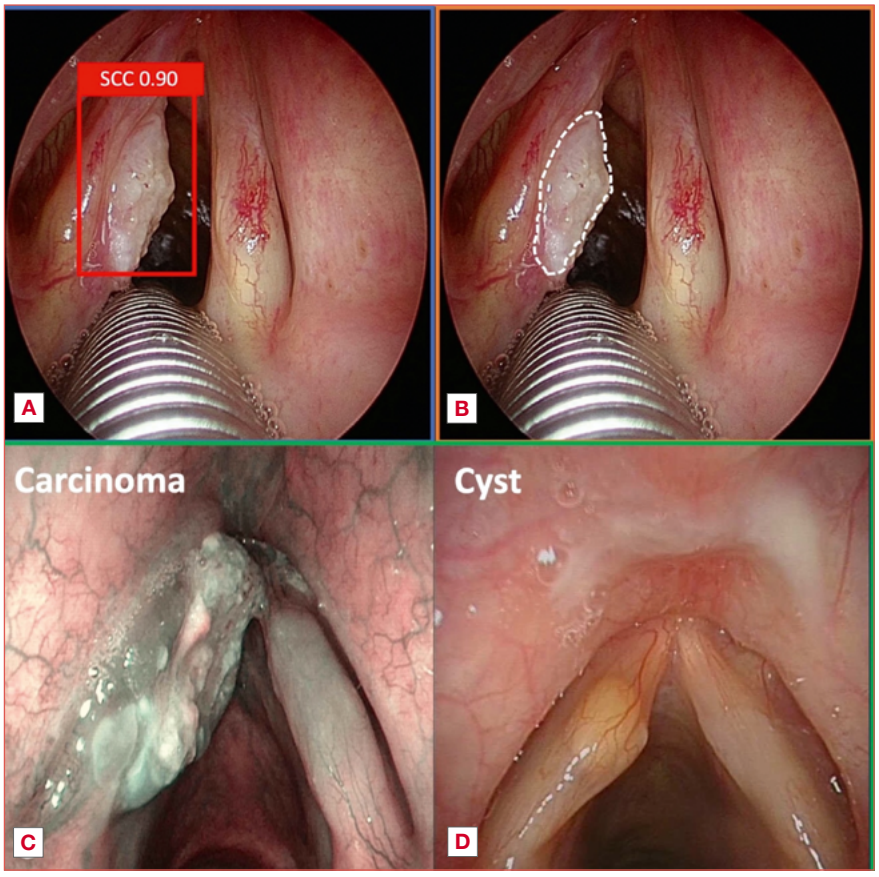
ACTA Otorhinolaryngologica Italica **2026**;46(Suppl. 1):S19-S33; doi: 10.14639/0392-100X-suppl.1-46-2026-A1967

Videomics and artificial intelligence in endoscopic diagnosis of laryngeal lesions: mapping current evidence through a scoping review



Cover figure. Main task of videomics applied to laryngeal endoscopy. The red box (**A**) represents the detection of a squamous cell carcinoma with a confidence score of 90%. The same lesion is evaluated, and a superficial segmentation is obtained in the white dotted box (**B**). **C** and **D** