62, F1=0.76
Matava (2020) ⁴⁹	A Convolutional Neural Network for Real Time Classification, Identification, and Labelling of Vocal Cord	Object detection	Larynx and trachea	MobileNetV1, Inception, ResNet	Identification and classification of vocal cords or tracheal rings on 457 bronchoscopy and laryngoscopy images	Sensitivity = 0.56-0.86; Specificity = 0.85-0.97; Inference time = 0.200-0.067s
Azam (2022) ⁵⁰	Deep Learning Applied to White Light and Narrow Band Imaging Videolaryngoscopy: Toward Real-Time Laryngeal Cancer Detection	Object detection	Larynx	YOLOV5	Laryngeal SCC detection on 657 frames extracted from both transnasal and intraoperative endoscopies (WL and NBI). Model tested on six videolaryngoscopies with good inference time	mean Average Precision (mAP) = 0.50-0.63; Recall = 0.56-0.62; Precision = 0.45-0.68 TP rate = 82.0%; Inference time = 0.026s
Inaba (2020) ⁵¹	Artificial intelligence system for detecting superficial laryngopharyngeal cancer with high efficiency of deep learning	Object detection	Larynx and hypopharynx	RetinaNet	Superficial laryngopharyngeal cancer detection in 1,200 NBI mucosal images.	Accuracy = 0.97; Sensitivity = 0.95; Specificity = 0.98
Tamashiro (2020) ⁵²	Artificial intelligence-based detection of pharyngeal cancer using convolutional neural networks	Object detection	Pharynx	Single Shot MultiBox Detector	Pharyngeal cancer detection model trained on 5,403 NBI and WL images. Independent validation set of 1,912 with and without cancer images.	Accuracy = 0.67; Sensitivity = 0.80; Specificity = 0.57; Inference time = 0.020s
Adamian (2021) ⁵³	An Open-Source Computer Vision Tool for Automated Vocal Fold Tracking From Videoendoscopy	Segmentation	Larynx	DeepLabCut (modified ResNet with deconvolutional layers for keypoint localization)	Vocal folds' anterior commissure key points localization to estimate the glottal opening angle and assess vocal fold paralysis	Pearson correlation coefficient (estimated angles vs. expert measured ones) = 0.97; AUC = 0.88 in estimating vocal fold paralysis
Parker (2021) ⁵⁴	Machine Learning in Laryngoscopy Analysis: A Proof of Concept Observational Study for the Identification of Post-Extubation Ulcerations and Granulomas	Segmentation	Larynx	Unet	Segmentation of ulcerations and granulomas on 127 laryngoscopic frames	Sensitivity = 0.63-0.82; Specificity = 0.99; AUC = 0.89-0.99

Paderno (2021) ⁵⁵	Deep Learning for Automatic Segmentation of Oral and Oropharyngeal Cancer Using Narrow Band Imaging: Preliminary Experience in a Clinical Perspective	Segmentation	Oral cavity and oropharynx	Unet, Unet3, ResNet	Segmentation of SCC on 110 oral cavity and 116 oropharynx NBI frames	Oral cavity: DSC = 0.59-0.66; Accuracy = 0.86-0.89; Recall = 0.67-0.75; Precision = 0.59-0.67; Oropharynx: DSC = 0.69-0.76; Accuracy = 0.78-0.84; Recall = 0.78-0.86; Precision = 0.62-0.80; Inference time = 0.059s-0.115s
Lin (2019) ⁵⁶	Quantification and Analysis of Laryngeal Closure from Endoscopic Videos	Segmentation	Larynx	Custom-made: Encoder (for detection of the region of interest) + fully convolutional network (for segmentation)	Real-time segmentation of glottic opening, vocal folds and. The dataset consists of 806 training and 400 testing frames. Temporal features like glottic opening angle and supraglottic index were also extracted to quantify tissue movements and assess the exercise-induced laryngeal obstruction.	IoU = 0.65-0.85; Inference time = 0.055-0.060s
Azam (2022) ⁵⁷	Videomics of the Upper Aero-Digestive Tract Cancer: Deep Learning Applied to White Light and Narrow Band Imaging for Automatic Segmentation of Endoscopic Images	Segmentation	Larynx	Custom-made (SegMENT)	Laryngeal SCC segmentation of 683 NBI and WL frames. Pre-training on ImageNet dataset. 2 external cohorts of 116 oropharyngeal SCC and 102 oral cavity SCC images were used to validate the model. Ensemble strategy and transfer learning were used to decrease errors considering also UNet and DenseNet121 networks.	DSC = 0.60-0.88; Accuracy = 0.93-0.97; IoU = 0.68-0.78; Recall = 0.90-0.95; Precision = 0.78-0.92; Inference time = 0.256s
Li (2018) ⁵⁸	Development and validation of an endoscopic images-based deep learning model for detection with nasopharyngeal malignancies	Segmentation	Nasopharynx	Inception-based fully convolutional network	Nasopharyngeal malignancy segmentation model on 27,536 biopsy-proven endoscopic images. Performance comparison with oncologist evaluations	DSC = 0.78; Accuracy = 0.89; Sensitivity = 0.91; Specificity = 0.83; AUC = 0.94; Inference time = 0.028s
Laves (2019) ⁵⁹	A dataset of laryngeal endoscopic images with comparative study on convolution neural network-based semantic segmentation	Segmentation	Larynx	UNet, SegNet, Enet, ErfNet	Segmentation of anatomical structures from a vocal fold dataset (536 images labeled as void, vocal folds, other tissue, glottal space, pathology, surgical tool, and intubation). Comparison between training from scratch, pre-training and fine-tuning, data augmentation, and ensemble configurations.	IoU = 0.54-0.84; Inference time = 0.048-0.008s
Pan (2022) ⁶⁰	RANT: A cascade reverse attention segmentation framework with hybrid transformer for laryngeal endoscope images	Segmentation	Larynx	Custom-made (RANT)	Segmentation of larynx structures (open vocal folds open, closed vocal folds, epiglottis, glottal space, pyriform fossa, other tissue) from 822 laryngoscopic images and (vocal folds, other tissue, glottal space, lesion, surgical tool, intubation, void) from 536 laryngoscopic images.	DSC = 0.83-0.93; IoU = 0.77-0.89; Inference time = 0.004s
Ji (2020) ⁶¹	A multi-scale recurrent fully convolution neural network for laryngeal leukoplakia segmentation	Segmentation	Larynx	Custom-made (Boldface Multi-scale Net)	Segmentation of leukoplakia on 649 laryngoscopic images.	Accuracy = 0.99; Recall = 0.89; Precision = 0.74; F1-score = 0.78; IoU = 0.83; Inference time = 0.205s

Hamad (2019) ⁶²	Automated Segmentation of the Vocal Folds in Laryngeal Endoscopy Videos using Deep Convolutional Regression Networks	Segmentation	Larynx	13-layers fully convolutional regression network	Segmentation of the vocal fold and the glottal region from 7,892 endoscopic frames.	DSC = 0.86; Accuracy = 0.95; IoU = 0.88;
Ding (2022) ⁶³	Automatic glottis segmentation for laryngeal endoscopic images based on U-Net	Segmentation	Larynx	Custom-made (CN-DA-Unet)	Segmentation of the glottal space on 928 laryngoscopic images.	DSC = 0.93; Sensitivity = 0.93; Specificity = 0.93

From the systematic review, it can be seen how AI applied to the endoscopy of UADT is still an emerging field, but on the other hand, the interest of the scientific community in the topic is continuously growing, with the number of published papers exponentially increasing in the last years, paving the way towards AI-based clinical tools. Nowadays, the state of the art of AI applied to UADT endoscopy is represented by DL models that are leveraged for achieving three previously mentioned main tasks: classification, detection, and segmentation.

Classification

AI has shown considerable potential for assisting clinicians in classifying laryngeal and UADT findings.

The first task that was investigated at the origin of AI application to laryngoscopy was the automatic evaluation of image quality during endoscopy, helping to exclude frames degraded by blurring, secretions, or poor illumination and enabling the selection and extraction of informative images from the ones hampered by examination artifacts. Several ML and DL approaches, such as those developed by Moccia et al., Patrini et al., and Galdran et al., have achieved high performance in classifying informative versus non-informative frames, with F1-scores above 0.9 and inference times suitable for live application^{44,45,47}. More recently, Yao et al. validated a ResNet-based classifier on over 22,000 frames, obtaining precision and recall values exceeding 90%⁴⁶. These tools may support physicians in performing cleaner examinations and in automatically documenting procedural quality.

Another major classification objective is the “optical biopsy,” in which DL models discriminate between benign and malignant lesions. Despite being challenging, this task has the potential to create computer-assisted diagnosis technologies able to support clinicians in deciding the optimal patient’s therapeutic pathway. Early approaches with traditional ML methods based on handcrafted texture and vascular features achieved modest accuracy^{32,40–43}. This approach was limited by the number and the quality of the features selected by the investigators, and it was less efficient compared to modern DL models. Current CNN-based models have been trained on large datasets to classify oral, nasopharyngeal, and laryngeal cancers, with performance compatible with experts. Regarding this task, Zhao et al., Xiong et al., and Dunham et al. demonstrated that binary classification between benign and malignant lesions yields higher accuracy than multiclass schemes, which remain limited by dataset heterogeneity^{25,30,35}. Intuitively, as it happens for humans, the more complicated is the prediction task is, the lower the accuracy. Ren et al. developed one of the largest laryngeal image datasets to date and reported diagnostic accuracies above 95%, outperforming human experts in distinguishing between nodules, polyps, leukoplakia, and carcinoma²⁷. Despite promising metrics, most studies lack external validation or video-based clinical deployment, and the “black box” nature of DL models continues to hinder their interpretability, although visualization tools such as Grad-CAM are beginning to address this issue^{36,64}.

Detection

Detection tasks focus on automatically localizing suspicious lesions within endoscopic frames. Compared to the classification task, the detection of lesions in UADT endoscopic images has been the focus of a small number of investigations so far. A few studies have applied DL architectures to identify mucosal abnormalities in the pharynx and larynx. Inaba et al. trained a model on NBI images achieving 0.97 accuracy and 0.98 AUC for laryngopharyngeal cancer detection⁵¹. Tamashiro et al. combined white-light and NBI data to detect pharyngeal cancers with 80% sensitivity and real-time

processing speed of 0.02 s per frame⁵². For the larynx specifically, Our group first developed a YOLO-based model capable of real-time cancer detection (0.026 s per frame), reaching a mean average precision (mAP@0.5) of 0.63 and true positive rate of 0.82 on both white-light and NBI frames⁵⁰. Even though yielding non-excellent results, especially with small lesions, this work demonstrated that this approach was promising for real-time laryngeal cancer automatic detection. All together, these studies demonstrate the feasibility of embedding DL detection systems into endoscopic workflows to assist in highlighting regions of concern during examinations.

Segmentation

Segmentation and localization models extend detection by outlining lesion boundaries, which has potential applications for biopsy guidance and surgical margin definition. Several CNN architectures, including U-Net, DeepLab, and Mask R-CNN, have been evaluated for this purpose. Li et al. achieved a Dice similarity coefficient (DSC) of 0.78 for nasopharyngeal cancer segmentation, while Paderno et al. reported DSC = 0.76 for oropharyngeal squamous cell carcinoma using NBI images^{55,58}. Our group recently introduced the SegMENT model for laryngeal cancer delineation on white-light and NBI endoscopic images, obtaining DSC = 0.81 and validating it also on datasets of oral and oropharyngeal cancers with comparable performance⁵⁷. These results suggest that a single well-trained DL model could generalize across UADT sites. In the operating room, accurate lesion segmentation could help define resection margins and improve intraoperative decision-making. Additionally, segmentation has been applied to anatomical structures such as the vocal folds and glottic space to quantify motion or airway patency⁵³. While current models show near real-time inference speeds (25–30 FPS), clinical validation remains the essential next step for their translation into practice. Considering the complexity of the clinical trials needed to validate a surgical margin delineation model, segmentation system are expected to be the last to arrive in clinical practice.

Conclusion

Overall, AI is demonstrating strong potential to enhance diagnostic accuracy and standardize clinical practice in UADT endoscopy. Classification models can assist in selecting diagnostically relevant frames and differentiating between benign and malignant lesions, supporting objective and reproducible interpretation. Detection algorithms have achieved real-time identification of suspicious areas, paving the way for computer-assisted endoscopic examinations that highlight lesions as the physician navigates the airway. Segmentation models can delineate tumor margins with precision comparable to that of trained otolaryngologists, suggesting a future role in surgical planning and intraoperative margin assessment. Despite these promising results, the clinical impact of AI remains limited by small, single-center datasets, a lack of external validation, and insufficient real-time prospective studies. Future research should focus on multicentric collaborations, integration into clinical endoscopy systems, and standardized reporting of performance metrics to ensure robust, transparent, and generalizable applications in routine otolaryngology practice.

PURPOSE OF THE THESIS

From this review, AI applied to UADT endoscopy emerges as a growing field that is rapidly progressing and approaching performance levels appropriate for clinical use. Nevertheless, only preclinical studies were reported so far in the literature, with single centers pursuing partial tasks with limited databases, often without external testing and clinical validation.

Building on this foundation, for this project, a large multicenter international dataset of endoscopic images and videos was assembled and manually annotated by expert clinicians, ensuring high-quality ground truth for model development. Using this dataset, several key computer vision tasks were methodically investigated:

- First, algorithms for frame quality assessment and automatic selection of informative frames were developed, aimed at reducing redundancy in videos and enhancing the efficiency of downstream analyses
- Second, real-time detection models were implemented to localize laryngeal lesions and tumors, supporting earlier recognition of suspicious areas during routine examinations
- Third, classification systems were trained to predict tumor histology (high risk vs. low risk) from endoscopic appearances, bridging the gap between visual diagnosis and histopathological confirmation
- Fourth, semantic segmentation approaches were applied to delineate tumor boundaries, with the goal of supporting surgical planning and intraoperative margin assessment.
- Fifth, motion assessment strategies were used to automatically distinguish normal laryngeal motility from laryngeal paralysis cases.

Each task was evaluated not only on internal data but also validated externally on independent cohorts, with additional assessment by clinical experts. This multi-step approach ensured that model performance was evaluated not only in technical terms but also with regard to clinical applicability.

Through this multi-step framework, the thesis aims to advance the translation of artificial intelligence into laryngeal endoscopy, ensuring that these models are critically benchmarked for robustness and reproducibility, with relevant benefits in clinical practice.

MATERIALS AND METHODS

STUDY DESIGN AND OVERVIEW

This thesis followed a multi-step design combining evidence synthesis and experimental development of artificial intelligence tools for laryngeal endoscopy. A systematic review was first conducted to assess the current evidence base and define unmet needs. A large multicenter image and video repository was then assembled and annotated by expert otolaryngologists. From this resource, task-specific datasets were derived and used to train, validate, and externally test models for five distinct CV-automatized tasks: (1) informative frame selection, (2) lesion detection, (3) histology prediction, (4) semantic segmentation, and (5) laryngeal motility assessment.

Dataset collection and framework

A multicenter, retrospective dataset of endoscopic images and videos was assembled to support the experimental phases of this thesis. Data were collected from several tertiary referral hospitals and academic centers across Europe and Asia, ensuring a heterogeneous patient population and broad representation of endoscopic practices. Ethical approval was obtained from all participating institutions (e.g., CER Liguria: 169/2022), and data collection complied with the Declaration of Helsinki. Participating institutions included:

- Unit of Otolaryngology – Head and Neck Surgery of the Hospital San Martino in Genova, Italy.
- Service of Otolaryngology – Head and Neck Surgery of the Hospital Clínic in Barcelona, Spain.
- Department of Otolaryngology – Head and Neck Surgery of the Hospital Severance in Seoul, Republic of Korea.
- Unit of Laryngology of the Sushrut Institute of Plastic Surgery & Super Specialty Hospital in Lucknow, India.

- Unit of Otolaryngology – Head and Neck Surgery of the Department of Neuroscience, Azienda Ospedale Università di Padova, Italy.
- Unit of Otolaryngology – Head and Neck Surgery of the University of Brescia, Italy
- Unit of Otolaryngology of the Oncologic Hospital Saint Savvas in Athens, Greece.

At all centers, endoscopic examinations were performed using both white light (WL) and NBI modalities. Examinations were conducted either in-office with flexible nasopharyngolaryngoscopes or intraoperatively with rigid telescopes, depending on clinical workflow at each site. The endoscopic videos were acquired either during outpatient evaluations using high-definition flexible nasopharyngolaryngoscopes (High-Definition Video Rhino-laryngoscope Olympus ENF-VH – Olympus Medical System Corporation, Tokyo, Japan) by transnasal approach, or during intraoperative examinations using 0°, 30°, and 70° endoscopes equipped with high-definition cameras.

From the recorded video-endoscopies, image frames were retrospectively extracted and annotated by expert otolaryngologists. Depending on the experimental objective, annotations included image quality classification, lesion presence and location, tumor boundaries, or laryngeal motility classification. Each time, annotations were independently reviewed by at least two experts, and disagreements were resolved by consensus. Depending on the specific task, datasets from one institution could serve as internal training/testing sets and from another as independent external validation sets. For dividing the datasets into training, validation, and test sets, a patient-wise split methodology was used in order to reduce overfitting probabilities.

INFORMATIVE FRAME SELECTION

Purpose

Automatic identification of diagnostically informative frames within videolaryngoscopies is essential to reduce redundancy, improve efficiency in reviewing examinations, and provide high-quality inputs for subsequent AI models such as detection, diagnosis, or segmentation. In clinical settings, many frames are blurred, obstructed, or poorly illuminated, limiting their diagnostic value. This experiment aimed to develop and validate DL models capable of classifying frames as *informative* or *non-informative*, with particular emphasis on generalizability across centers and feasibility for real-time use during laryngeal endoscopy.

Dataset

Two datasets were used for this experiment. First, the publicly available NBI-InfFrames dataset from Moccia et al. was included⁴⁵. It consists of 720 NBI frames from 18 patients with laryngeal squamous cell carcinoma, categorized into four groups: informative (n = 180), dark (n = 180), blurred (n = 180), and artifact-contaminated (n = 180). Images were acquired with an Olympus Visera Elite S190 system and an ENF-VH video rhino-laryngoscope.

To extend beyond carcinoma cases and NBI-only images, a new Frame Quality Dataset was assembled from 51 videolaryngoscopies performed at the San Martino Hospital in Genoa (Italy) and the Hospital Clínic de Barcelona (Spain). Patients included both healthy subjects and individuals with diverse laryngeal conditions (carcinoma, papillomatosis, polyps, Reinke's edema, granulomas, and nodules). At both centers, examinations were conducted in-office using a flexible transnasal high-definition video nasopharyngolaryngoscope with both WL and NBI modalities, performed sequentially.

Frame extraction was performed with VideoLAN VLC v3.0.18 software. Frames were labeled as *informative* if they provided a clear and close view of the larynx or its subsites without blur, saliva, glare, or underexposure. All other frames were labeled as *non-informative*. At the San Martino hospital, three junior otolaryngologists performed independent labeling with final approval by a senior laryngologist. At the Hospital Clínic de Barcelona, a similar process was followed with independent labeling by one junior otolaryngologist and final review by a senior specialist.

Model development

Three CNNs were compared:

1. A lightweight shallow CNN with three convolutional blocks and <0.3 million parameters.
2. MobileNetV2, optimized for low-resource environments with depthwise separable convolutions (2.4 million parameters).
3. ResNet-50, a deeper residual architecture with skip connections (23.8 million parameters).

All models were adapted for binary classification between informative and non-informative.

A self-supervised pretext task based on image rotation prediction was used for initialization instead of standard ImageNet weights. From 232 sampled frames, 512×512 patches were cropped and rotated by 0° , 90° , 180° , or 270° . Networks were pretrained for 200 epochs to classify rotation angles (Adam optimizer, learning rate = 0.0001, batch size = 8). The resulting weights were then used to initialize the models for the real classification task.

Evaluation metrics

Performance was evaluated by comparing predicted labels against ground truth, creating a confusion matrix. True positives (TPs) represent the number of frames accurately predicted as belonging to the informative class, while true negatives (TNs) account for the number of uninformative frames

correctly predicted. The false negatives (FNs) correspond to informative frames not identified by the model, and conversely, false positives (FPs) refer to frames wrongly categorized as part of that class. *Precision* (also known as positive predictive value) measures the percentage of objects that are correctly recognized among all the predicted outputs.

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

Recall (also known as sensitivity) is computed as the fraction of the objects that are correctly recognized, compared to all the ones that should have been recognized.

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

The simultaneous evaluation of both these metrics is paramount to deciding whether the model is appropriate for the clinical needs. The *F1 score* is a way of looking at both parameters at the same time to assess the model's quality more comprehensively. The *F1 score* is defined as the harmonic mean of *Precision* and *Recall* as the relative contribution of both these metrics is equal. It is higher when both measures are similarly high.

$$F1score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN}$$

The Receiver Operator Characteristic (ROC) curve is a probability curve that plots the TP rate (recall) against the FP rate. The more the curve approaches the upper left corner the higher is the TP rate, the lower is the FP rate, and the wider will be the Area Under the Curve (AUC). In fact, the AUC of the model provides an aggregate measure of performance. For instance, in a classifier, an AUC equal to 1 corresponds to a classifier able to perfectly distinguish between all positive and negative cases. When the AUC is equal to 0, the model predicts all negatives as positives and vice versa. When AUC is 0.5, the classifier is not able to distinguish between the Positive and Negative classes.

Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to visualize salient image regions contributing to model predictions.

Video testing

To assess real-time performance, four unedited preoperative videolaryngoscopies not included in the training datasets were analyzed. Frame-level predictions were displayed with a colored border (green = informative, red = uninformative). The Video Informativeness Level (VIL), defined as the fraction of frames classified as informative, was calculated for each video.

$$VIL = \frac{\text{number of informative frames}}{\text{total number of frames}}$$

The ResNet-50 model (with pretext initialization) was deployed in ONNX format, achieving an inference time of <0.020 s per frame on an NVIDIA RTX A6000 GPU.

Statistical analysis

Categorical variables were reported as frequencies and percentages, continuous variables as medians with interquartile ranges. The McNemar test was used to compare the three CNNs on the internal test set. ROC curves and AUC were compared between the internal and external test sets using Z-tests. A two-sided $p < 0.05$ was considered significant for both tests. Analyses were performed in R studio version 2024.09.1+394.

LARYNGEAL TUMORS DETECTION

Purpose

Accurate and early detection of laryngeal lesions is critical to improving patient outcomes, however visual recognition can be challenging, particularly for small or flat lesions. This experiment aimed to develop and validate an artificial intelligence–based system capable of automatically detecting benign and malignant laryngeal lesions from both WL and NBI endoscopic images. The goal was to design a computationally efficient model that could operate in real time and support clinicians during endoscopic assessment, even in resource-limited settings.

Dataset

The study employed a multi-center dataset comprising three internal datasets and one external validation set:

1. ENDO-LC1 — collected at the IRCCS Ospedale Policlinico San Martino, University of Genoa (Italy). It included 2037 high-definition WL and NBI images. Data included malignant lesions, i.e., Squamous Carcinoma, and benign lesions such as cysts, granuloma, leukoplakia, papilloma, polyps, and Reinke’s edema. Acquisitions were performed as previously explained during outpatient assessments using a flexible nasopharyngolaryngoscope and during intraoperative examinations using 0°, 30°, or 70° rigid telescopes.
2. ENDO-LC2 — collected at the ENT Center, Voice & Airway Clinic, Sushrut Institute of Plastic Surgery and Super Specialty Hospital, Lucknow (India). This set comprised 133 images (70 WL and 63 NBI) from both healthy and pathological cases, acquired using a flexible nasopharyngolaryngoscope during in-office examinations.
3. Laryngoscope8 — a publicly available dataset from the Sixth Medical Center of PLA General Hospital, Beijing (China)³⁸. A subset of 1700 WL images was selected to balance classes,

representing both normal and pathological conditions (glottic cancer, Reinke’s edema, granuloma, leukoplakia, papilloma, polyps, cysts, and nodules). These images were acquired using a flexible nasopharyngolaryngoscope during in-office examinations.

Frames from all datasets were reviewed by an otolaryngologist, who excluded poor-quality images (blurred, overexposed, saliva-contaminated) before including the images in the dataset. Remaining frames were annotated using the CVAT online tool (Intel Corporation, USA) ⁶⁵, with bounding boxes (BBs) drawn around all visible lesions—malignant and benign—grouped under the unified label “lesion.”

After quality control, data were divided with a patient-wise method into training (89%), validation (7%), and test (4%) sets. WL and NBI modalities were mixed across splits to ensure balanced representation.

An external dataset (ENDO-LC ext) was independently collected at collected at the IRCCS Ospedale Policlinico San Martino, University of Genoa (Italy), including 149 WL and NBI frames from different patients than those in the internal datasets. This dataset was collected so that it contained only small ($<32\times32$ px) and medium-sized (32×32 – 96×96 px) lesions and was used exclusively for external validation.

Model development

The Super-Resolution Enhanced YOLO model was developed to improve detection of small lesions without compromising real-time performance⁶⁶. It was based on the lightweight YOLOv8-nano architecture, designed for high-speed image analysis. During training, an auxiliary *super-resolution* branch reconstructed higher-resolution representations of input frames to enhance the model’s sensitivity to fine details. the original YOLOv8n model was then compared to the the Super-Resolution Enhanced YOLO model for assessing the benefit of this strategy.

Both models were trained using transfer learning with initialization from the COCO dataset⁶⁷, further adapted through task-specific fine-tuning. The COCO dataset (Common Objects in Context) is a large-scale, publicly available image dataset developed for computer vision research. It contains over 200,000 labeled images with more than 80 object categories and detailed pixel-level annotations. Because of its size and diversity, COCO is commonly used for pretraining DL models in object detection, allowing networks to learn general visual features before being fine-tuned on domain-specific medical datasets such as our laryngeal endoscopy datasets. After the pretraining on COCO dataset, the models were fine-tuned on the training partition of the ENDO-LC1, ENDO-LC2, and laryngoscope8 datasets, with data augmentation (spatial transformations, scaling, flipping, and mosaic composition) to enhance generalization.

Evaluation metrics

The output result of a detection model is typically a frame around the object called bounding box (BB). To define a TP for this task, we must introduce the concept of *Intersection over Union* (IoU) of the BBs explained in **Figure 15**. To define a TP a threshold of IoU is selected (we considered an $\text{IoU} > 50\%$ is considered a TP).

The object detection metric which takes into account both *Precision* and *Recall* is the *Average Precision* (AP). The AP is the area under the *Precision-Recall* curve for a single class. Performance was separately reported for small, medium, and large lesions to assess robustness across lesion sizes. Real-time efficiency was assessed by computing frames per second (FPS), comparing SRE-YOLO with standard YOLOv8 and YOLOv5 models.

Especially in real-time applications, it is crucial to know how fast an algorithm can process image frames. Defining “real-time” is not straightforward. Generally, real-time performances are defined if the algorithm process images at least at the rate of the source supplying the frames (i.e. the endoscopic camera). As endoscopy videos usually run at 25-30 *frames-per-second* (FPS), it is generally accepted

that algorithms processing images around 25 FPS are considered real-time. This metric of course depends on the algorithm's complexity but also on the hardware characteristics. A lighter algorithm running on a powerful computer will result in better inference times.

Statistical analysis

Categorical variables were reported as frequencies and percentages, and continuous variables as means with standard deviation. The McNemar test was used to compare the two CNNs on the internal test set. To compare the selected Super-Resolution Enhanced YOLO model on the internal and external dataset, bootstrap resampling over images within each dataset was used to generate a sampling distribution of AP, and then the results were used to assess whether the confidence intervals overlapped or the difference's 95% CI included 0. A two-sided $p < 0.05$ was considered significant for both tests. Analyses were performed in R studio version 2024.09.1+394.

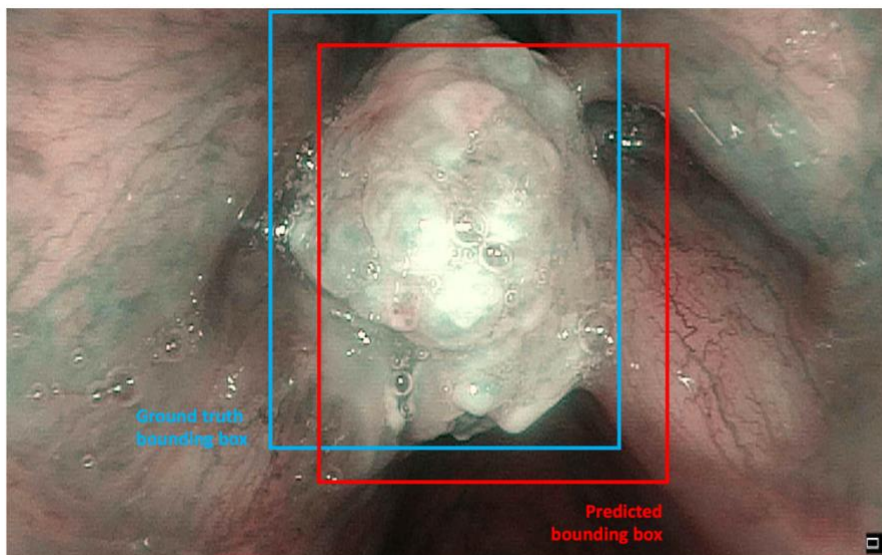


Figure 15. Intersection over union (IoU) explanation

LARYNGEAL TUMORS DIAGNOSIS

Purpose

This experiment aimed to develop and externally validate a computer-aided diagnosis (CADx) model based on artificial intelligence to classify laryngeal lesions from laryngoscopy images into high-risk (HR) or low-risk (LR) categories. The goal was to create a clinically reliable and generalizable system that could distinguish malignant or premalignant lesions (HR) from benign or inflammatory ones (LR), supporting earlier diagnosis and improving workflow efficiency. The performance of the CADx model was compared on an external cohort with that of several categories of physicians for assessing the clinical reliability of the model.

Dataset

A retrospective multicenter study was performed using laryngoscopy images collected between 2014 and 2023 from five tertiary referral centers in Europe and Asia. All centers obtained local ethical approval prior to data collection:

- Hospital San Martino – University of Genoa (Italy): 8,777 images from 889 patients.
- Sushrut Institute of Plastic Surgery & Super Speciality Hospital, Lucknow (India): 1,728 images from 960 patients.
- Hospital Clínic of Barcelona (Spain): 150 images from 150 patients.
- Oncologic Hospital Saint Savvas, Athens (Greece): 909 images from 133 patients.
- Sixth Medical Center of PLA General Hospital, Beijing (China): 3,057 images from 1,950 patients (publicly available *Laryngoscope8* dataset by Yin et al³⁸).

Images were acquired during routine clinical practice using high-definition flexible or rigid Olympus endoscopes, capable of both white light (WL) and narrow-band imaging (NBI). Video frames were

manually extracted and annotated by expert otolaryngologists (≥ 5 years of experience in laryngology and NBI interpretation) at each center.

Each lesion was delineated using a bounding box (BB) encompassing its visible surface and labeled according to final histopathological diagnosis or expert consensus. If multiple lesions were present, multiple BBs were added. If no lesion was present, no BB was added. The lesion label was based on final postoperative histopathological confirmation, except for healthy images. A dedicated internal annotation software was used to allow secure labeling without third-party upload, storing annotations in COCO JSON format. All the images were validated by a second clinician for quality assurance.

Seven lesion types were annotated: *cyst*, *granuloma*, *leukoplakia*, *polyp*, *papilloma*, *Reinke's edema*, and *carcinoma*. For classification purposes, leukoplakia (consisting of keratosis with several degrees of dysplasia) and carcinoma lesions were subsequently grouped as HR lesions, while all the others were grouped as LR lesions. Laryngeal frames characterized by the absence of such lesions were organized into a single category named "Healthy".

An additional Pre-Training dataset of 8,651 frames was automatically extracted from the videos of the endoscopies of the same Italian patient cohort. Frame extraction followed an AI-based frame quality filter using the model developed by our group⁶⁸ to ensure diagnostic clarity. One frame was extracted every ten frames, and images were automatically labeled based on the patient's confirmed diagnosis. These images were used as an in-domain pre-training strategy to enhance the model's performance without requiring manual labeling by the physicians.

The Internal dataset combined the Italian, Indian, and Chinese datasets and was used for the CADx model development (training, validation, and internal testing). The Greek and Spanish datasets served as external validation cohorts.

Model development

The internal dataset was divided into training (70%), validation (20%), and testing (10%) subsets to prevent data leakage. This division followed a patient-level strategy to avoid having images from the same patients in different sets. The CADx model used the EfficientNetV2S architecture⁶⁹, selected for its balance between accuracy and computational efficiency. Instead of following a general-domain pre-training method (i.e. using the ImageNet dataset), this process was based on our custom *Pre-Training* dataset. After this initial step, the *Internal* dataset was used to fine-tune the CADx model, introducing the ability to distinguish between LR and HR lesions.

Model training was performed on cropped lesion images corresponding to the annotated bounding boxes to emulate real clinical deployment, where the CADx model would operate downstream of a computer-aided detection (CADE) system like the one previously published by our group⁵⁰. This latter model would flag the lesion on the image and submit the image within the BB to the CADx model, which would ultimately analyze that cropped image. Consequently, both the training and the test sets were annotated with BBs from which the cropped images were automatically extracted. All cropped images were resized to 224×224 pixels using zero-padding to preserve spatial proportions. Data augmentation expanded the dataset size by 30% using small image transformations (rotations up to 10° , horizontal/vertical shifts up to 10%, zoom 0–10%, shear 10%, brightness variation, and flipping). Training parameters included 100 epochs, batch size of 8, Adam optimizer with an initial learning rate of 1×10^{-4} , and exponential decay (0.96). Training and inference were performed in Python (TensorFlow) on an NVIDIA Tesla V100 GPU (16 GB).

Evaluation metrics

The performance was evaluated by comparing the predicted classes with the ground-truth classes, resulting in confusion matrices as described in **Figure 16**. For every class, a confusion matrix with

task-specific definitions of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) was used.

		Predicted labels		
		High-risk	Low-risk	Healthy
Ground truth labels	High-risk	Correctly predicted as High-risk	Wrongly predicted as Low-risk	Wrongly predicted as Healthy
	Low-risk	Wrongly predicted as High-risk	Correctly predicted as Low-risk	Wrongly predicted as Healthy
	Healthy	Wrongly predicted as High-risk	Wrongly predicted as Low-risk	Correctly predicted as Healthy

Figure 16. Scheme displaying the confusion matrix of a model's or rater's predicted instances across three categories: Low-risk, High-risk, and Healthy. The actual labels are on the vertical axis, and the predicted labels are on the horizontal. The diagonal cells contain correctly classified instances, while the others contain the errors.

Since the CADx model was specifically trained to distinguish between LR and HR classes, and the Barcelona dataset included healthy cases without annotated BBs, these images were processed using a previously developed lesion detection system⁷⁰ to extract potential classifier inputs. The image's final prediction was considered healthy if no detection occurred. As the CADx model was trained only to distinguish between LR or HR, and the healthy cases were very few images (n=12; 8%), and their resulting classification depended on the detection algorithm (if it flagged them or not), they were excluded from the classification metrics computation.

The outcomes of this study were also analyzed using standard diagnostic metrics for classification in the CV field:

Precision (also known as positive predictive value) measures the percentage of objects that are correctly recognized among all the predicted outputs.

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

Recall (also known as sensitivity) is computed as the fraction of the objects that are correctly recognized, compared to all the ones that should have been recognized.

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

F1 score is a way of looking at both *Precision* and *Recall* parameters at the same time to assess the model's quality more comprehensively. The *F1 score* is defined as the harmonic mean of *Precision* and *Recall* as the relative contribution of both these metrics is equal. It is higher when both measures are similarly high.

$$F1score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN}$$

Accuracy is a metric that evaluates the overall correctness of a model's predictions by measuring the percentage of correctly classified instances above all the images. However, it can be significantly affected by class imbalance, which occurs when the different classes of the dataset are disproportionally distributed. Considering our test sets are evenly distributed, accuracy represents a robust measure making it the primary outcome for assessing the overall diagnostic performance of the CADx model.

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

-Receiver Operator Characteristic (ROC) curve: This is a probability curve that plots the TP rate (recall) against the FP rate. The more the curve approaches the upper left corner, the higher is the TP

rate, the lower is the FP rate, and the wider will be the Area Under the Curve (AUC). The AUC of a classifier provides an aggregate measure of performance.

Comparison with physicians and large language models

To benchmark AI performance, the CADx model was compared against clinicians with different expertise levels and against the ChatGPT-4o large language model.

For this, physicians with varying degrees of expertise in laryngoscopy were selected from the faculty and residents of the Hospital Clínic of 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* dataset was used for this purpose. Plane images without BB were shown to the raters, and they were asked to classify each image as HR (lesions that looked like a leukoplakia or a carcinoma), LR (all the other benign lesions), or healthy (no lesion detected). 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 (September 2024 version) was prompted in English: “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?” Images were submitted individually, and responses were recorded for analysis. A scheme summarizing the whole experiment is reported in **Figure 17**.

Statistical analysis

Statistical comparisons between the CADx model and human raters used McNemar’s test on classification accuracy (primary outcome). Comparisons of ROC AUC between the internal and

external test sets employed pairwise Z-tests following the Hanley and McNeil method. 95% confidence intervals (CIs) were estimated via bootstrap resampling. Significance was set at $p < 0.05$. Analyses were performed using Python 3.9 and R Studio 2024.09.1+394.

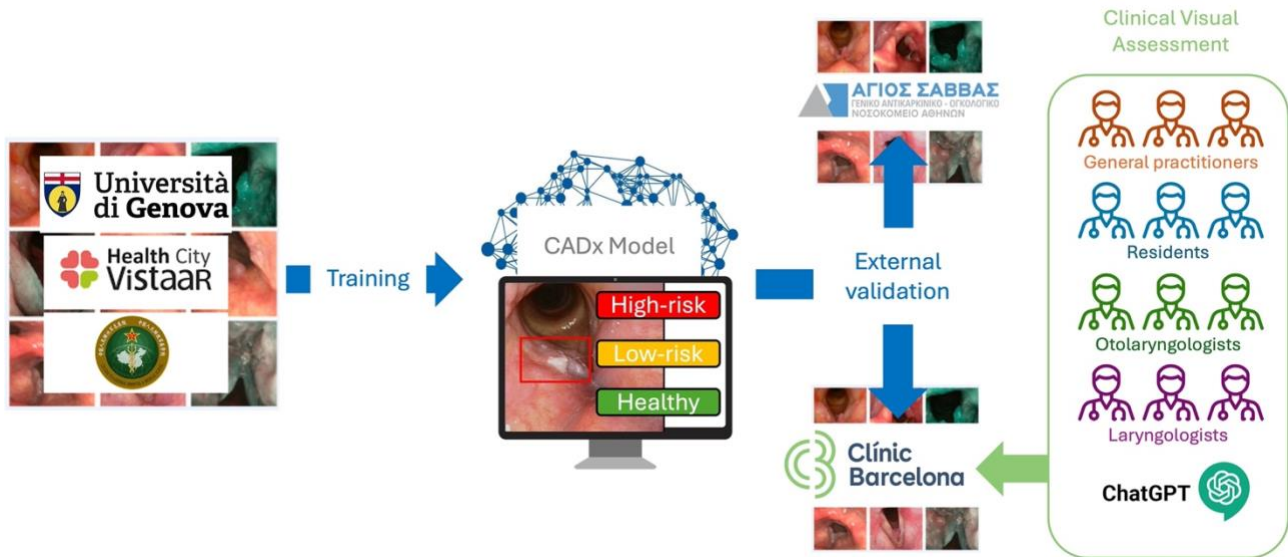


Figure 17. In this scheme, the experiment is summarized. The Genova, Lucknow, and Laryngoscope8 datasets were used as the Internal dataset for model development. The model pipeline consisted of outputting a detection bounding box when a lesion was found and then classifying the bounding box content as a Low-risk or High-risk lesion. When no bounding box was outputted, the image was considered Healthy. The CADx model was tested on two external datasets: the Greek dataset and the Spanish dataset. On the latter dataset, 4 categories of physicians and ChatGPT-4o independently rated the images for clinical comparisons.

LARYNGEAL TUMORS SEGMENTATION

Purpose

The precise delineation of tumor margins is a key challenge in the surgical management of laryngeal cancer, particularly in transoral laser or robotic surgery, where inadequate margin identification can lead to residual disease. This experiment aimed to develop and validate a DL model for segmentation, named SegMENT-Plus, capable of automatically delineating the superficial extent of laryngeal cancer on both WL and NBI laryngoscopy. The goal was to achieve high segmentation accuracy and real-time feasibility, while assessing the model's generalizability on independent external datasets and comparing it with human physicians.

Dataset

Videolaryngoscopy recordings of patients with biopsy-proven laryngeal carcinoma who underwent pre-treatment endoscopic examinations between 2014 and 2020 at the Unit of Otolaryngology and Head and Neck Surgery, IRCCS Ospedale Policlinico San Martino – University of Genoa (Italy) were retrospectively collected.

Videolaryngoscopies were performed both in-office using a flexible nasopharyngolaryngoscope (HD Video Rhino-Laryngoscope Olympus ENF-VH) and intraoperatively using 0°, 30°, or 70° rigid telescopes (Visera Elite CLV-S190, Olympus Medical Systems, Tokyo, Japan). Both WL and NBI modalities were recorded sequentially.

From each videolaryngoscopy, five WL and five NBI representative frames were manually selected, prioritizing high-quality images with clearly visible tumor margins. The final dataset contained images of variable resolution (768×576 to 1920×1072 pixels).

Image annotation was performed using the CVAT annotation software (Intel Corporation, USA)⁶⁵ by three expert laryngologists, each with ≥ 5 years' experience in laryngology and NBI interpretation.

Tumor borders were manually outlined pixel by pixel, generating binary ground-truth masks labeled “squamous cell carcinoma”. Cases with uncertain margins were reviewed jointly, and final segmentation was defined by consensus. Both WL and NBI frames were included to ensure the model learned from multimodal visual features. This dataset, hereafter referred to as the University of Genoa segmentation dataset, was randomly split by patient into training/validation (90%) and test (10%) sets.

To test model generalizability, two external datasets were used:

1. University of Brescia dataset: acquired at the unit of otolaryngology of the Spedali Civili - University of Brescia (Italy)
2. Yonsei University dataset: collected at the department of otolaryngology of Severance Hospital - Yonsei University, Seoul (Republic of Korea).

Both centers followed the same annotation protocol, with segmentation performed by a single experienced laryngologist at each institution. These datasets were used exclusively for external validation.

Additionally, five complete preoperative videolaryngoscopies (not used for training) were selected from the University of Genoa to test computational speed and simulate real-time segmentation performance.

Model development

The segmentation model, SegMENT-Plus, was an improved version of the previously published, by our group, SegMENT architecture⁵⁷. SegMENT-Plus was designed to be more efficient, with 38.5% fewer parameters (17.7 million fewer) than the original version, improving computation speed while preserving accuracy.

Architectural enhancements included a single modified Atrous Spatial Pyramid Pooling (m-ASPP) module incorporating Squeeze-and-Excitation and Convolution Block Attention Modules

(CBAM) to capture finer local features, and a new residual link in the decoder to preserve early feature details during upsampling.

The model was implemented in Python 3.9 using Keras 2.11.0 and TensorFlow 2.11.0, trained for 30 epochs with a batch size of 8, initial learning rate = 0.001, and reduction by a factor of 0.1 after three consecutive stagnant epochs. The Adam optimizer was used to minimize the Dice Similarity Coefficient loss.

Training and testing were performed on a workstation equipped with dual Intel Xeon Gold 5222 CPUs (3.8 GHz), 128 GB RAM, and an NVIDIA RTX A6000 GPU (48 GB). All input images were resized to 256×256 pixels using bilinear interpolation (OpenCV 3.9) to standardize spatial dimensions and reduce computational load.

Evaluation metrics

Model performance was assessed by comparing the predicted masks with expert ground truth ones using pixel-wise analysis. Each pixel was categorized as true positive (TP), true negative (TN), false positive (FP), or false negative (FN). From these, the following metrics were derived:

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

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

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

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

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

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

- *Intersection over Union (IoU)*: is a metric that evaluates the similarity between the predicted segmentation and the ground truth segmentation (**Figure 18**). It measures the proportion of the intersection of the predicted and ground truth segmentations with respect to their union:

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

- *Dice similarity coefficient (DSC)*: is a metric that evaluates the correspondence between the predicted segmentation and the ground truth segmentation, similarly to the IoU. It is calculated as the ratio of twice the intersection of the predicted and ground truth segmentation and the sum of the areas of the predicted and ground truth segmentations.

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

DSC and IoU were selected as the primary outcome measures for statistical comparisons.

Inference speed, measured in *frames per second* (FPS), was used to assess real-time feasibility. Given that typical videolaryngoscopies are recorded at ~25 FPS, a similar inference speed was considered indicative of real-time capability (citing references 14 and 15 from the manuscript).

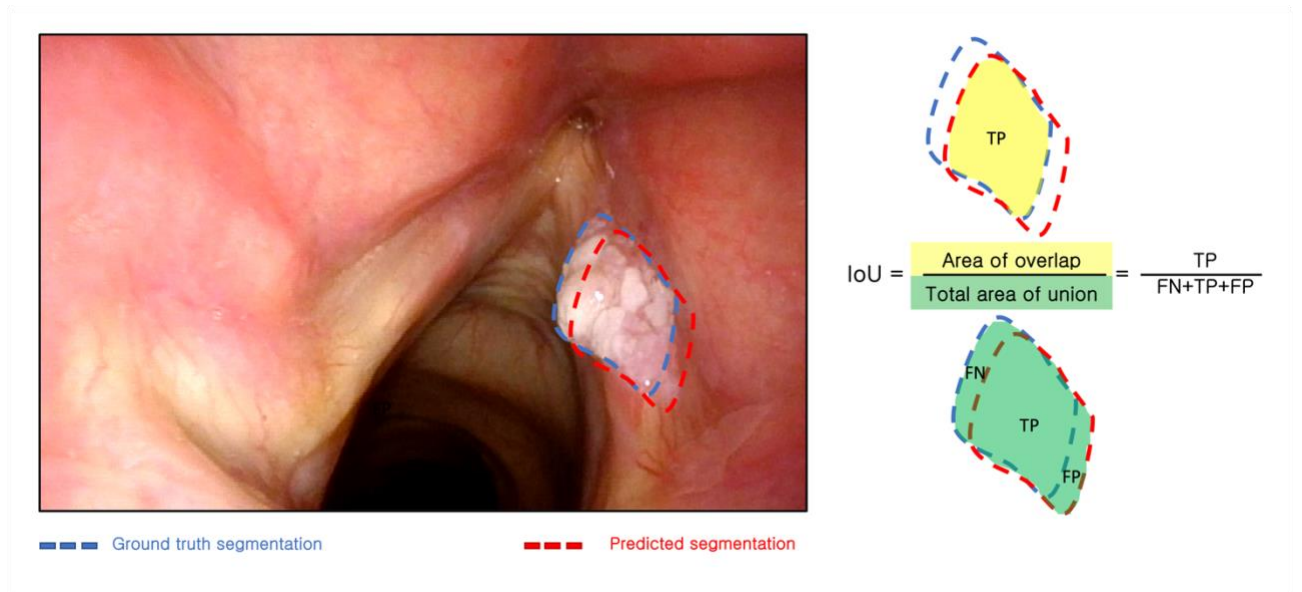


Figure 18. Graphical representation of the intersection over union (IoU) The boundary traced in light blue represents the ground truth segmentation, while the red one represents the model prediction. The IoU is calculated by dividing the overlapping area (containing the true positive pixels) by the total area of union (encompassing the false negative, true positive, and false positive pixels).

Comparison with human physicians

To better evaluate the efficacy of *SegMENT-Plus* we compared its segmentation performance with those of human physicians. A subset of 100 representative test images from the University of Genoa dataset was selected for human comparison. Three experts jointly reviewed and confirmed the ground truth. Two otolaryngology residents (one junior, second year; one senior, fourth year) independently segmented all images after training on the CVAT software. *SegMENT-Plus* predictions were compared with the residents' segmentations using DSC and IoU metrics.

Statistical analysis

Categorical variables were summarized as counts and percentages, while continuous variables (e.g., DSC, IoU, Recall, Precision, Accuracy) were reported as median $\pm$ interquartile range (IQR). Normality was assessed using the Shapiro–Wilk test. Comparisons between two independent groups were conducted using the Mann–Whitney U test. Comparisons among three or more independent groups used the one-way Anova test. Post-hoc analysis was performed using Tukey's multiple comparisons test. A two-sided $p < 0.05$ was considered statistically significant. All analyses were performed in Python 3.9 (packages: *scipy.stats*, *statsmodels 0.13.2*).

LARYNGEAL MOTILITY ASSESSMENT

Purpose

The accurate evaluation of laryngeal motility is crucial for diagnosing and managing vocal fold paralysis, yet traditional endoscopic assessments are often subjective and affected by inter- and intra-observer variability. This study developed and validated an artificial intelligence–based pipeline capable of automatically assessing vocal fold motility and classifying patients into three categories: healthy, unilateral paralysis, and bilateral paralysis.

Dataset

A multicenter dataset of in-office video-laryngoscopies was retrospectively collected from the Hospital Clínic de Barcelona (Spain) and the Azienda Ospedale – Università di Padova (Italy). Each video represented a different patient examined using transnasal flexible video-laryngoscopy recorded at 35 frames per second. Patients were instructed to phonate the vowel /i:/ followed by a deep inhalation to display full adduction and abduction of the vocal folds.

Inclusion criteria were the availability of complete video recordings and adequate visualization of the vocal folds. Laryngeal cancer cases were excluded to ensure homogeneity, as mucosal irregularities could confound motion analysis. Stroboscopic examinations were also excluded due to their lower temporal resolution.

Each case was reviewed and classified into one of three diagnostic categories: unilateral paralysis, bilateral paralysis, or normal motility. Three expert otolaryngologists classified each video in one of the three categories. When consensus was not achieved, the most frequent classification among reviewers was adopted. All data were anonymized, and ethical approval was obtained from both centers (Hospital Clínic de Barcelona Reg. HCB/2023/0897; Azienda Ospedale – Università di Padova 190n/AO/21).

Model development

Feature extraction

Video-laryngoscopies were first analyzed using the Automated Glottic Action Tracking by Artificial Intelligence (AGATI) toolbox⁵³. Built on the DeepLabCut framework, AGATI enabled the automatic tracking of vocal fold motion across frames and extraction of quantitative parameters, including the anterior glottic angle (AGA), left and right vocal fold angles, and their velocities and accelerations.

The AGA was chosen as the main quantitative parameter for assessing vocal fold motion. The AGA measures the opening angle formed by the free edges of the two vocal folds at their anterior commissure. As the vocal folds move during respiration or phonation, the AGA widens during abduction and narrows during adduction, providing a continuous and objective measure of glottic dynamics. This parameter was considered particularly reliable because it is less affected by patient or camera movements than other geometric measurements such as vocal fold displacement or area. Even when the endoscope slightly shifts or rotates, the angle between the two folds remains relatively stable, making the AGA a robust indicator of true physiological motion.

However, the AGA signal varies continuously throughout the recording and may include minor, clinically irrelevant fluctuations caused by breathing irregularities or subtle laryngeal tremor. To focus only on the major opening–closing movements that correspond to clinically meaningful abduction and adduction cycles, we applied a Gaussian Mixture Model (GMM) to the temporal distribution of AGA values (**Figure 19**).

The GMM statistically separates the AGA signal into two main components: one representing frames in which the glottis is relatively open (abduction phase) and another corresponding to the closed or nearly closed positions (adduction phase). This step allowed the algorithm to isolate distinct motion cycles, improve signal quality, and ensure that subsequent quantitative analysis reflected the true functional excursions of the vocal folds rather than random noise or minor movements.

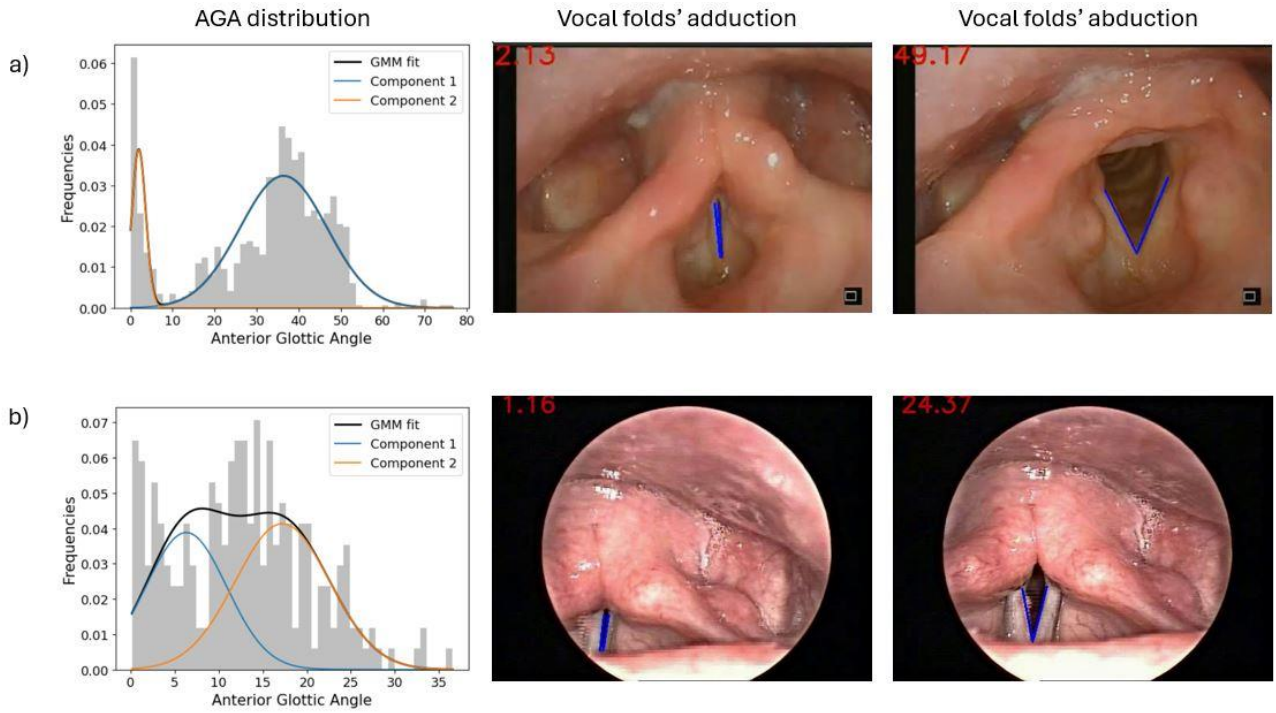


Figure 19. Fit of the Anterior Glottic Angle (AGA)'s distribution using a Gaussian Mixture Model (GMM) consisting of a mixture of two Gaussian distributions. Row (a): subject with healthy motility; row (b): patient affected by laryngeal paralysis, where abduction and adduction are inevitably more similar. AGA values are shown in red in the top-left part of each frame.

By means of this model, the frames where the vocal folds were at maximum abduction and adduction were automatically identified. This kind of feature selection was conducted for each patient independently of their class.

The constraints we imposed to accept a trial were the following: it should be composed of 100 frames, i.e., 4 seconds, in which at least 10% of the AGA values were lower than the lowest of the two Gaussian means, and another at least 10% should have been larger than the largest of the two Gaussian means, avoiding considering superposed trials. In this manner, trials representing the entire excursion of the vocal folds were extracted, from complete abduction to complete adduction. This allowed to identify relevant segments for laryngeal motility assessment from each video. On average, this method automatically extracted 4.21 trials per video.

Definition of quantitative features

For each trial, a set of quantitative descriptors was computed to capture the degree and symmetry of vocal fold motion. Features included the maximum and minimum AGA values (reflecting full abduction and adduction), the AGA standard deviation, and the difference between the two Gaussian means obtained by GMM fitting.

To describe inter-fold coordination, Kendall's correlation and Granger's causality were calculated between left and right vocal fold angles. These measures quantified synchronization and directional dependence of motion, enabling differentiation between unilateral and bilateral paralysis.

In addition to quantitative motion parameters, an image-based feature was computed to visually represent the areas of maximal movement during the glottic cycle. The frames corresponding to maximal abduction and maximal adduction were first identified. The adduction frame was then subtracted from the abduction frame, producing a difference map that highlighted the regions of greatest displacement between these two extremes of motion (**Figure 20**).

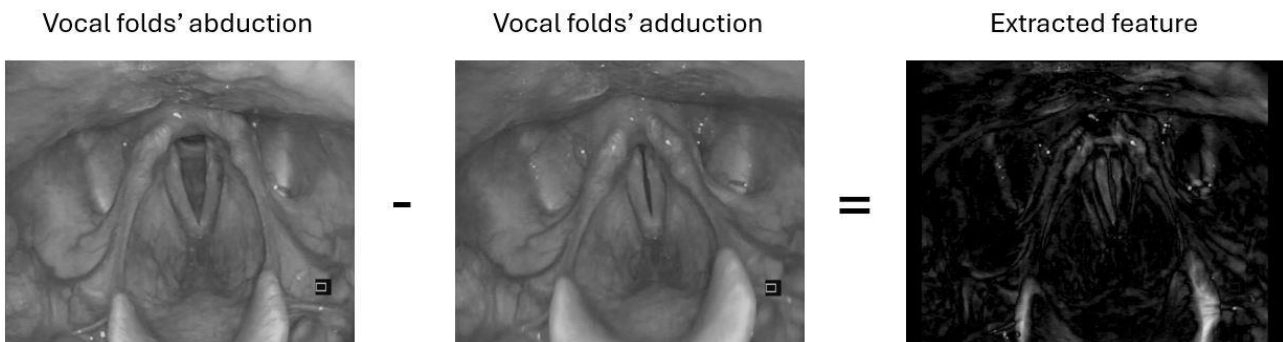


Figure 20. In the right image we show the image feature obtained comparing the two left images representing the vocal folds in maximal abduction and maximal adduction. The resulting feature is an image representing the motion and the changes in the anatomical configuration between the two poses, where white pixels represent the ones with the highest difference between the two frames.

In the resulting maps, regions exhibiting substantial motion appeared as bright areas, while static structures such as the laryngeal framework remained dark. This visualization provided an intuitive summary of vocal fold dynamics, emphasizing the amplitude and symmetry of movement. In healthy subjects, both vocal folds produced symmetrical motion patterns, whereas in cases of unilateral paralysis, changes were visible predominantly on one side, and in bilateral paralysis, motion was markedly reduced or absent on both sides. The inclusion of this feature enabled the model to capture

spatial information about motility, complementing the numerical parameters derived from temporal motion analysis.

Classification model

Because the feature set included both statistical and image-derived data, a multimodal architecture combining a CNN and a multilayer perceptron (MLP) was implemented. The CNN component, based on the ResNet-18⁷¹ backbone pretrained on the ImageNet dataset¹⁷, processed the image features, while the MLP integrated the statistical variables, i.e, the maximum and minimum AGA values, the AGA standard deviation, and the difference between the two Gaussian means obtained by GMM fitting.

The model architecture is illustrated in **Figure 21**. The combined output was used for two tasks:

1. Binary classification (healthy vs paralyzed)
2. Multi-class classification (healthy vs unilateral vs bilateral paralysis)

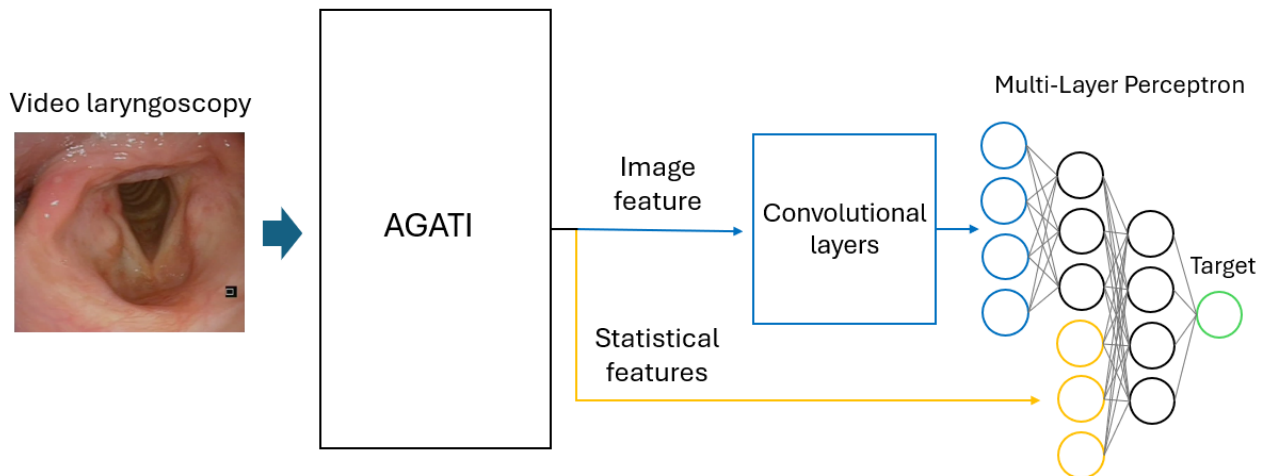


Figure 21. The figure illustrates the classification pipeline utilized for patient categorization. Initially, video-laryngoscopies were processed through the AGATI toolbox, followed by trial definition and feature extraction. Subsequently, a combined CNN/MLP model was used to optimize predictions by leveraging the extracted image feature together with the other statistical features in an end-to-end model. This model passed the image feature to the Resnet-18 convolutional layers. The output of the CNN was then passed to an MLP layer for dimensionality reduction and, finally, after having concatenated all the remaining features, the model produced the prediction after another few additional MLP layers.

Model optimization was performed through a 5-fold cross-validation procedure. Each patient's trials were kept within a single fold to avoid data leakage. Class imbalance was addressed through weighted loss functions proportional to class representation.

Training was performed using the Adam optimizer and early stopping after 20 epochs without improvement in validation loss, with a maximum of 1000 epochs per run. Models were evaluated both at the trial level and at the aggregated patient level, where patient classification was derived from the mean probability of their trials.

Evaluation metrics

The classification model was evaluated at two levels: individual motion trials and aggregated patient outcomes. Performance was quantified using standard previously explained CV classification metrics, including *accuracy*, *precision*, *recall*, and *F1-score*. Balanced accuracy was used as the primary performance indicator to account for the unequal number of subjects across diagnostic categories. For the binary task, distinguishing healthy from paralyzed cases, accuracy and balanced accuracy were both reported, while for the three-class task (healthy, unilateral paralysis, bilateral paralysis), class-wise metrics were computed and averaged. Confusion matrices were generated to illustrate misclassification patterns, and ROC curves were plotted for each diagnostic category to assess discrimination ability. The area under the ROC curve (AUC) was used to summarize model performance.

To evaluate the relative contribution of each feature group, an ablation study using the Shapley value method was conducted. Shapley values were used to quantify the contribution of each feature group to overall classification accuracy. Four groups ("players") were defined:

1. The image feature from the CNN;
2. The traditional AGA-based features;
3. Statistical descriptors of AGA variability;

4. Correlation-based features between left and right folds.

Each model variant was retrained using all possible feature combinations, and Shapley values quantified the contribution of each group to classification accuracy.

Statistical analysis

Descriptive statistics were used to summarize the clinical distribution of the dataset and the quantitative features derived from motion tracking. Continuous variables were expressed as median $\pm$ median absolute deviation, and categorical variables as frequencies and percentages. For model evaluation, the median and the median absolute deviation of each performance metric were computed across the five cross-validation folds. Statistical comparisons between models were performed using the McNemar test for paired categorical outcomes. Statistical significance was set at $p < 0.05$. All analyses were carried out in Python 3.9 using Scikit-learn and Statsmodels libraries.

RESULTS

INFORMATIVE FRAME SELECTION

Dataset overview

The final internal dataset combined the NBI-InfFrames set with the Genoa dataset, yielding 5147 frames. Data were split at the patient level into training ($n = 3126$), validation ($n = 1317$), and testing ($n = 704$; 346 informative, 358 uninformative) subsets. Data augmentation included vertical and horizontal shifts (10%), zoom (10%), small rotations (0° – 10°), and horizontal flips.

The external test set comprised 646 frames (306 informative, 340 uninformative) collected only at the Hospital Clínic de Barcelona. This independent cohort, annotated by a different clinical team, was used to evaluate model robustness across institutions and annotation processes. The breakdown and content of the dataset is reported in **Table 2**

Table 2. Dataset breakdown

Datasets	Allocation	No. of frames	Optical modality	Laryngeal conditions
NBI-InfFrames	Internal dataset	720	NBI	carcinoma
Hospital San Martino, Genova	Internal dataset	4427	WL+NBI	Carcinoma, benign lesions, healthy
Hospital Clínic, Barcelona	External test set	646	WL+NBI	Carcinoma, benign lesions, healthy

Model performance on internal dataset

The classification metrics obtained by the three CNNs when predicting which frames were informative in the internal test set (704 frames) are reported in **Table 3**. The ResNet-50 model achieved the best performance during the internal testing phase of the classification model. Out of the

704 test frames, 673 were correctly classified, resulting in precision = 95%, recall, = 97%, and the F1-score= 96%. The F1 scores of the Shallow CNN vs. MobileNetv2 and Shallow CNN vs. ResNet-50 were significantly different (p -value<0.05), while MobileNetv2 and ResNet-50 performed closely on the internal set.

Table 3. Performance of the deep learning models on the internal and external datasets

	True positives	True negatives	False positives	False negatives	Precision	Recall	F1score
Internal test set							
Shallow CNN	264 (76.3%)	313 (87.4%)	45 (12.6%)	82 (23.7%)	0.85	0.76	0.81
MobileNetv2	333 (96.2%)	339 (94.7%)	19 (5.3%)	13 (3.8%)	0.95	0.96	0.95
ResNet-50	334 (96.5%)	339 (94.7%)	19 (5.3%)	12 (3.5%)	0.95	0.97	0.96
External test set							
ResNet-50	272 (88.9%)	331 (97.4%)	9 (2.6%)	34 (11.1%)	0.97	0.89	0.93

External validation

Given the superior F1 score, the ResNet-50 model was chosen for additional external test evaluation and video testing. The pretext ResNet-50 model performed similarly on the external set achieving precision, recall, and F1-score were equal to 97%, 89%, and 93%, respectively, with no statistical difference compared to the internal dataset performance for F1 (p -value = 0.062). **Figure 22** shows some graphical examples of the output of the pretext ResNet-50 model on the external test set.

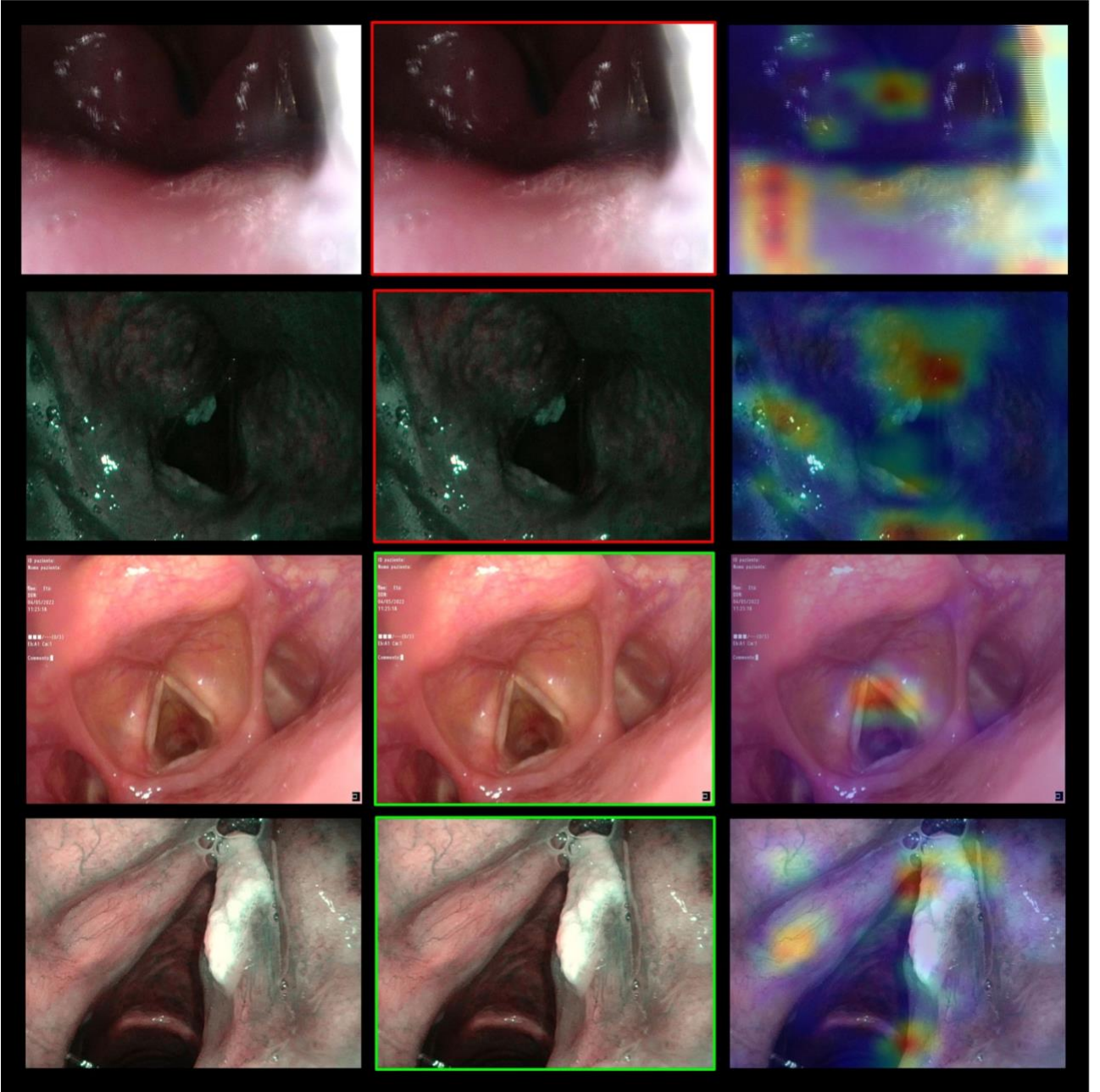


Figure 22. Examples of frames from the external test set. In the first column, the original image is shown. In the second column, the prediction is output as a green frame for informative frames and as a red frame for uninformative frames. In the third column, the areas used by the model for the prediction are output according to a gradCAM heatmap.

Model performance on the video dataset

Table 4 resumes the video characteristics. With the current experiment hardware, the median computing times for the five videos was 0.015 [0.012-0.018] milliseconds per frame. **Figure 23** shows the evolution of the Video Informativeness Level (VIL) during the different test videos.

Table 4. Characteristics of the tested laryngoscopy videos. Inference times and the Video Informativeness Level (VIL) are calculated after applying the Res-Net50 deep learning model.

ID video	Length	Frames per second	Median inference time	VIL	Clinical setting
1	00' 42''	29.97	0.0155s	60%	In-office
2	00' 53''	29.97	0.0156s	8%	In-office
3	00' 53''	29.97	0.0155s	55%	In-office
4	01' 07''	29.97	0.0155s	33%	In-office

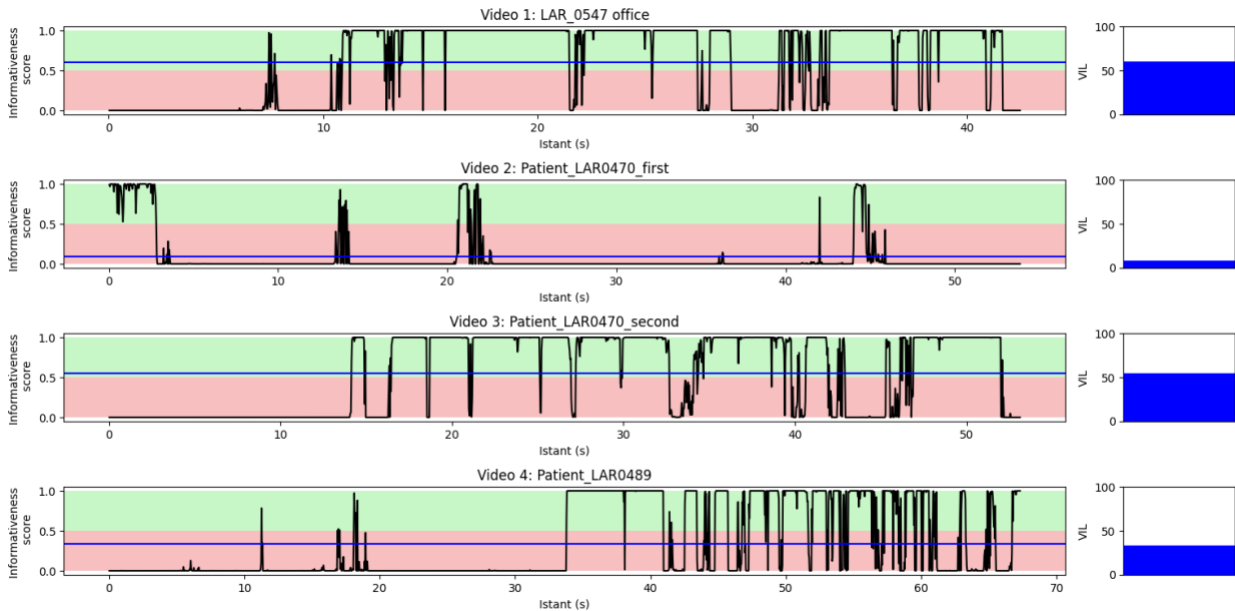


Figure 23. Plots of the model's (ResNet-50) predictions trend during each test video. The black line represents the probability for the frame to be classified as "informative" (when in the green background) or "uninformative" (when in the red background). The blue line represents the average probability for the frames of the whole video to be assigned to the category "informative". At the end of each bar plot a histogram with the Video Informativeness Level (VIL) value for the same video is reported. Video 2 and Video 3 report the in-office diagnostic laryngoscopy of the same patient before and after local anesthesia.

Study interpretation

Overall, the model demonstrated high reliability in distinguishing diagnostically useful frames from non-informative ones in laryngeal endoscopic videos, both on internal and external test sets. Its real-time performance confirmed feasibility for clinical integration, permitting continuous selection of relevant frames during examination, and allowing it to be used in combination with other models.

The Video Informativeness Level (VIL) might support more efficient and standardized video interpretation, giving an objective score of how informative a laryngoscopic examination is.

LARYNGEAL TUMORS DETECTION

Dataset overview

A total of 3,870 images from the internal datasets (ENDO-LC1, ENDO-LC2, and Laryngoscope8) were used for model development and validation (**Table 5**). The independent external dataset (ENDO-LC ext) included 149 images from the University of Genoa. After quality control, all images showed consistent annotation of lesion locations and a good balance between WL and NBI modalities.

Table 5. Breakdown of the images composition of the three different were organized into internal training, validation, and test sets. ENDO-LC ext was used for external validation.

Dataset	Optical modality	Number of patients	Number of frames
ENDO-LC 1	WL, NBI	233	2,037
Laryngoscope8	WL	1,297	1,700
ENDO-LC 2	WL, NBI	63	155
Training set	WL, NBI	188 ENDO-LC 1; 1164 Laryngoscope8; 60 ENDO-LC 2	3,452 (1,816 pathologic; 1,636 healthy)
Validation set	WL, NBI	36 ENDO-LC 1; 47 Laryngoscope8; 3 ENDO-LC 2	269 (pathologic)
Test set	WL, NBI	9 ENDO-LC 1; 86 Laryngoscope8	171 (pathologic)
ENDO-LC ext	WL, NBI	73	149 (pathologic)

Model performance on internal dataset

On the internal test set, the Super-Resolution Enhanced YOLO reached an AP of 0.83 ± 0.005 outperforming the original YOLOv8n ($AP = 0.79 \pm 0.016$) even though not significantly ($p = 0.10$). Performance varied according to lesion size. For medium lesions (32–96 pixels), the Super-Resolution Enhanced YOLO model achieved the best detection with an AP of 0.85 ± 0.006 , whereas for small lesions (<32 pixels), its AP, though reducing, remained acceptable at 0.82 ± 0.018 , showing a clear gain over the standard YOLOv8n without the Super-Resolution branch ($AP = 0.66 \pm 0.02$; p

< 0.05). These findings demonstrate that the integrated super-resolution strategy improved the model's sensitivity for subtle or early lesions while maintaining robustness across lesion scales.

When the two imaging modalities were analyzed separately, there was no substantial difference between them at bootstrapping sampling, with NBI images yielding slightly higher accuracy ($AP = 0.88 \pm 0.003$) than WL images ($AP = 0.86 \pm 0.005$), confirming that if adequately trained with enough images, the model can adapt to both imaging modalities without issues.

External validation

On the external dataset, the Super-Resolution Enhanced YOLO maintained high generalizability with an AP of 0.82 ± 0.05 , showing no statistically significant difference from its internal performance (no difference at bootstrapping sampling). Detection remained reliable across both small ($AP = 0.80 \pm 0.02$) and medium lesion categories ($AP = 0.82 \pm 0.03$). By reviewing the BBs generated by the model in the images, visual inspection by expert otolaryngologists confirmed correct lesion localization in nearly all true-positive detections, including cases of early glottic carcinoma and benign vocal fold lesions (**Figure 24**).

Real-time feasibility

The Super-Resolution Enhanced YOLO model demonstrated excellent computational efficiency. Despite the added super-resolution branch during training, the inference phase remained lightweight, achieving 81 FPS on a single NVIDIA RTX A6000 GPU. This result indicates suitability for real-time deployment on standard clinical hardware. Qualitative inspection of inference videos showed that the model reliably drew bounding boxes around lesions without flickering or delay, preserving stable predictions across consecutive frames.

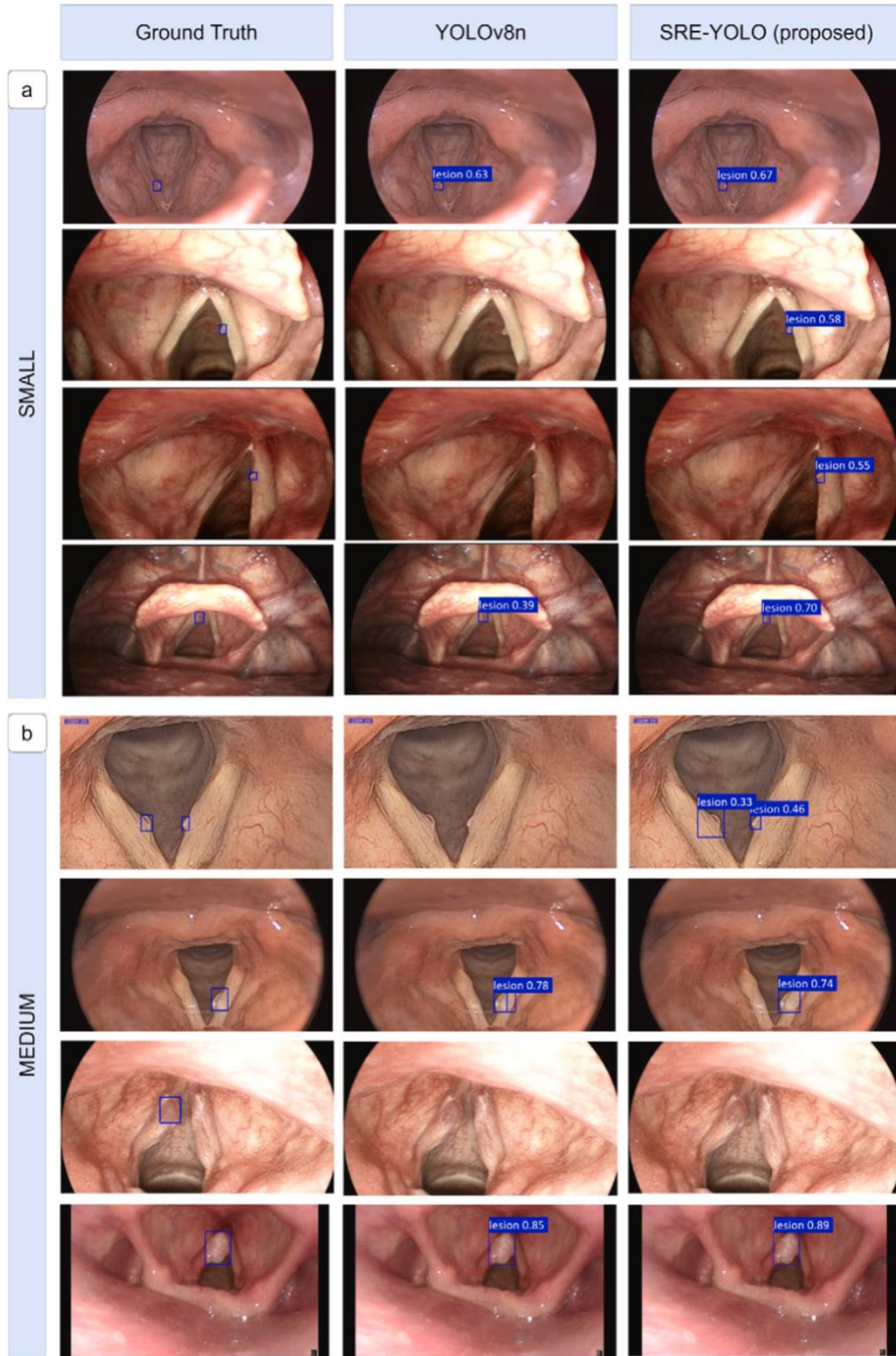


Figure 24. Visual results of the detection models on different lesions according to their sizes

Study interpretation

Overall, the proposed detection system achieved robust lesion detection across multiple centers and imaging modalities. The model accurately localized both benign and malignant lesions, including small or early-stage abnormalities, while maintaining real-time processing speed. This balance between precision and computational efficiency supports potential clinical use as an assistive tool for early lesion recognition during routine endoscopic assessment.

LARYNGEAL TUMORS DIAGNOSIS

Dataset overview

The total amount of frames used for the CADx model development was 21,834, consisting of 13,184 frames cropped based on the annotations and 8,651 unlabeled full-size frames. The *Internal* dataset included 12,118 cropped frames, which were split into 8,477 for training, 2,415 for validating, and 1,226 for testing. The *External* dataset included a total of 1,079 images, 909 from the *Greek* dataset, cropped into 929 frames, and 150 from the *Spanish* dataset, of which 13 healthy cases were considered only for the comparisons with physicians and ChatGPT. Demographics were not prospectively recorded during the dataset collection, and due to the anonymization of the images, they were not retrospectively available.

A comprehensive delineation of the number of frames allocated to each category within the *Internal* and *External* datasets is provided in **Figure 25**.

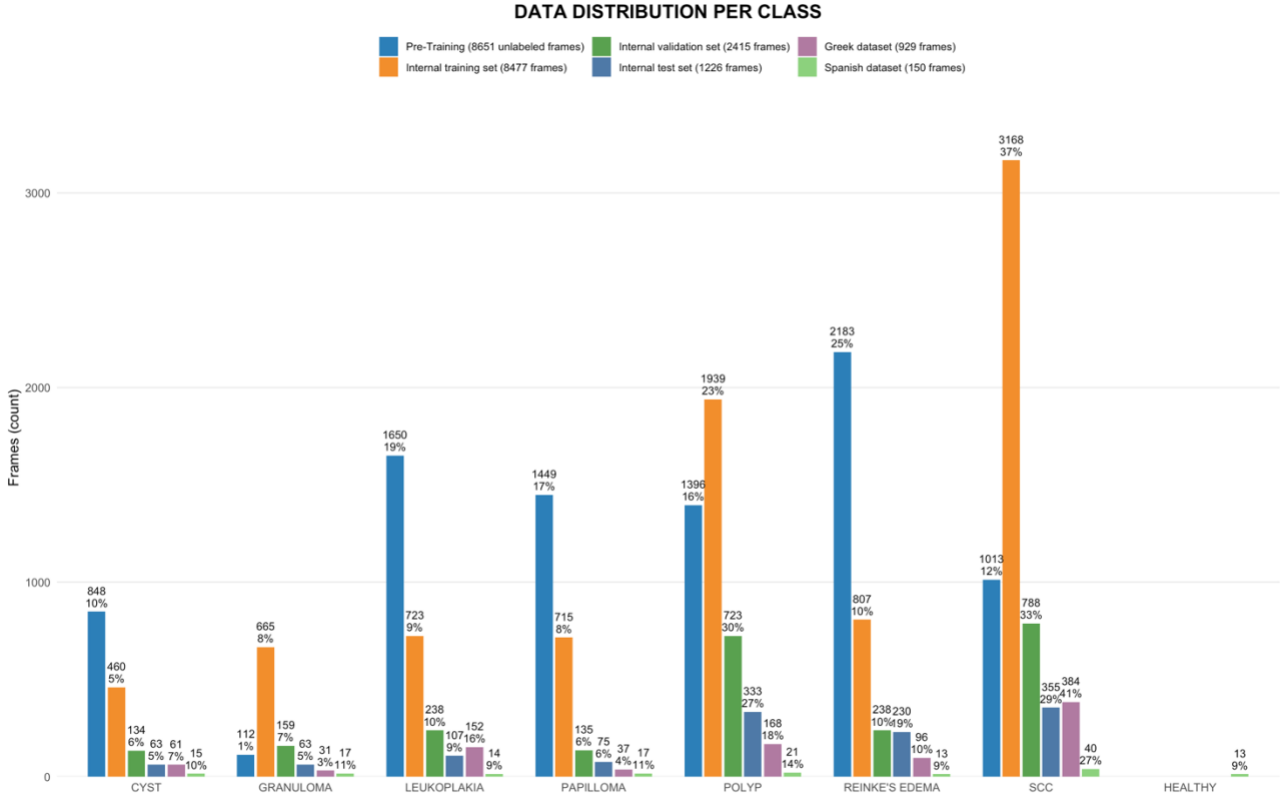


Figure 25. A representation of the datasets used in the experiment, divided according to the image categories. Percentages refer to the related dataset.

CADx model performances on internal and external datasets

On the *Internal* test set, the CADx model performed well, reaching an accuracy and AUC of 0.9 and 0.89, respectively. On the *External* datasets, the models performed similarly with an accuracy of 0.85 and an AUC of 0.85 ($p=0.075$) on the *Greek* dataset, and an accuracy of 0.88 and an AUC of 0.88 ($p=0.73$) on the *Spanish* dataset (Table 6 and Figure 26A).

Table 6. Performance metrics comparisons between the CADx model on the Internal test set and on the External datasets (Greek dataset and Spanish dataset). AUC= area under the receiving-operator curve. [95% confidence intervals]

CADx model	Internal test set (n=1,226)			External Greek dataset (n=929)			External Spanish dataset (n=150)		
	Low-risk	High-risk	Weighted average	Low-risk	High-risk	Weighted average	Low-risk	High-risk	Weighted average
Precision	0.91 [0.87, 0.93]	0.9 [0.87, 0.92]	0.9 [0.88, 0.92]	0.79 [0.75, 0.83]	0.9 [0.87, 0.93]	0.86 [0.83, 0.88]	0.92 [0.86, 0.97]	0.83 [0.72, 0.93]	0.89 [0.83, 0.94]
Recall	0.94 [0.93, 0.96]	8.84 [0.80, 0.87]	0.9 [0.88, 0.92]	0.88 [0.84, 0.91]	0.83 [0.80, 0.86]	0.85 [0.80, 0.95]	0.88 [0.80, 0.95]	0.89 [0.81, 0.96]	0.88 [0.82, 0.93]
F1	0.92 [0.91, 0.94]	0.87 [0.84, 0.89]	0.9 [0.87, 0.92]	0.83 [0.80, 0.86]	0.87 [0.84, 0.89]	0.85 [0.83, 0.87]	0.9 [0.85, 0.95]	0.86 [0.78, 0.92]	0.88 [0.83, 0.93]
Accuracy	0.9 [0.87, 0.91]			0.85 [0.83, 0.87]			0.88 [0.83, 0.93]		
AUC	0.89 [0.87, 0.91]			0.85 [0.83, 0.88]			0.88 [0.83, 0.93]		

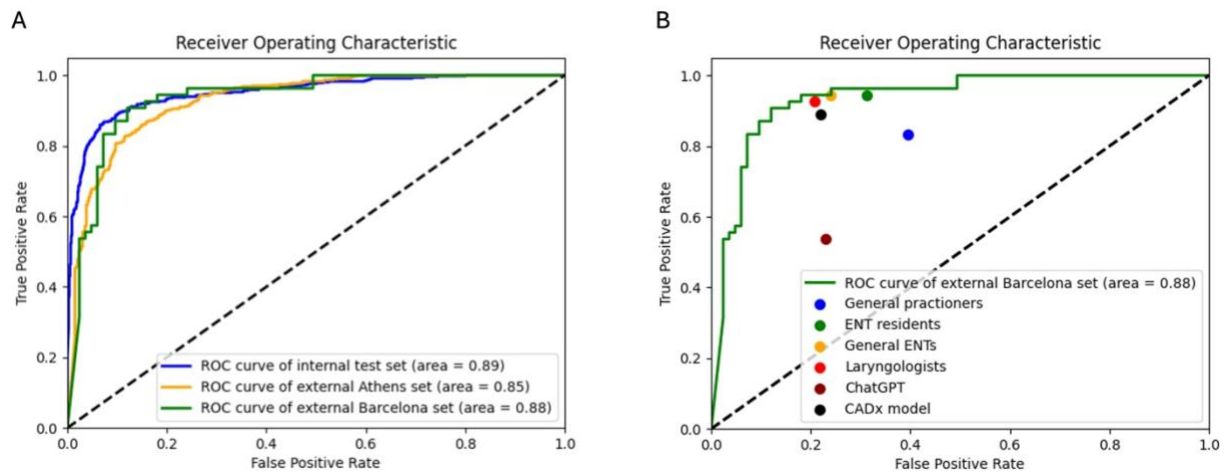


Figure 26. The plot shows the Receiver Operator Characteristic (ROC) curves of the CADx model performance on the three datasets. B: The plot shows the comparison between clinical raters, ChatGPT, and the CADx model.

CADx model performance comparison with clinical raters and Large Language Models

The comparison of the CADx model with the physicians and ChatGPT-4o is summarized in **Table 7**. The model performed comparably to the otolaryngology residents (Accuracy 0.88. vs. 0.87; $p=0.87$), the general otolaryngologists (Accuracy 0.88. vs. 0.89; $p=1.00$), and the expert laryngologists (Accuracy 0.88. vs. 0.93; $p=0.28$), while it performed significantly better compared to the general practitioners (Accuracy 0.88. vs. 0.69; $p<0.001$), and ChatGPT-4o (Accuracy 0.88. vs. 0.68; $p=0.001$). **Figure 26B** depicts the ROC curve comparisons between the CADx model and the raters. The confusion matrices of the model, the clinical raters, and ChatGPT-4o are reported in **Figure 27**.

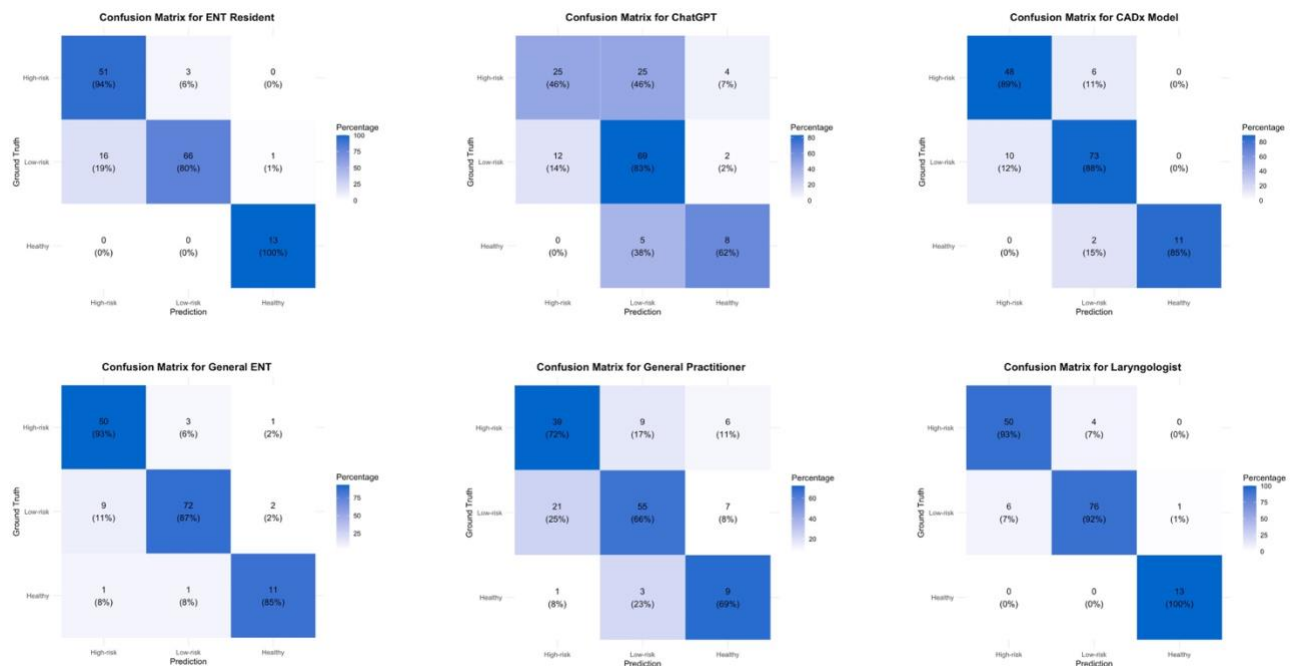


Figure 27. The confusion matrices of the CADx model, the 4 clinical raters, and ChatGPT-4o on the Barcelona external dataset. The actual labels are on the vertical axis, and the predicted labels are on the horizontal axis. The diagonal cells contain correctly classified instances, while the others contain errors.

Results

Table 7. Performance metrics comparisons between the CADx model, the physicians, and ChatGPT on the Spanish dataset (n=150). The p-values are calculated using McNemar's test comparing each rater's Accuracy (primary outcome) with the CADx model's value. P-values in bold are statistically significant. [95% confidence intervals]

Rater	Accuracy	Precision			Recall			F1		
		Healthy	High-risk	Low-risk	Healthy	High-risk	Low-risk	Healthy	High-risk	Low-risk
CADx model	0.88 [0.83, 0.93]	1.00 [1.00, 1.00]	0.83 [0.72, 0.92]	0.90 [0.83, 0.96]	0.85 [0.61, 1.00]	0.89 [0.79, 0.96]	0.88 [0.81, 0.95]	0.92 [0.76, 1.00]	0.86 [0.78, 0.92]	0.89 [0.84, 0.97]
General Practitioner	0.69 [0.63, 0.76] (p<0.0001)	0.41 [0.21, 0.62]	0.64 [0.52, 0.76]	0.82 [0.73, 0.91]	0.69 [0.45, 0.93]	0.72 [0.60, 0.84]	0.66 [0.56, 0.76]	0.51 [0.29, 0.71]	0.68 [0.57, 0.77]	0.73 [0.65, 0.80]
Otolaryngology Resident	0.87 [0.81, 0.92] (p= 0.87)	0.93 [0.78, 1.00]	0.76 [0.65, 0.86]	0.96 [0.91, 1.00]	1.00 [1.00, 1.00]	0.94 [0.88, 1.00]	0.80 [0.70, 0.88]	0.96 [0.87, 1.00]	0.84 [0.77, 0.91]	0.87 [0.80, 0.92]
General Otolaryngologist	0.89 [0.84, 0.93] (p= 1.00)	0.79 [0.56, 1.00]	0.83 [0.74, 0.93]	0.95 [0.89, 0.99]	0.85 [0.62, 1.00]	0.93 [0.84, 0.98]	0.87 [0.79, 0.94]	0.81 [0.63, 0.95]	0.88 [0.81, 0.93]	0.91 [0.86, 0.95]
Laryngologist	0.93 [0.88, 0.97] (p=0.28)	0.93 [0.78, 1.00]	0.89 [0.81, 0.96]	0.95 [0.90, 0.99]	1.00 [1.00, 1.00]	0.93 [0.85, 0.98]	0.92 [0.85, 0.97]	0.96 [0.87, 1.00]	0.91 [0.84, 0.96]	0.93 [0.89, 0.97]
ChatGPT	0.68 [0.61, 0.76] (p=0.0001)	0.57 [0.30, 0.82]	0.68 [0.52, 0.82]	0.70 [0.61, 0.79]	0.62 [0.30, 0.86]	0.46 [0.33, 0.60]	0.83 [0.75, 0.91]	0.59 [0.31, 0.79]	0.55 [0.42, 0.67]	0.77 [0.68, 0.82]

Study interpretation

Overall, the CADx model successfully classified laryngeal lesions into high-risk, low-risk, and healthy categories across diverse centers and imaging modalities. The system performed comparably to expert clinicians and outperformed less experienced raters, confirming strong diagnostic capability and generalizability. These findings highlight the potential role of AI-based optical biopsy as a supportive instrument for rapid, reproducible risk stratification of laryngeal lesions.

LARYNGEAL TUMORS SEGMENTATION

Dataset overview

A total of 3933 annotated frames from 557 patients were included in the University of Genoa dataset. Of these, 2025 (51.5%) were acquired under WL and 1908 (48.5%) under NBI. Images were equally distributed between in-office (flexible endoscopy) and intraoperative (rigid endoscopy) examinations. Following patient-level splitting, 3539 images (90%) were assigned to the training/validation set, and 394 (10%) to the internal test set. The detailed dataset composition, including acquisition setting (in-office vs. intraoperative) and modality (WL vs. NBI), is reported in **Table 8**.

Table 8. Composition of the University of Genova segmentation image dataset. The subdivision of frames in training-validation and test datasets is reported. The acquisition setting (in-office vs. intra-operative) and imaging modality (white light = WL vs. Narrow Band Imaging = NBI) are described.

Dataset Distribution	Type of Examination	No.	Imaging	No.	No.
Training- Validation	In-office	1208	WL	1884	3146
	Intra-operative	676			
	In-office	1153	NBI	1655	
	Intra-operative	502			
Test	In-office	129	WL	208	394
	Intra-operative	79			
	In-office	122	NBI	186	
	Intra-operative	64			

The two external test sets, which were used for generalization analysis, were composed as follows:

- University of Brescia dataset: 200 WL and NBI images from 146 patients.
- Yonsei University dataset: 156 WL and NBI images from 63 patients.

Internal test performance

On the University of Genova test set, the model achieved median values of DSC = 0.827 (0.70-0.90), IoU = 0.828 (0.73-0.90), Recall = 0.881 (0.74-0.96), Precision = 0.855 (0.70-0.94), Accuracy = 0.972

(0.95-0.99), and median inference speed of 25.6 fps. A further comparison was made by analyzing separately the WL and NBI images of the testing set to evaluate if the model performance could be influenced by the imaging modality. No significant differences in the model performance were noticed between WL and NBI images in terms of DSC = 0.81 (0.66-0.88) and 0.81 (0.70-0.89), respectively, $p = 0.06$; and IoU = 0.83 (0.72-0.88) and 0.82 (0.73-0.89), respectively, $p = 0.78$.

External validation

When evaluated on independent datasets, SegMENT-Plus maintained high performance without additional fine-tuning (**Table 9**). In both the external datasets employed, *SegMENT-Plus* showed comparable performances to the University of Genova test set for the IoU ($p = 0.05$), while for DSC, on the Yonsei University dataset the model showed slightly superior performances (0.838, 0.68-0.89) compared to the University of Genova (0.828, 0.73-0.90; $p = 0.001$) and University of Brescia (0.807, 0.70-0.87; $p = 0.001$). These findings support the model's generalizability across imaging devices and acquisition protocols. Qualitative examples of external cases are shown in **Figure 28**.

Table 9. Results of SegMENT-Plus on the University of Genova testing dataset and on the two external testing datasets (University of Brescia and Yonsei University) are reported as median (first and third quartiles). IoU = Intersection over Union; DSC = dice similarity coefficient; FPS = frames per second.

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

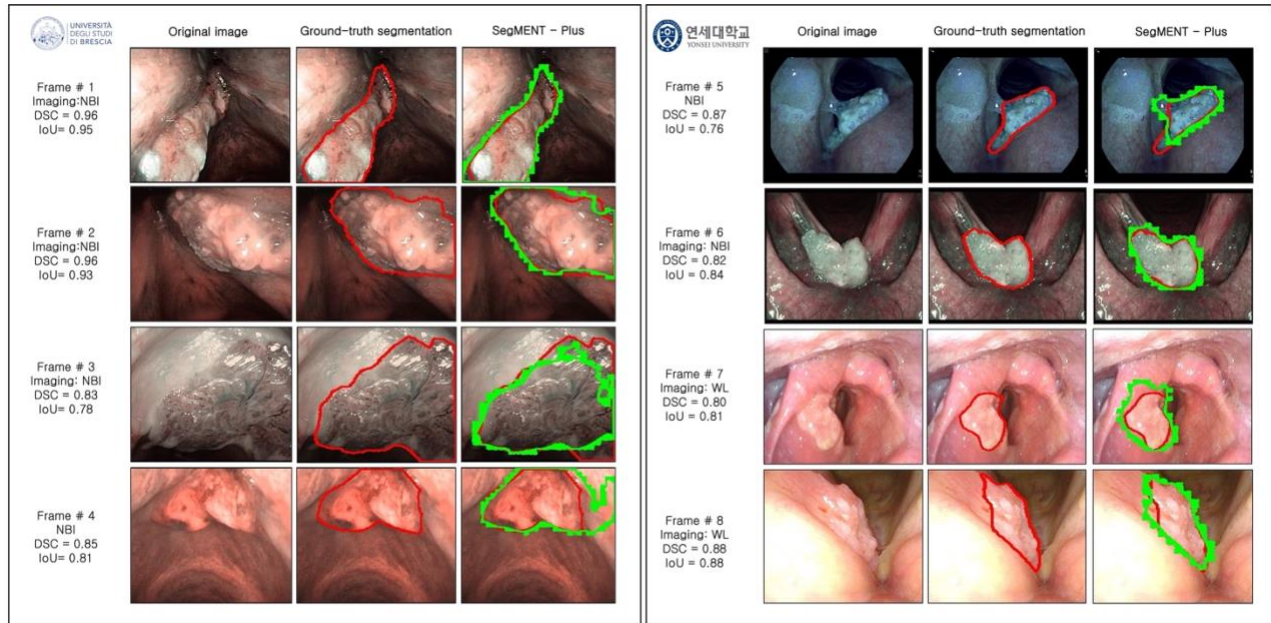


Figure 28. Examples of SegMENT-Plus automatic segmentation of laryngeal carcinoma. The panel on the left reports four examples from the University of Brescia dataset. The panel on the right reports four examples from the Yonsei University dataset. Red segmentations represent the ground truth provided by the experts, while green annotations are the areas predicted by the model. DSC, dice similarity coefficient; IoU, intersection over union; WL, white light; NBI, Narrow Band Imaging.

Comparison with human physicians

The 100 images subcohort selected from the University of Genova test set was independently segmented by a junior resident, a senior resident, and *SegMENT-Plus*. Their results, in comparisons with the ground truth annotations, were the following: senior resident, DSC = 0.91 (0.84-0.95), IoU= 0.91 (0.84-0.94); junior resident, DSC = 0.88 (0.79-0.93), IoU= 0.88 (0.81-0.92); *SegMENT-Plus* DSC = 0.89 (0.81-0.93), IoU= 0.89 (0.83-0.92). **Figure 29** shows the quantitative performance achieved by the model and the residents. No significant differences were observed between the segmentation performances of the three groups (DSC, $p = 0.41$; IoU, $p = 0.33$). Qualitative examples of the masks of the three raters and the model are provided in **Figure 30**.

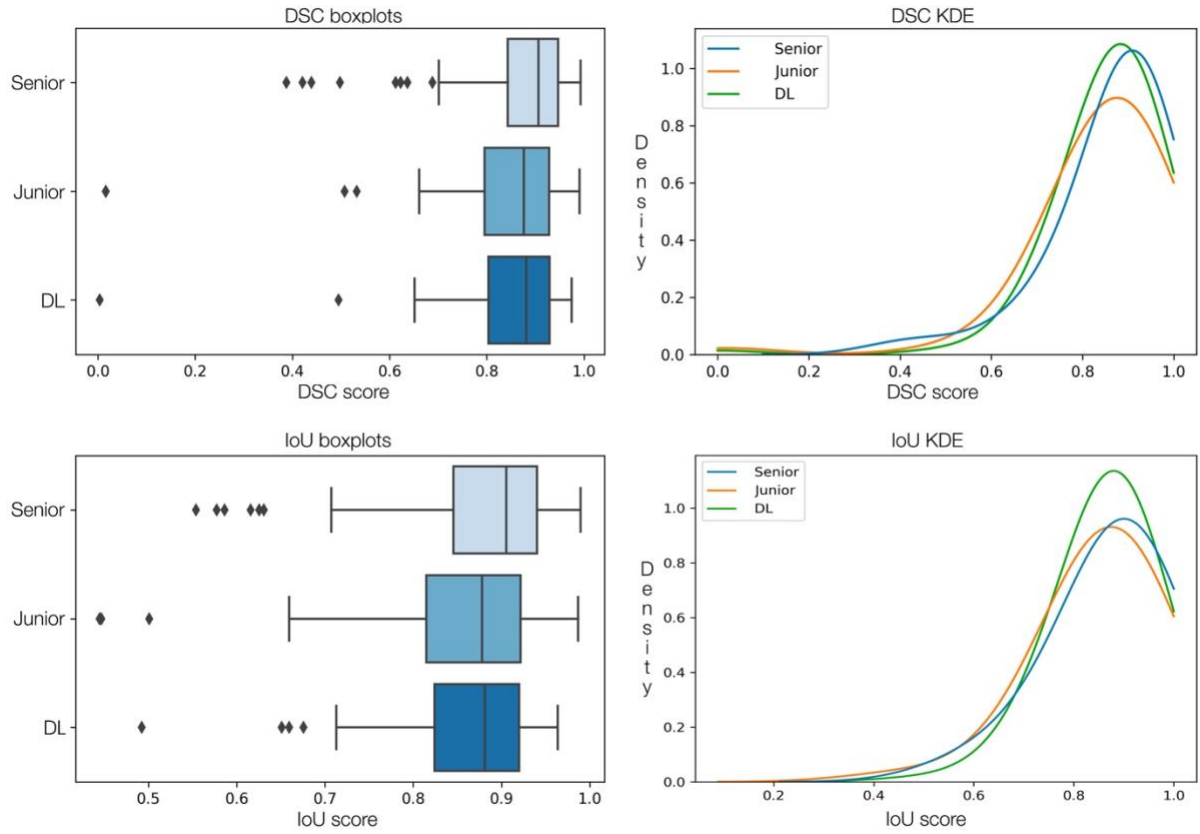


Figure 29. The performance results of residents and SegMENT-Plus are represented by boxplots and kernel density estimates (KDE) curves. In the left figures, the vertical bar inside the boxes represents the median value, while the box represents 50% of the distribution within the 1st and the 3rd quartiles. DSC, dice similarity coefficient; IoU, Intersection over union; DL, deep learning model.

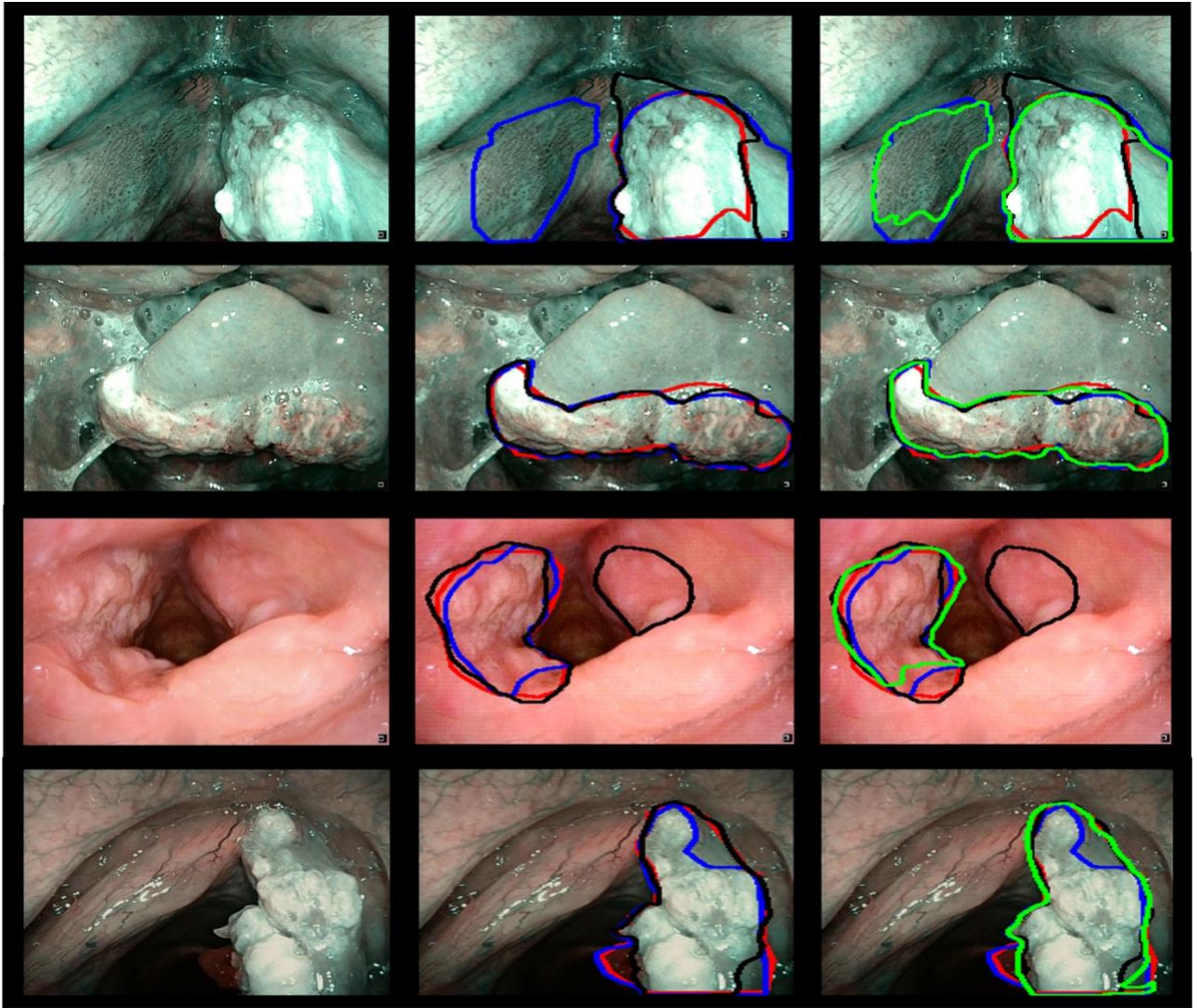


Figure 30. Examples of the segmentations performed by the human physicians and SegMENT-Plus on the University of Genova test dataset. The column on the left shows the original frames, the column in the middle reports the annotation performed by the human physicians, and in the column on the right the DL model prediction is added. Blue = ground truth; Red = senior resident; black = junior resident; green = SegMENT-Plus.

Real-time video performance

The model was applied to five complete preoperative videolaryngoscopies to assess real-time applicability. SegMENT-Plus successfully delineated tumor boundaries continuously across frames, maintaining smooth contour transitions without temporal instability or flickering. The mean inference time per frame was 0.041, equivalent to 25 FPS, confirming true real-time

capability (**Table 10**). Examples of the original video frames and the corresponding ones processed by the DL model are shown in **Figure 31**.

Table 10. Testing videos characteristics and computation time after applying SegMENT-Plus.

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

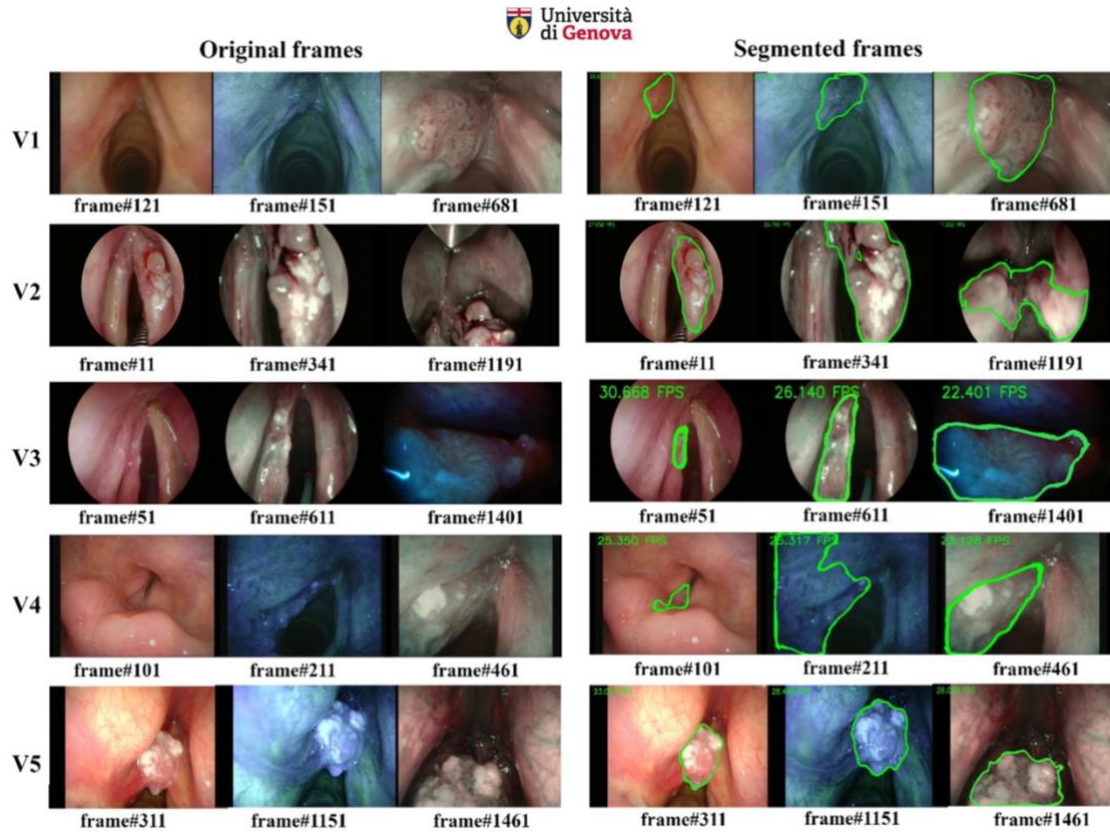


Figure 31. Testing video frames extracted from five videolaryngoscopies. Each row represents a different video: the pictures in the panel on the left belong to the original video, while the pictures in the panel on the right are the same frames after the prediction of the deep learning model.

Study interpretation

Overall, the SegMENT-Plus model architecture provided precise and consistent delineation of tumor boundaries on both WL and NBI images. The model maintained high performance across internal and external datasets and achieved real-time inference speeds, demonstrating both robustness and clinical applicability. Its ability to reproduce otolaryngology resident-level segmentation suggests potential value for surgical planning, intraoperative guidance, and standardized tumor mapping in laryngeal surgery.

LARYNGEAL MOTILITY ASSESSMENT

Dataset overview

The retrospective dataset collected at the two centers comprised 155 in-office video laryngoscopies, each one belonging to a different patient. After the revision by the panel of experienced otolaryngologists, they were divided into three categories: normal motility ($n = 37$), unilateral vocal fold paralysis ($n = 68$), or bilateral paralysis ($n = 50$). On average, 4.21 trials were extracted per video, giving a larger dataset comprising 404 trials (135 referring to subjects with normal healthy motility, 198 referring to patients with unilateral paralysis, and 71 referring to patients with bilateral paralysis).

Binary classification performance

The binary classification model was designed to distinguish between healthy subjects and those affected by laryngeal paralysis (including both unilateral and bilateral cases). After automatic trial identification and feature extraction, the model achieved a median balanced accuracy of 0.73 across trials and 0.85 at the patient level (**Table 11**). Precision, sensitivity, and F1-score values at the patient level were 0.83, 0.85, and 0.84, respectively, demonstrating high reliability in discriminating normal from pathological motility patterns. **Figure 32** illustrates the corresponding confusion matrices, highlighting that most misclassifications occurred at the trial level, whereas patient-level aggregation markedly improved accuracy.

Table 11. Binary classification performance (healthy vs paralyzed).

Median $\pm$ Median Absolute Value of the performance obtained by the model across the 5 folds. Performances are shown both for trials and patients, where the predictions on several trials related to specific subjects are aggregated.

Metric	Trial	Patient
Precision	0.73 ± 0.03	0.83 ± 0.05
Sensitivity	0.73 ± 0.01	0.85 ± 0.07
F1-score	0.73 ± 0.01	0.84 ± 0.05
Balanced Accuracy	0.73 ± 0.01	0.85 ± 0.07

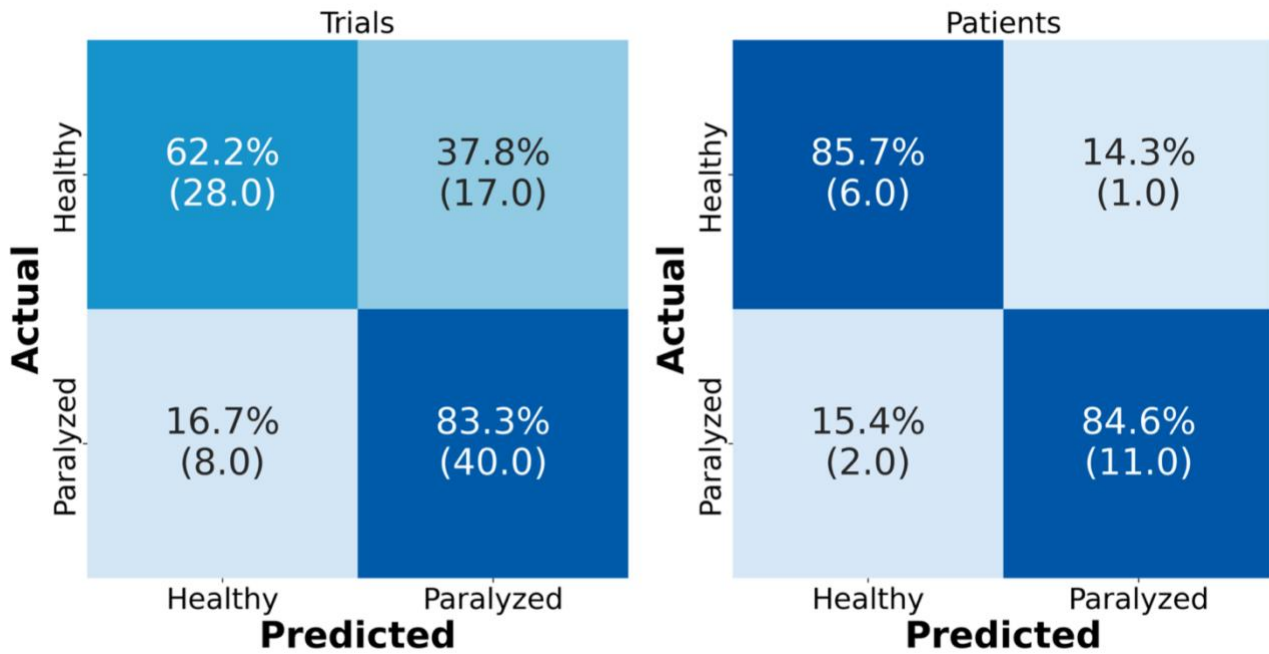


Figure 32. Binary classification (healthy vs paralyzed) confusion matrices. Median confusion matrices obtained by the model across folds. Here, each entry shows the percentage of subjects actually categorized in the class characterizing its row and predicted as the class assigned to its column. The median is computed between the different percentages obtained across folds.

The improvement in balanced accuracy ($p < 0.0001$) is clinically relevant, as averaging predictions across multiple trials for the same patient reduces the influence of local noise and short-term movement artifacts commonly seen during endoscopy. The model's robust performance at the patient level therefore indicates potential for clinical translation in objective paralysis recognition.

Multi-class classification performance

The same architecture was then applied to a more complex diagnostic task i.e., classifying subjects into three categories: healthy, unilateral paralysis, and bilateral paralysis. The model achieved a balanced accuracy of 0.70 across trials and 0.83 at the patient level (**Table 12**). Patient-level sensitivity was 0.86 for healthy cases, 0.62 for unilateral paralysis, and 1.00 for bilateral paralysis. These results indicate that the model effectively distinguished between complete and partial loss of motion, with the strongest discrimination achieved for bilateral immobility. The moderate sensitivity

for unilateral paralysis reflects the greater diagnostic complexity of asymmetric motion patterns, which often require more nuanced interpretation even for experienced clinicians.

Table 12. Multi-class classification performance (healthy, unilateral paralysis, bilateral paralysis). Median $\pm$ Median Absolute Value of the performance obtained by the model across the 5 folds. Performances are shown both for trials and patients, where the predictions on several trials related to specific subjects are aggregated.

Metric	Trial			Patient		
	Healthy	Unilateral	Bilateral	Healthy (Patients)	Unilateral (Patients)	Bilateral (Patients)
Precision	0.78 $\pm$ 0.02	0.59 $\pm$ 0.01	0.55 $\pm$ 0.05	0.75 $\pm$ 0.05	0.83 $\pm$ 0.08	0.83 $\pm$ 0.05
Sensitivity	0.64 $\pm$ 0.04	0.67 $\pm$ 0.04	0.89 $\pm$ 0.12	0.86 $\pm$ 0.00	0.62 $\pm$ 0.08	1.00 $\pm$ 0.10
F1-score	0.69 $\pm$ 0.02	0.60 $\pm$ 0.02	0.67 $\pm$ 0.07	0.80 $\pm$ 0.03	0.71 $\pm$ 0.07	0.91 $\pm$ 0.07
Balanced Accuracy	0.70 $\pm$ 0.03			0.83 $\pm$ 0.06		

The confusion matrices in **Figure 33** confirm these trends, showing clear differentiation between healthy and bilaterally paralyzed cases, with some overlap between healthy and unilateral paralysis predictions. Nevertheless, patient-level aggregation again improved performance consistency, underscoring the robustness of the classification framework across diagnostic categories.

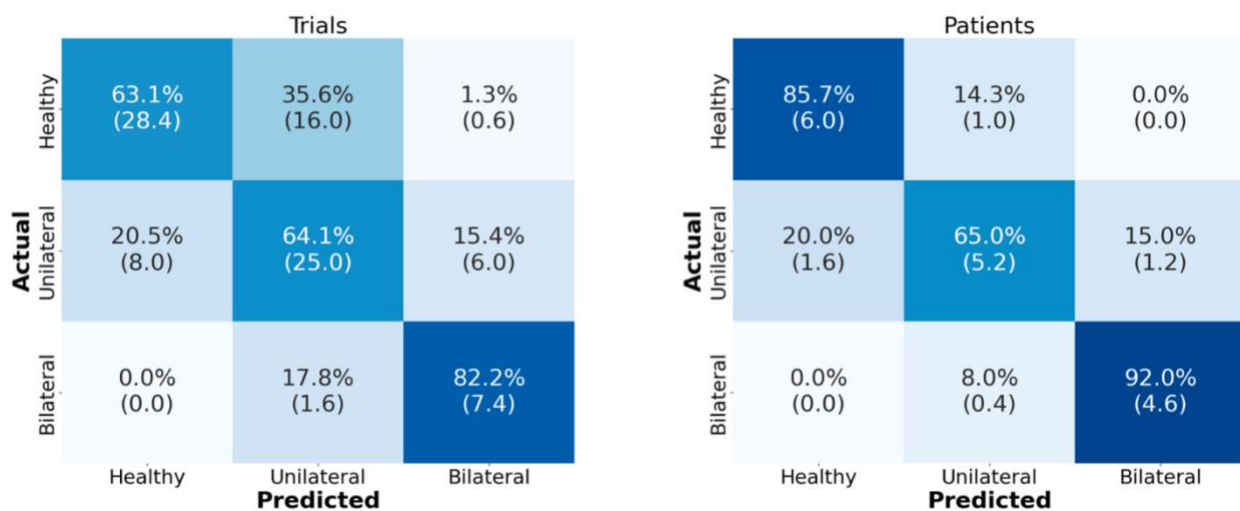


Figure 33. Median confusion matrices obtained by the multi-class classification model (healthy, unilateral paralysis, bilateral paralysis) across folds. Here, each entry shows the percentage of subjects actually categorized in the class characterizing its row and predicted as the class assigned to its column. The median is computed between the different percentages obtained across folds.

Ablation study and feature relevance

An ablation analysis using the Shapley value approach was conducted to quantify the contribution of each feature group to model performance (**Figure 34**). In this analysis, the model was repeatedly retrained after selectively removing different groups of features. The impact of each feature group on the model’s performance was then quantified using Shapley values, which measure how much each group contributes to the model’s correct classifications (a higher Shapley value means that the feature group provided more useful information to the model).

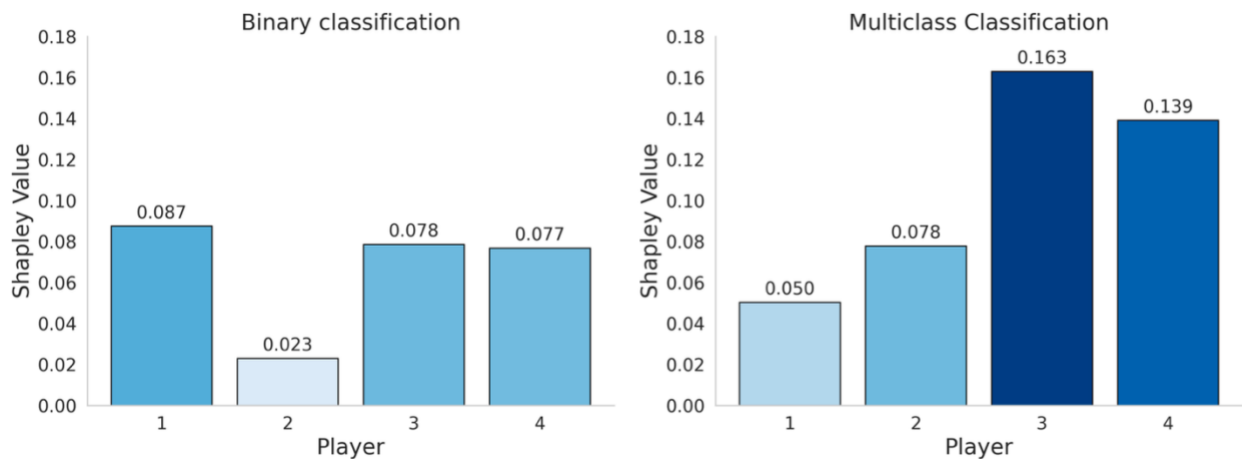


Figure 34. Shapley value analysis illustrating the contribution of each feature group to model performance. In the binary task (left), overall glottic motion was most informative, while in the multiclass task (right), motion variability and inter-fold coordination provided the strongest diagnostic cues.

The bar charts summarize the results of this analysis. In both graphs, the four “players” represent distinct groups of features used by the model:

1. Image-based feature – the difference map between frames of maximal abduction and adduction, capturing the overall degree of movement of the vocal folds.
2. Anterior Glottic Angle (AGA) features – the minimum and maximum angles measured during the glottic cycle, representing the range of opening and closing.
3. Variability features – descriptors that quantify how much the AGA changes over time, reflecting the regularity and completeness of vocal fold motion.

4. Coordination features – measures such as correlation and Granger causality, describing how synchronized or asymmetric the movements of the left and right vocal folds are.

In binary classification, the image-based feature—derived from the subtraction of frames at maximal abduction and adduction—emerged as the most influential factor. This finding is consistent with clinical expectations, as this feature effectively summarizes the total degree of vocal fold movement and highlights immobile or restricted areas that characterize paralysis.

For multi-class classification, the most relevant features were those describing temporal coordination between the two vocal folds, including correlation and Granger causality measures. These features captured asymmetry and delay between left and right fold motion, which are hallmarks of unilateral paralysis. Features traditionally reported in the literature, such as the maximum and minimum AGA values, remained relevant but were less dominant, suggesting that the integration of newer, dynamic parameters provided additional discriminative power.

Study interpretation

Overall, the classification framework achieved state-of-the-art results in the automated analysis of laryngeal motility. Both binary and multi-class models demonstrated similar strong diagnostic performance, particularly at the patient level, despite multiclass classification is a more demanding task. The inclusion of multimodal features enhanced robustness against variability in video quality and endoscopic positioning. The combination of quantitative motion descriptors and image-based features resulted in a clinically interpretable and objective approach for assessing laryngeal paralysis, with potential applications in real-time or retrospective diagnostic support.

DISCUSSION

General considerations

This thesis investigated the role of artificial intelligence in the endoscopic evaluation of laryngeal pathology, encompassing the full spectrum of clinically relevant computer vision tasks: frame selection, lesion detection, diagnostic classification, tumor segmentation, and motility assessment. Across these five studies, a coherent methodological framework was adopted, characterized by multicenter image collection, rigorous data curation with expert annotation, and validation on external datasets. This structured approach ensured that each experiment contributed a distinct yet complementary step toward a comprehensive and clinically applicable ecosystem of AI tools for laryngeal endoscopy. Together, these studies demonstrate that DL models, when developed and validated under appropriate methodological standards, can reach or human-level performance in several diagnostic subtasks while operating at real-time speed.

The collective findings highlight three major advances. First, the integration of AI into laryngeal imaging workflows can enhance objectivity by transforming subjective visual judgments into reproducible quantitative measures. Tasks such as assessing image quality, identifying suspicious mucosal areas, or evaluating vocal fold motion are traditionally dependent on clinician experience and subject to variability. The proposed models reduce this variability by providing standardized, data-driven outputs that correlate with expert assessment. Second, AI can improve efficiency in both data management and clinical practice. Informative frame selection, for example, reduces the burden of manual video review, while automated detection and classification tools accelerate diagnostic interpretation and documentation. Third, the inclusion of multicenter and multimodal data demonstrated that these systems can achieve robust generalizability, performing reliably across imaging modalities (WL and NBI), clinical settings, and devices. This aspect is crucial for translation to real-world use, where heterogeneity of acquisition conditions often limits the reproducibility of AI models²¹.

From a methodological perspective, the thesis followed best practices for AI research in medicine, including patient-level data partitioning to prevent leakage, separate internal and external validation, and use of explainability tools such as class activation maps, Grad-CAM, and Shapley methodology^{72,73}. These strategies ensured transparency and interpretability, which are essential for clinician trust and regulatory compliance⁷⁴. The use of self-supervised pretraining, transfer learning, and lightweight architectures further demonstrated that clinically relevant performance can be achieved even with moderately sized medical datasets, provided that data quality and annotation consistency are prioritized.

Clinically, the five experiments collectively address key challenges in laryngeal oncology and functional assessment. In oncology, early detection, accurate risk stratification, and precise delineation of lesion margins directly influence prognosis and treatment outcomes. In functional disorders, objective quantification of motility offers the potential to complement or replace subjective grading systems. By demonstrating that AI models can perform these tasks automatically and in real time, this thesis supports the integration of computer vision as a decision-support tool rather than a replacement for clinical judgment.

This thesis also gives valuable insights into how AI can offer numerous advantages in the context of endoscopy for diagnosing HNSCC. AI algorithms can significantly enhance the accuracy of lesion detection by analyzing endoscopic images in real-time, identifying subtle abnormalities that might be missed by human eyes. This capability can standardize and improve diagnostic accuracy, reducing variability due to operator skill and experience. When used for differentiating between benign and malignant lesions, AI can provide decision support to clinicians and potentially reduce the need for unnecessary biopsies as it has happened for gastrointestinal endoscopy with a so-called “diagnose-and-leave” strategy⁷⁵.

Integrating AI into endoscopic procedures means that even healthcare providers with less specialized training, such as nurses, could effectively enhance the use of these systems. This democratization of

technology is particularly beneficial for massive screening programs in rural or underserved areas, where access to specialized medical professionals is limited. AI-powered endoscopy systems could facilitate early detection of HNSCC, leading to prompt referral for treatment and significantly improving patient outcomes. These systems can be cost-effective, making advanced diagnostic capabilities accessible to low-cost healthcare providers and enhancing the overall reach and impact of cancer screening programs. Moreover, recently, modern portable video-endoscopy digital equipment has become more common and affordable, allowing broader access to this essential diagnostic tool, together with the digitalization of the images⁷⁶. These two technological advances if combined, could dramatically change the scenario for HNSCC.

Several real-world initiatives have already shown how AI-based screening can improve healthcare access, particularly in resource-limited environments. For instance, a recent *JAMA* study reported that individuals with no formal training in obstetric ultrasound were able to estimate gestational age with accuracy comparable to experienced sonographers when assisted by a low-cost, point-of-care AI tool⁷⁷. Another example comes from India, where Google and the Aravind Eye Hospital developed a deep learning system for diabetic retinopathy detection that performs at the level of trained ophthalmologists. This technology has been integrated into a large-scale screening network across 71 vision centers in rural Tamil Nadu, where technicians capture retinal images and transmit them digitally for remote specialist review⁷⁸. Similar equity-oriented efforts have also emerged in gastrointestinal endoscopy: Medtronic, in partnership with the American Society for Gastrointestinal Endoscopy and with support from Amazon Web Services, launched a Health Equity Assistance Program aimed at delivering AI-assisted colonoscopy systems to underserved populations⁷⁹.

Early detection remains a key determinant of survival in HNSCC. Detecting tumors at stage I–II can increase five-year survival from below 50% to approximately 70–90%. If AI technology could improve early detection rates by even a conservative estimate of 20%, and assuming an average survival of 80% for early-stage disease according to the SEER database⁸⁰, based on the predicted

number of the International Agency for Research on Cancer (IARC), up to 300,000 patients could potentially experience significantly improved outcomes by 2050⁵. In addition to its clinical impact, earlier diagnosis enabled by AI could yield major economic benefits. Timely treatment typically reduces the need for extensive surgical procedures, adjuvant therapies, and long hospitalizations, all of which are associated with higher healthcare costs. Patients diagnosed with localized disease generally incur substantially lower treatment expenditures than those presenting with advanced tumors managed with surgery and (chemo)radiotherapy⁸¹.

Financial burden is another critical consideration. Individuals with HNSCC often face significant out-of-pocket expenses that can cause financial hardship and negatively influence recovery and quality of life⁸². These costs include medications, nutritional supplements, travel to treatment centers, and dental rehabilitation. Depending on the healthcare system, treatment of advanced cancer can cost up to nine times more than that of early-stage disease, making the economic advantage of early detection even more evident⁸³. This issue is particularly severe in low- and middle-income countries with limited insurance coverage. In India, for instance, where much of healthcare expenditure is paid directly by patients, the cost of medicines alone frequently results in household impoverishment and catastrophic health spending⁸⁴.

Informative frames selection

The first study in this thesis addressed one of the most fundamental yet often overlooked challenges in laryngeal endoscopy: the presence of large numbers of non-informative frames within recorded videos. These include frames affected by blur, poor illumination, saliva artifacts, or transient obstructions, all of which hinder diagnostic interpretation and complicate the use of endoscopic data for research or AI training. The experiment demonstrated that DL can be effectively applied to automate this task, providing reliable and real-time discrimination between informative and non-informative content. This capability represents an important step toward standardized data quality control and efficient downstream analysis in both clinical and research settings.

The proposed system achieved high performance on both internal and external datasets, with F1-scores ranging between 93% and 96%. This high accuracy is explicable considering this is intuitively the easiest one among the classification tasks. This performance is consistent with previous literature on informative frame classification in endoscopic imaging. Early studies by Moccia et al.⁴⁵ (2018) and Patrini et al.⁴⁴ first established the feasibility of using traditional and deep learning models for this task, while Yao et al.⁴⁶ extended this concept to a large laryngoscopic dataset of more than 22,000 images, achieving F1-scores between 0.69 and 0.92. The results obtained in this thesis confirmed and expanded upon these findings by validating the approach across two independent institutions and by demonstrating its real-time application to continuous video streams, simulating real clinical workflow.

Beyond accuracy, the model introduced a novel real-time visualization interface capable of displaying an informative or non-informative label for each frame in less than 0.02 seconds. This enabled continuous quality monitoring during endoscopic examination and allowed the calculation of the Video Informativeness Level (VIL), a global score summarizing the proportion of diagnostically useful frames within a video. In the study, the mean VIL did not exceed 60%, underscoring that a substantial portion of laryngoscopic recordings is redundant or diagnostically limited. Such

quantification can guide future optimization of recording protocols and inform storage or annotation strategies.

From a broader perspective, informative frame selection serves as a foundation for all subsequent AI-based endoscopic applications. Several studies have shown that the accuracy and stability of detection, segmentation, and classification algorithms depend heavily on the quality of input data^{30,55,64}. By filtering out uninformative content, this system ensures that subsequent models are trained and tested on high-quality and diagnostically relevant images, thereby enhancing robustness and generalizability. In clinical practice, this process can also assist less experienced clinicians by automatically highlighting the most relevant video segments, reducing the time required for manual review and improving the consistency of endoscopic documentation.

Nevertheless, certain limitations should be acknowledged. The distinction between informative and non-informative frames remains partly subjective, as no universally accepted criteria exist to define optimal diagnostic visibility. Occasional misclassifications occurred in borderline cases, such as mildly blurred or partially obstructed frames that still retained clinical value. Future work should aim to standardize annotation protocols and assess model performance across different endoscopic systems, light sources, and anatomical regions.

In conclusion, this study demonstrated that automatic informative frame classification in laryngeal endoscopy is feasible, accurate, and generalizable, even across centers using different imaging platforms. Its ability to operate in real time and to provide objective video-quality metrics highlights its potential integration as the first stage in an AI-assisted diagnostic pipeline. Within the broader context of this thesis, informative frame selection represents the essential foundation upon which reliable lesion detection, classification, and segmentation systems can be built, ensuring that AI models are trained on consistently high-quality, clinically meaningful data.

Laryngeal tumors detection

The second study of this thesis focused on the automatic detection of laryngeal tumors in endoscopic images, addressing the clinical need for early and objective recognition of suspicious mucosal lesions during examination. In routine laryngoscopy, lesion identification relies heavily on the clinician's experience, and subtle neoplastic changes—particularly small or early-stage tumors—can be missed due to limited visibility, operator fatigue, or distracting anatomic and optical factors. The study demonstrated that DL can be effectively applied to assist clinicians by automatically localizing lesions in both white light WL and narrow band imaging NBI modalities, providing real-time diagnostic support.

The developed system, termed SRE-YOLO (Super-Resolution Enhanced YOLO), was based on the YOLOv8 detection framework but integrated a super-resolution preprocessing module to enhance the visibility of small lesions and improve the quality of fine vascular and surface details. This design choice was particularly relevant to laryngeal tumors, as early lesions are often subtle and occupy a small proportion of the image. In terms of performance, the model achieved high accuracy across datasets, with a mean average precision (AP) of 0.83 on internal testing and 0.82 on external validation, indicating strong generalizability. When comparing lesion sizes, SRE-YOLO maintained stable detection for both small (<32 px) and medium (32–96 px) lesions, outperforming standard architectures. These findings are clinically relevant, as early-stage or minimally invasive tumors frequently fall within this small-lesion category and are often the most challenging to identify during standard endoscopy. The study also compared model performance across imaging modalities, observing slightly higher accuracy for NBI (AP = 0.88) than for WL (AP= 0.86). This aligns with existing evidence demonstrating that NBI provides superior visualization of microvascular patterns associated with neoplastic transformation^{12–14}. The model's ability to maintain consistent performance across both modalities suggests potential for multimodal integration, where AI could simultaneously process and compare WL and NBI frames to maximize diagnostic yield.

The study's results are in line with recent applications of AI for lesion detection in gastroesophageal endoscopy and colonoscopy, where YOLO-based networks have demonstrated high sensitivity and real-time usability^{85,86}. However, while such models have become increasingly common in digestive endoscopy, their use in laryngology has been limited, making our group the first one to develop a real-time implementation tailored to the UADT. By confirming that high accuracy and inference speed can be achieved even in this anatomically complex region, the study extends the application of AI-based detection to a new clinical domain.

From a clinical standpoint, one of the key advantages of SRE-YOLO was its computational efficiency and real-time feasibility. The model achieved an inference speed of approximately 81 FPS on standard GPU hardware, far exceeding the 25–30 FPS threshold of clinical video systems. This ensures that detection results can be displayed live during endoscopic examination, highlighting suspicious areas without delay or workflow disruption. In practice, this capability could provide instant visual feedback to the examiner, improving lesion recognition, especially for non-specialist clinicians or in high-volume screening contexts. Automated lesion detection could serve as a “second observer” during routine endoscopy, supporting diagnostic vigilance and standardizing examination quality across centers. This may be particularly valuable in training environments or resource-limited settings, where specialized expertise in NBI interpretation is not always available. Moreover, by improving early lesion detection, such systems could contribute to reducing diagnostic delays and improving oncological outcomes.

Nevertheless, certain limitations should be recognized. The dataset was composed of static frames rather than continuous videos, which may not fully capture the variability encountered during real clinical procedures. Future work should therefore focus on extending the model to full video input and evaluating its robustness under different recording conditions and image qualities. Additionally, no clinical raters were asked to confront the model in a detection challenge so that the clinical implications insights of this study remain limited.

In summary, the SRE-YOLO model demonstrated that AI-based detection of laryngeal tumors is accurate, fast, and generalizable across imaging modalities and centers. By integrating a super-resolution module, the system improved sensitivity for small and early lesions, achieving clinically meaningful performance in real time. Within the broader context of this thesis, this study provides an essential step toward AI-assisted visual diagnosis, establishing the foundation for subsequent experiments on classification and segmentation of laryngeal tumors.

Laryngeal tumors diagnosis

The third study of this thesis focused on the diagnostic classification of laryngeal lesions using a CADx model based on artificial intelligence. The goal was to develop a system capable of distinguishing HR from low-risk LR lesions during diagnostic flexible laryngoscopy, assisting clinicians in identifying malignant or pre-malignant changes. This study represents the largest multicenter validation to date of a CADx model for laryngeal disease, integrating data from five countries and systematically comparing AI performance with clinicians and large language models. This study addressed a central limitation of previous works in the field: the scarcity of multicentric data and the absence of robust external validation. While earlier studies on AI-assisted laryngeal classification demonstrated promising diagnostic accuracy, most were conducted on small, single-center datasets, often without independent validation or clinical comparison^{21,87}. The present work overcame these limitations by incorporating more than 20,000 images collected across tertiary referral centers in Italy, Spain, India, Greece, and China, ensuring diversity in imaging conditions and patient populations.

The developed CADx model was based on the EfficientNetV2S architecture, optimized through an in-domain pretraining strategy using an auxiliary dataset automatically extracted from clinical videos. This pretraining step leveraged the previously validated models for informative frame selection and lesion detection in this thesis to ensure that the classification algorithm was automatically trained on clear, high-quality, and properly localized lesion regions. Such an integrated workflow simulated a realistic diagnostic pipeline where AI sequentially filters, localizes, and classifies the lesion. A critical methodological contribution of this study was its rigorous approach to avoiding overfitting. Images from the same patient were never included in both training and testing sets, and independent validation was performed using datasets from different geographical regions and imaging systems. This strategy directly addressed a recurrent issue in AI for medical imaging, where models often perform well only on internal data but fail to generalize to new clinical environments. Furthermore,

the inclusion of diverse populations helped minimize ethnicity-related bias, an important step toward equitable AI deployment.

The model achieved high diagnostic accuracy on both internal and external datasets, with accuracy and AUC values of 0.90/0.89 on the internal test set, 0.85/0.85 on the Greek dataset, and 0.88/0.88 on the Spanish dataset. Importantly, on the Spanish dataset, the CADx model's diagnostic performance was comparable to that of otolaryngologists and expert laryngologists (accuracy 0.88–0.93), and significantly higher than that of general practitioners and ChatGPT-4o (accuracy 0.68–0.69). This finding demonstrated not only the robustness of the CADx model but also its potential to standardize diagnostic accuracy across varying levels of expertise.

From a clinical perspective, these results underscore the potential of AI-based diagnosis to function as an assistive “second observer” during laryngoscopy, particularly for non-specialists or in resource-limited settings or for screening campaigns. By distinguishing HR from LR lesions in real time, AI could help prioritize cases requiring biopsy or referral, thus reducing diagnostic delays and unnecessary interventions. This way, AI-assisted endoscopy may help democratize cancer detection by improving early identification and care access worldwide, especially where trained laryngologists are scarce. In the broader context of this thesis, this study represents the transition from automated visual recognition to diagnostic interpretation, bridging the gap between computer vision and clinical decision-making. The CADx model not only demonstrated technical reliability but also achieved a level of interpretive performance comparable to human experts, indicating that AI systems may eventually complement clinical judgment in routine diagnostic workflows.

Certain limitations should be acknowledged, including the retrospective design and the focus on static images rather than live video analysis. Moreover, while the CADx model was conceived as part of a multi-step AI pipeline, its clinical benefits have yet to be fully validated prospectively. For this reason, a multicenter clinical trial is currently underway to evaluate the real-world implementation of this framework.

In conclusion, the laryngeal tumor diagnosis study demonstrated that AI can achieve expert-level accuracy in classifying laryngeal lesions across multiple institutions and populations. By validating the model externally and benchmarking it against clinicians, this work established a new standard for the clinical evaluation of AI tools in otolaryngology. Within this thesis, it represents the culmination of a methodological progression—from data quality optimization and lesion localization to diagnostic inference—providing a strong foundation for the clinical integration of AI-assisted optical biopsy systems in laryngology.

Laryngeal tumors segmentation

The fourth study of this thesis investigated the use of DL for the automatic segmentation of laryngeal cancer, a task of major clinical relevance for improving surgical precision and standardizing the delineation of tumor margins. The study introduced an improved DL model, SegMENT-Plus, designed to provide accurate, real-time delineation of laryngeal cancer boundaries on both WL and NBI endoscopic images and videos.

The accurate definition of surgical margins is critical in transoral minimally invasive surgery, as positive or close margins are still reported in up to 40% of cases despite advances in technique^{88–90}. NBI has proven to improve intraoperative margin visualization^{91,92}, but its interpretation remains highly operator-dependent and requires a long learning curve⁹³. The introduction of automated computer vision systems capable of recognizing and delineating tumor margins may therefore enhance surgical accuracy and mitigate inter-operator variability, offering consistent, objective visual guidance.

SegMENT-Plus builds upon a previous pilot version⁵⁷ which demonstrated the feasibility of tumor segmentation on a smaller dataset. In this expanded study, the model architecture was refined to improve computational efficiency while being trained on a substantially larger dataset of 3,933 endoscopic images obtained from 557 patients at the University of Genoa. Two additional external datasets—from the University of Brescia (Italy) and Yonsei University (Republic of Korea)—were used for independent validation. The model achieved high segmentation accuracy across all datasets, with median Dice similarity coefficients (DSC) ranging from 0.81 to 0.83 and Intersection over Union (IoU) scores between 0.82 and 0.84, demonstrating excellent generalization capability. The model processed images at a median speed of 25–26 frames per second, compatible with real-time endoscopic visualization.

Importantly, SegMENT-Plus performed equally well on both WL and NBI modalities, suggesting that it was able to interpret complex vascular patterns and mucosal textures independently of the

optical filter used. This ability to integrate information from different imaging modalities is particularly relevant for laryngeal surgery, where the margins of tumors can vary markedly depending on the vascular enhancement. When compared with human performance, SegMENT-Plus achieved segmentation accuracy equivalent to that of senior otolaryngology residents (DSC 0.89 vs. 0.91; $p > 0.05$), demonstrating that AI can approximate the competence of clinicians with several years of training.

From a clinical perspective, these findings suggest that automatic segmentation may play multiple roles in the management of laryngeal cancer. During surgery, such models could provide real-time margin mapping to assist surgeons in identifying tumor boundaries more accurately, reducing the risk of residual disease and positive margins. In the diagnostic and follow-up setting, segmentation could facilitate objective monitoring of lesion progression or treatment response by quantifying mucosal changes over time⁹⁴. Moreover, automated delineation may accelerate image annotation for large-scale datasets, thus supporting future research and training of AI systems.

When compared with the broader literature, this work represents one of the most comprehensive applications of AI to laryngeal segmentation. Ji et al. achieved comparable results in the segmentation of laryngeal leukoplakia (DSC 0.78, IoU 0.83) on a smaller dataset of 649 WL images⁶¹. However, leukoplakias are morphologically simpler and more contrast-rich than invasive carcinomas, which makes the current study's results even more notable given the visual heterogeneity of laryngeal cancer. By validating its model on multi-institutional and multimodal datasets, this work strengthens the evidence that AI-driven segmentation is generalizable and clinically reliable across different acquisition systems and populations.

Nevertheless, several limitations should be acknowledged. The model was trained on retrospective data, and the subjective nature of manual annotations introduces variability in the definition of tumor borders. Furthermore, while real-time processing was successfully simulated, prospective validation

during live surgery remains to be conducted. Future developments should include larger annotated datasets and testing on other upper aerodigestive tract sites to enhance generalization.

In conclusion, SegMENT-Plus demonstrated robust, accurate, and real-time delineation of laryngeal tumors, achieving performance comparable to trained clinicians and maintaining consistency across multiple institutions. Within the framework of this thesis, this study represents a pivotal step toward the translation of AI into intraoperative support tools, offering the potential to improve surgical precision, reduce variability in margin assessment, and contribute to more standardized and data-driven surgical oncology.

Laryngeal motility assessment

The fifth and final study of this thesis explored the use of artificial intelligence for the automatic assessment of laryngeal motility, with a specific focus on diagnosing vocal fold paralysis. This condition remains one of the most frequent and clinically challenging laryngeal disorders, often requiring expert interpretation of endoscopic videos. Traditional evaluation is highly subjective and suffers from significant inter- and intra-observer variability, with studies reporting low concordance among clinicians, particularly when assessing subtle cases of impaired movement⁹⁵. This study proposed a quantitative and automated framework for laryngeal motility evaluation, aiming to provide a more objective, reproducible, and efficient diagnostic tool.

The developed pipeline combined computer vision and machine learning to analyze 155 video-laryngoscopies collected across two centers: the Hospital Clínic de Barcelona (Spain) and the University Hospital of Padua (Italy). Patients were categorized into three groups: healthy ($n = 37$), unilateral paralysis ($n = 68$), and bilateral paralysis ($n = 50$). The analysis relied first on a vocal tracking model, which automatically tracked the movement of the vocal folds and quantified their motion through the Anterior Glottic Angle (AGA), a measure of the angle between the free edges of the vocal folds. This metric is particularly valuable because it remains stable despite minor changes in camera orientation or lighting, allowing for reliable quantification of the degree of vocal fold mobility.

To focus on the most informative moments of each examination, a ML model was applied to automatically identify video segments containing the full cycle of vocal fold opening and closing. This process filtered out irrelevant frames, enhanced data quality, and generated multiple usable “trials” per patient for analysis. The model then extracted several features from the video-trials describing both the symmetry and coordination of vocal fold movement. The classification system combined a CNN to capture pixel variations during movements, allowing a multimodal understanding of laryngeal motion. Two tasks were addressed: a binary classification (healthy vs paralyzed) and a

multiclass classification (healthy vs unilateral vs bilateral paralysis). The binary model achieved a balanced accuracy of 0.85 at the patient level, while the multiclass model reached 0.83, confirming that the system could distinguish between different degrees of motility impairment with high reliability. Notably, this represents one of the first successful attempts to automatically differentiate unilateral and bilateral paralysis—a clinically significant distinction that guides both prognosis and management.

An ablation study based on Shapley values provided interpretability to the model's predictions, identifying which features most influenced the diagnostic outcome. Image-derived motion features contributed the most to distinguishing healthy from paralyzed patients, while temporal asymmetry measures (such as correlation and causality between left and right movements) were most relevant for identifying unilateral paralysis. These findings confirmed that the system captured clinically meaningful patterns aligned with human expert reasoning.

This study marks an important step toward objective motility assessment. By quantifying laryngeal movement and eliminating subjective interpretation, AI could substantially reduce inter-rater variability and standardize evaluation across institutions. Moreover, automated motion analysis could enhance clinical efficiency, allowing for rapid screening of large video archives or integration into live examinations as a “second observer.” Similar AI-based motion analysis tools have already shown clinical utility in other specialties, such as facial paralysis grading⁹⁶, suggesting that these advances are transferable to laryngology.

However, in this study, there are several limitations. The work was retrospective, and the exclusion of laryngeal cancer cases was necessary to ensure dataset homogeneity. The algorithm's performance in more complex scenarios, such as tumor-related fixation, remains to be tested. Videos were also prone to noise due to handheld camera movement, variations in anatomy, and inconsistent lighting conditions, which were not fully addressed, probably leading to the suboptimal results in

classification. Furthermore, bilateral paralysis cases presented challenges for automatic segmentation due to limited movement range, suggesting the need for more refined motion descriptors.

Future research directions include applying this approach to oncological cases where motility impairment may signal deep tumor invasion, as well as extending the analysis to integrate additional modalities such as stroboscopy, acoustic data, or clinical reports. Ultimately, this AI-driven motility assessment provides a foundation for developing quantitative, reproducible tools that could improve diagnostic accuracy, optimize treatment planning, and support longitudinal monitoring of vocal fold function in diverse clinical contexts.

In the context of this thesis, this work completes the progression from visual classification of static lesions to dynamic functional analysis, demonstrating how AI can capture both structural and kinematic aspects of laryngeal disease. The study provides an essential step toward comprehensive, multimodal AI-assisted endoscopic evaluation, aligning technological development with clinically meaningful outcome.

Conclusions

This thesis investigated the integration of artificial intelligence into the endoscopic evaluation of the larynx, encompassing the full spectrum of clinically relevant computer vision tasks—from image quality assessment to lesion detection, histologic risk classification, tumor boundary segmentation, and motility analysis. Starting from a systematic review and progressing through five complementary studies, it demonstrated that DL models can not only efficiently automate but also somehow enhance core diagnostic functions traditionally performed by clinicians, offering improved objectivity, reproducibility, and efficiency. Each experiment built upon the preceding one, forming a coherent methodological and clinical framework that spans the entire diagnostic workflow in laryngeal diagnosis and functional assessment.

The first study on informative frame selection established the technical foundation for this framework, introducing real-time automation to distinguish diagnostically valuable images from redundant or low-quality content. This innovation addressed one of the main limitations in endoscopic imaging, data redundancy, and ensured that subsequent analyses relied on standardized, high-quality visual input. The second study extended this concept to lesion localization, with a DL model that accurately detected benign and malignant laryngeal lesions across imaging modalities and different centers. By achieving real-time inference, it demonstrated that AI can feasibly operate as a “second observer” during live endoscopy, improving vigilance and supporting early detection. The third study advanced toward diagnostic interpretation, developing and externally validating a multicenter CADx model that classified lesions, identifying malignant and pre-malignant disease with accuracy comparable to expert clinicians, contributing to a more equitable and data-driven optical biopsy tool. The fourth study focused on tumor segmentation, introducing a DL model capable of delineating laryngeal cancer margins in standard light and NBI. By matching the accuracy of trained clinicians and achieving real-time processing, it provided a concrete example of how AI could assist intraoperative margin assessment and surgical planning, addressing one of the most critical challenges

in transoral oncologic surgery. Finally, the fifth study explored the functional dimension of laryngeal diagnostic assessment, applying AI to quantify vocal fold motion and classify normal motility, unilateral paralysis, and bilateral paralysis. This work demonstrated that AI can extend beyond visual diagnosis to capture kinematic patterns, translating qualitative clinical impressions into quantitative, reproducible metrics.

Across these studies, several common strengths emerged. Each model was developed using rigorously annotated, multicenter datasets and evaluated through patient-wise data partitioning and external validation to minimize bias and overfitting. The use of both WL and NBI ensured robustness across modalities, while real-time processing demonstrated the feasibility of clinical deployment. Importantly, all systems were designed with interpretability in mind, employing visualization tools or explainability analyses to align model behavior with established clinical reasoning. The multi-institutional and international collaborations that supported this research further reinforced the generalizability and practical relevance of its findings.

As discussed more specifically for each task, this thesis also underscores several limitations common to AI in medical imaging. Current models remain dependent on high-quality labeled data and may underperform in rare pathologies or atypical presentations. Dataset heterogeneity across centers, although beneficial for generalizability, introduces variability in illumination, endoscope type, and acquisition protocols that can affect model stability. Although the models performed well on still frames and preselected sequences, continuous video analysis and integration with live endoscopic systems will be essential for full clinical translation. Annotation subjectivity remains another constraint, particularly for tasks such as defining tumor boundaries or labeling frame quality, underscoring the need for consensus-based and standardized annotation protocols. Moreover, while the experiments included external validation, prospective testing in real-time clinical environments is still required to confirm safety and usability. These limitations point to the need for standardized

imaging protocols, larger shared datasets, and continued collaboration between clinicians, engineers, and data scientists.

From a clinical perspective, the findings presented in this thesis provide a strong rationale for the progressive integration of AI into otolaryngologic practice. The demonstrated systems could assist clinicians by automating repetitive and time-consuming tasks, supporting objective quality control, and improving early detection and risk stratification of laryngeal lesions. When embedded into endoscopic platforms, such models could offer real-time decision support during diagnostic and surgical procedures, enhancing precision and reducing inter-operator variability. Beyond tertiary centers, AI-assisted endoscopy could democratize access to high-quality diagnosis, particularly in regions where trained specialists are limited, echoing the broader goal of technology-enabled healthcare equity.

Looking ahead, the next phase of research should focus on prospective, real-time validation of these models within clinical workflows, the integration of multimodal data (combining imaging with acoustic, histopathologic, or genomic information), and the development of explainable and regulatory-compliant AI systems tailored to clinical needs. Future directions also include the creation of federated learning frameworks that enable privacy-preserving collaboration across institutions and continuous learning.

In conclusion, this thesis demonstrates that AI can serve as a reliable and transformative ally in the field of laryngeal endoscopy, capable of enhancing both the accuracy and accessibility of diagnosis and treatment. By uniting technical innovation with clinical relevance, in the future, we are expecting a new era of AI-assisted precision head and neck oncology, where decision-making is increasingly guided by standardized, transparent, and data-driven evidence.

LIST OF ABBREVIATIONS

HNSCC: head and neck squamous cell carcinoma

HPV: human papilloma virus

UADT: upper aerodigestive tract

NBI: narrow-band imaging

AI: artificial intelligence

ML: machine learning

DL: deep learning

CNNs: convolutional neural network

CV: computer vision

WL: white light

BB: bounding box

AP: average precision

FPS: frames per second

CADx: computer-aided diagnosis

HR: high-risk

LR: low-risk

TP: true positive

TN: true negative

FP: false positive

FN: false negative

AGA: anterior glottic angle

GMM: gaussian mixture model

MLP: multilayer perceptron

REFERENCES

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin.* 2021;71(3):209-249. doi:10.3322/caac.21660
2. Chen AM. Epidemiological Trends in the Characterization of Human Papillomavirus Virus-Related Oropharyngeal Cancer: A Global Perspective. *Curr Oncol Rep.* Published online August 20, 2025. doi:10.1007/s11912-025-01656-4
3. Curado MP, Hashibe M. Recent changes in the epidemiology of head and neck cancer. *Curr Opin Oncol.* 2009;21(3):194-200. doi:10.1097/CCO.0B013E32832A68CA
4. Conway DI, Petticrew M, Marlborough H, Berthiller J, Hashibe M, Macpherson LMD. Socioeconomic inequalities and oral cancer risk: A systematic review and meta-analysis of case-control studies. *Int J Cancer.* 2008;122(12):2811-2819. doi:10.1002/IJC.23430
5. Cancer Tomorrow. Accessed September 24, 2025. <https://gco.iarc.who.int/today/>
6. Cancer of the Oral Cavity and Pharynx - Cancer Stat Facts. SEER. Accessed September 24, 2025. <https://seer.cancer.gov/statfacts/html/oralcav.html>
7. Karamanou M, Markatos K, Lymperi M, Agapitos E, Androutsos G. A historical overview of laryngeal carcinoma and the first total laryngectomies. *J BUON.* 2017;22(3):807-811.
8. Zeitels SM, Burns JA, Dailey SH. Suspension laryngoscopy revisited. *Ann Otol Rhinol Laryngol.* 2004;113(1):16-22. doi:10.1177/000348940411300104
9. Carobbio ALC, Missale F, Fragale M, et al. Transoral laser microsurgery: feasibility of a new exoscopic HD-3D system coupled with free beam or fiber laser. *Lasers Med Sci.* 2021;36(9):1865-1872. doi:10.1007/S10103-020-03221-W

10. Hopkins HH. The Application of Frequency Response Techniques in Optics. <https://doi.org/10.1117/127971484>. 1969;7(5):131-136. doi:10.1117/12.7971484
11. Silberman HD, Wilf H, Tucker JA. Flexible fiberoptic nasopharyngolaryngoscope. *Ann Otol Rhinol Laryngol*. 1976;85(5):640-645. doi:10.1177/000348947608500513
12. Arens C, Piazza C, Andrea M, et al. Proposal for a descriptive guideline of vascular changes in lesions of the vocal folds by the committee on endoscopic laryngeal imaging of the European Laryngological Society. *Eur Arch Otorhinolaryngol*. 2016;273(5):1207-1214. doi:10.1007/s00405-015-3851-y
13. Vilaseca I, Valls-Mateus M, Nogués A, et al. Usefulness of office examination with narrow band imaging for the diagnosis of head and neck squamous cell carcinoma and follow-up of premalignant lesions. *Head Neck*. 2017;39(9):1854-1863. doi:10.1002/HED.24849
14. Piazza C, Cocco D, Benedetto LD, Bon FD, Nicolai P, Peretti G. Narrow band imaging and high definition television in the assessment of laryngeal cancer: a prospective study on 279 patients. *Eur Arch Oto-Rhino-Laryngol* 2009 2673. 2010;267(3):409-414. doi:10.1007/S00405-009-1121-6
15. Filauro M, Paderno A, Perotti P, et al. Role of narrow-band imaging in detection of head and neck unknown primary squamous cell carcinoma. *Laryngoscope*. 2018;128(9):2060-2066. doi:10.1002/lary.27098
16. Hinton GE, Osindero S, Teh YW. A Fast Learning Algorithm for Deep Belief Nets. *Neural Comput*. 2006;18(7):1527-1554. doi:10.1162/NECO.2006.18.7.1527

17. Deng J, Dong W, Socher R, Li LJ, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In: Institute of Electrical and Electronics Engineers (IEEE); 2010:248-255. doi:10.1109/cvpr.2009.5206848
18. Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: Understanding transfer learning for medical imaging. *Adv Neural Inf Process Syst*. 2019;32(NeurIPS).
19. Sampieri C, Ioppi A, Costantino A. Comments on: “Vocal Cord Leukoplakia Classification Using Siamese Network Under Small Samples of White Light Endoscopy Images.” *Otolaryngol - Head Neck Surg U S*. Published online October 1, 2024. doi:10.1002/ohn.690
20. Paderno A, Holsinger FC, Piazza C. Videomics: bringing deep learning to diagnostic endoscopy. *Curr Opin Otolaryngol Head Neck Surg*. 2021;29(2):143-148. doi:10.1097/MOO.0000000000000697
21. Wilmes CMLH, BSc AG, Marres HAM, Wellenstein DJ, Van Den Broek GB. A Systematic Review of the Clinical Impact of Implementing Artificial Intelligence in Upper Aerodigestive Tract Endoscopy. *Head Neck*. Published online June 18, 2025:hed.28213. doi:10.1002/hed.28213
22. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. doi:10.1136/bmj.n71
23. Hanif U, Kezirian E, Kiar EK, Mignot E, Sorensen HBD, Jennum P. Upper Airway Classification in Sleep Endoscopy Examinations using Convolutional Recurrent Neural Networks. *Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Int Conf*. 2021;2021:3957-3960. doi:10.1109/EMBC46164.2021.9630098

24. Cho WK, Lee YJ, Joo HA, et al. Diagnostic Accuracies of Laryngeal Diseases Using a Convolutional Neural Network-Based Image Classification System. *Laryngoscope*. 2021;131(11):2558-2566. doi:10.1002/lary.29595
25. Xiong H, Lin P, Yu JG, et al. Computer-aided diagnosis of laryngeal cancer via deep learning based on laryngoscopic images. *EBioMedicine*. 2019;48:92-99. doi:10.1016/j.ebiom.2019.08.075
26. Ay B, Turker C, Emre E, Ay K, Aydin G. Automated classification of nasal polyps in endoscopy video-frames using handcrafted and CNN features. *Comput Biol Med*. 2022;147:105725. doi:10.1016/j.compbiomed.2022.105725
27. Ren J, Jing X, Wang J, et al. Automatic Recognition of Laryngoscopic Images Using a Deep-Learning Technique. *Laryngoscope*. 2020;130(11):E686-E693. doi:10.1002/lary.28539
28. Girdler B, Moon H, Bae MR, Ryu SS, Bae J, Yu MS. Feasibility of a deep learning-based algorithm for automated detection and classification of nasal polyps and inverted papillomas on nasal endoscopic images. *Int Forum Allergy Rhinol*. 2021;11(12):1637-1646. doi:10.1002/alr.22854
29. Cho WK, Choi SH. Comparison of Convolutional Neural Network Models for Determination of Vocal Fold Normality in Laryngoscopic Images. *J Voice*. Published online 2020. doi:10.1016/j.jvoice.2020.08.003
30. Dunham ME, Kong KA, McWhorter AJ, Adkins LK. Optical Biopsy: Automated Classification of Airway Endoscopic Findings Using a Convolutional Neural Network. *Laryngoscope*. 2022;132(S4):S1-S8. doi:10.1002/lary.28708

31. Witt DR, Chen H, Mielens JD, et al. Detection of chronic laryngitis due to laryngopharyngeal reflux using color and texture analysis of laryngoscopic images. *J Voice*. 2014;28(1):98-105. doi:10.1016/j.jvoice.2013.08.015
32. Mascharak S, Baird BJ, Holsinger FC. Detecting oropharyngeal carcinoma using multispectral, narrow-band imaging and machine learning. *Laryngoscope*. 2018;128(11):2514-2520. doi:10.1002/lary.27159
33. Heo J, Lim JH, Lee HR, et al. Deep learning model for tongue cancer diagnosis using endoscopic images. *Sci Rep*. 2022;12(1). doi:10.1038/s41598-022-10287-9
34. He Y, Cheng Y, Huang Z, et al. A deep convolutional neural network-based method for laryngeal squamous cell carcinoma diagnosis. *Ann Transl Med*. 2021;9(24):1797-1797. doi:10.21037/atm-21-6458
35. Zhao Q, He Y, Wu Y, et al. Vocal cord lesions classification based on deep convolutional neural network and transfer learning. *Med Phys*. 2022;49(1):432-442. doi:10.1002/mp.15371
36. Xu J, Wang J, Bian X, et al. Deep Learning for nasopharyngeal Carcinoma Identification Using Both White Light and Narrow-Band Imaging Endoscopy. *Laryngoscope*. 2022;132(5):999-1007. doi:10.1002/lary.29894
37. Ye G, Du C, Lin T, Yan Y, Jiang J. Deep learning for laryngopharyngeal reflux diagnosis. *Appl Sci Switz*. 2021;11(11). doi:10.3390/app11114753
38. Yin L, Liu Y, Pei M, Li J, Wu M, Jia Y. Laryngoscope8: Laryngeal image dataset and classification of laryngeal disease based on attention mechanism. *Pattern Recognit Lett*. 2021;150:207-213. doi:10.1016/j.patrec.2021.06.034

39. Wang YY, Hamad AS, Palaniappan K, Lever TE, Bunyak F. LARNet-STC: Spatio-temporal orthogonal region selection network for laryngeal closure detection in endoscopy videos. *Comput Biol Med.* 2022;144. doi:10.1016/j.compbimed.2022.105339
40. Araújo T, Santos CP, De Momi E, Moccia S. Learned and handcrafted features for early-stage laryngeal SCC diagnosis. *Med Biol Eng Comput.* 2019;57(12):2683-2692. doi:10.1007/s11517-019-02051-5
41. Irem Turkmen H, Elif Karşligil M, Kocak I. Classification of laryngeal disorders based on shape and vascular defects of vocal folds. *Comput Biol Med.* 2015;62:76-85. doi:10.1016/j.compbimed.2015.02.001
42. Moccia S, De Momi E, Guarnaschelli M, Savazzi M, Laborai A. Confident texture-based laryngeal tissue classification for early stage diagnosis support. *J Med Imaging.* 2017;4(03):1. doi:10.1117/1.jmi.4.3.034502
43. Barbalata C, Mattos LS. Laryngeal tumor detection and classification in endoscopic video. *IEEE J Biomed Health Inform.* 2016;20(1):322-332. doi:10.1109/JBHI.2014.2374975
44. Patrini I, Ruperti M, Moccia S, Mattos LS, Frontoni E, De Momi E. Transfer learning for informative-frame selection in laryngoscopic videos through learned features. *Med Biol Eng Comput.* 2020;58(6):1225-1238. doi:10.1007/s11517-020-02127-7
45. Moccia S, Vanone GO, Momi ED, et al. Learning-based classification of informative laryngoscopic frames. *Comput Methods Programs Biomed.* 2018;158:21-30. doi:10.1016/j.cmpb.2018.01.030

46. Yao P, Witte D, Gimonet H, et al. Automatic classification of informative laryngoscopic images using deep learning. *Laryngoscope Investig Otolaryngol*. 2022;7(2):460-466. doi:10.1002/liv.2754
47. Galdran A, Costa P, Campilho A. Real-time informative laryngoscopic frame classification with pre-trained convolutional neural networks. *Proc - Int Symp Biomed Imaging*. 2019;2019:87-90. doi:10.1109/ISBI.2019.8759511
48. Kim GH, Sung ES, Nam KW. Automated laryngeal mass detection algorithm for home-based self-screening test based on convolutional neural network. *Biomed Eng Online*. 2021;20(1). doi:10.1186/s12938-021-00886-4
49. Matava C, Pankiv E, Raisbeck S, Caldeira M, Alam F. A Convolutional Neural Network for Real Time Classification, Identification, and Labelling of Vocal Cord and Tracheal Using Laryngoscopy and Bronchoscopy Video. *J Med Syst*. 2020;44(2). doi:10.1007/s10916-019-1481-4
50. Azam MA, Sampieri C, Ioppi A, et al. Deep Learning Applied to White Light and Narrow Band Imaging Videolaryngoscopy: Toward Real-Time Laryngeal Cancer Detection. *Laryngoscope*. 2022;132(9):1798-1806. doi:10.1002/lary.29960
51. Inaba A, Hori K, Yoda Y, et al. Artificial intelligence system for detecting superficial laryngopharyngeal cancer with high efficiency of deep learning. *Head Neck*. 2020;42(9):2581-2592. doi:10.1002/hed.26313
52. Tamashiro A, Yoshio T, Ishiyama A, et al. Artificial intelligence-based detection of pharyngeal cancer using convolutional neural networks. *Dig Endosc*. 2020;32(7):1057-1065. doi:10.1111/den.13653

53. Adamian N, Naunheim MR, Jowett N. An Open-Source Computer Vision Tool for Automated Vocal Fold Tracking From Videoendoscopy. *Laryngoscope*. 2021;131(1):E219-E225. doi:10.1002/lary.28669
54. Parker F, Brodsky MB, Akst LM, Ali H. Machine Learning in Laryngoscopy Analysis: A Proof of Concept Observational Study for the Identification of Post-Extubation Ulcerations and Granulomas. *Ann Otol Rhinol Laryngol*. 2021;130(3):286-291. doi:10.1177/0003489420950364
55. Paderno A, Piazza C, Del Bon F, et al. Deep Learning for Automatic Segmentation of Oral and Oropharyngeal Cancer Using Narrow Band Imaging: Preliminary Experience in a Clinical Perspective. *Front Oncol*. 2021;11(March):1-12. doi:10.3389/fonc.2021.626602
56. Lin J, Walsted ES, Backer V, Hull JH, Elson DS. Quantification and Analysis of Laryngeal Closure from Endoscopic Videos. *IEEE Trans Biomed Eng*. 2019;66(4):1127-1136. doi:10.1109/TBME.2018.2867636
57. Azam MA, Sampieri C, Ioppi A, et al. Videomics of the Upper Aero-Digestive Tract Cancer: Deep Learning Applied to White Light and Narrow Band Imaging for Automatic Segmentation of Endoscopic Images. *Front Oncol*. 2022;12. doi:10.3389/fonc.2022.900451
58. Li C, Jing B, Ke L, et al. Development and validation of an endoscopic images-based deep learning model for detection with nasopharyngeal malignancies 08 Information and Computing Sciences 0801 Artificial Intelligence and Image Processing. *Cancer Commun*. 2018;38(1). doi:10.1186/s40880-018-0325-9
59. Laves MH, Bicker J, Kahrs LA, Ortmaier T. A dataset of laryngeal endoscopic images with comparative study on convolution neural network-based semantic segmentation. *Int J Comput Assist Radiol Surg*. 2019;14(3):483-492. doi:10.1007/s11548-018-01910-0

60. Pan X, Bai W, Ma M, Zhang S. RANT: A cascade reverse attention segmentation framework with hybrid transformer for laryngeal endoscope images. *Biomed Signal Process Control*. 2022;78. doi:10.1016/j.bspc.2022.103890
61. Ji B, Ren J, Zheng X, et al. A multi-scale recurrent fully convolution neural network for laryngeal leukoplakia segmentation. *Biomed Signal Process Control*. 2020;59:101913. doi:10.1016/j.bspc.2020.101913
62. Hamad A, Haney M, Lever TE, Bunyak F. Automated segmentation of the vocal folds in laryngeal endoscopy videos using deep convolutional regression networks. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. Vol 2019-June. IEEE Computer Society; 2019:140-148. doi:10.1109/CVPRW.2019.00023
63. Ding H, Cen Q, Si X, Pan Z, Chen X. Automatic glottis segmentation for laryngeal endoscopic images based on U-Net. *Biomed Signal Process Control*. 2022;71. doi:10.1016/j.bspc.2021.103116
64. Girdler B, Moon H, Bae MR, Ryu SS, Bae J, Yu MS. Feasibility of a deep learning-based algorithm for automated detection and classification of nasal polyps and inverted papillomas on nasal endoscopic images. *Int Forum Allergy Rhinol*. 2021;11(12):1637-1646. doi:10.1002/alr.22854
65. CVAT - image anotator. Accessed October 5, 2025. <https://www.cvat.ai/>
66. SuperYOLO: Super Resolution Assisted Object Detection in Multimodal Remote Sensing Imagery. Accessed October 7, 2025. <https://ieeexplore.ieee.org/document/10075555>
67. Lin TY, Maire M, Belongie S, et al. Microsoft COCO: Common Objects in Context. *arXiv*. Preprint posted online February 21, 2015. doi:10.48550/arXiv.1405.0312

68. Baldini C, Azam MA, Sampieri C, et al. An automated approach for real-time informative frames classification in laryngeal endoscopy using deep learning. *Eur Arch Oto-Rhino-Laryngol Off J Eur Fed Oto-Rhino-Laryngol Soc EUFOS Affil Ger Soc Oto-Rhino-Laryngol - Head Neck Surg*. Published online 2024. doi:10.1007/S00405-024-08676-Z
69. Tan M, Le QV. EfficientNetV2: Smaller Models and Faster Training. *arXiv*. Preprint posted online June 23, 2021. doi:10.48550/arXiv.2104.00298
70. Baldini C, Migliorelli L, Berardini D, et al. Improving real-time detection of laryngeal lesions in endoscopic images using a decoupled super-resolution enhanced YOLO. *Comput Methods Programs Biomed*. 2025;260. doi:10.1016/j.cmpb.2024.108539
71. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. *arXiv*. Preprint posted online December 10, 2015. doi:10.48550/arXiv.1512.03385
72. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Int J Comput Vis*. 2020;128(2):336-359. doi:10.1007/s11263-019-01228-7
73. Peña MF de la, Gómez ÁLP, Maimó LF. ShaTS: A Shapley-based Explainability Method for Time Series Artificial Intelligence Models applied to Anomaly Detection in Industrial Internet of Things. *arXiv*. Preprint posted online June 2, 2025. doi:10.48550/arXiv.2506.01450
74. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. Published online April 16, 2024:e078378. doi:10.1136/bmj-2023-078378

75. Rondonotti E, Hassan C, Tamanini G, et al. Artificial intelligence-assisted optical diagnosis for the resect-and-discard strategy in clinical practice: the Artificial intelligence BLI Characterization (ABC) study. *Endoscopy*. 2023;55(1):14-22. doi:10.1055/a-1852-0330
76. Moon Y, Oh J, Hyun J, et al. Cost-Effective Smartphone-Based Articulable Endoscope Systems for Developing Countries: Instrument Validation Study. *JMIR MHealth UHealth*. 2020;8(9). doi:10.2196/17057
77. Stringer JSA, Pokaprakarn T, Prieto JC, et al. Diagnostic Accuracy of an Integrated AI Tool to Estimate Gestational Age From Blind Ultrasound Sweeps. *JAMA*. Published online August 1, 2024. doi:10.1001/JAMA.2024.10770
78. How AI is saving Indians from going blind: Four case studies. Accessed August 4, 2024. <https://indiaai.gov.in/article/how-ai-is-saving-indians-from-going-blind-four-case-studies>
79. Medtronic to provide underserved communities with artificial intelligence (AI)-assisted colonoscopy technology through Health Equity Assistance Program for colon cancer screening - Feb 22, 2022. Accessed August 4, 2024. <https://news.medtronic.com/2022-02-22-Medtronic-to-provide-underserved-communities-with-artificial-intelligence-AI-assisted-colonoscopy-technology-through-Health-Equity-Assistance-Program-for-colon-cancer-screening>
80. Cancer of the Larynx - Cancer Stat Facts. SEER. Accessed October 19, 2025. <https://seer.cancer.gov/statfacts/html/laryn.html>
81. Wissinger E, Griebisch I, Lungershausen J, Foster T, Pashos CL. The economic burden of head and neck cancer: a systematic literature review. *Pharmacoeconomics*. 2014;32(9):865-882. doi:10.1007/S40273-014-0169-3

82. Raggini E, Mattavelli D, Zigliani G, Bossi P, Piazza C. Measuring financial toxicity in head and neck cancer: a systematic review. *Acta Otorhinolaryngol Ital Organo Uff Della Soc Ital Otorinolaringol E Chir Cerv-facc.* 2024;44(1):1-12. doi:10.14639/0392-100X-N2762
83. Pisu M, Henrikson NB, Banegas MP, Yabroff KR. Costs of cancer along the care continuum: What we can expect based on recent literature. *Cancer.* 2018;124(21):4181-4191. doi:10.1002/CNCR.31643
84. Dar MA, Chauhan R, Murti K, Trivedi V, Dhingra S. Development and Validation of Subjective Financial Distress Questionnaire (SFDQ): A Patient Reported Outcome Measure for Assessment of Financial Toxicity Among Radiation Oncology Patients. *Front Oncol.* 2022;11. doi:10.3389/FONC.2021.819313
85. Luo H, Xu G, Li C, et al. Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study. *Lancet Oncol.* 2019;20(12):1645-1654. doi:10.1016/S1470-2045(19)30637-0
86. Seager A, Sharp L, Neilson LJ, et al. Polyp detection with colonoscopy assisted by the GI Genius artificial intelligence endoscopy module compared with standard colonoscopy in routine colonoscopy practice (COLO-DETECT): a multicentre, open-label, parallel-arm, pragmatic randomised controlled trial. *Lancet Gastroenterol Hepatol.* 2024;9(10):911-923. doi:10.1016/S2468-1253(24)00161-4
87. Marrero-Gonzalez AR, Diemer TJ, Nguyen SA, Camilon TJM, Meenan K, O'Rourke A. Application of artificial intelligence in laryngeal lesions: a systematic review and meta-analysis. *Eur Arch Otorhinolaryngol.* Published online 2024. doi:10.1007/s00405-024-09075-0

88. Fiz I, Koelmel JC, Sittel C. Nature and role of surgical margins in transoral laser microsurgery for early and intermediate glottic cancer. *Curr Opin Otolaryngol Head Neck Surg.* 2018;26(2):78-83. doi:10.1097/MOO.0000000000000446
89. Gorphe P, Simon C. A systematic review and meta-analysis of margins in transoral surgery for oropharyngeal carcinoma. *Oral Oncol.* 2019;98:69-77. doi:10.1016/J.ORALONCOLOGY.2019.09.017
90. Iandelli A, Gabella G, Marchi F, et al. The impact of margins in laryngeal cancer patients treated with transoral laser microsurgery: a systematic review and meta-analysis. *Eur Arch Oto-Rhino-Laryngol Off J Eur Fed Oto-Rhino-Laryngol Soc EUFOS Affil Ger Soc Oto-Rhino-Laryngol - Head Neck Surg.* 2024;281(9):4485-4494. doi:10.1007/s00405-024-08610-3
91. Garofolo S, Piazza C, Del Bon F, et al. Intraoperative narrow band imaging better delineates superficial resection margins during transoral laser microsurgery for early glottic cancer. *Ann Otol Rhinol Laryngol.* 2015;124(4):294-298. doi:10.1177/0003489414556082
92. Zwakenberg MA, Westra JM, Halmos GB, Wedman J, van der Laan BFAM, Plaat BEC. Narrow-Band Imaging in Transoral Laser Surgery for Early Glottic Cancer: A Randomized Controlled Trial. *Otolaryngol Neck Surg.* 2023;169(3):606-614. doi:10.1002/OHN.307
93. Valls-Mateus M, Nogués-Sabaté A, Blanch JL, Bernal-Sprekelsen M, Avilés-Jurado FX, Vilaseca I. Narrow band imaging for head and neck malignancies: Lessons learned from mistakes. *Head Neck.* 2018;40(6):1164-1173. doi:10.1002/hed.25088
94. Buchwald M, Nogal P, Nowak J, et al. Standardisation of an AI-based vocal fold assessment tool on a recurrent respiratory papillomatosis model. *Acta Otorhinolaryngol Ital Organo Uff Della Soc Ital Otorinolaringol E Chir Cerv-facc.* 2025;45(4):244-251. doi:10.14639/0392-100X-N2896

95. Ferrari M, Mularoni F, Taboni S, et al. How reliable is assessment of true vocal cord-arytenoid unit mobility in patients affected by laryngeal cancer? a multi-institutional study on 366 patients from the ARYFIX collaborative group. *Oral Oncol.* 2024;152:106744. doi:10.1016/j.oraloncology.2024.106744
96. Ten Harkel TC, de Jong G, Marres HAM, Ingels KJAO, Speksnijder CM, Maal TJJ. Automatic grading of patients with a unilateral facial paralysis based on the Sunnybrook Facial Grading System - A deep learning study based on a convolutional neural network. *Am J Otolaryngol.* 2023;44(3):103810. doi:10.1016/j.amjoto.2023.103810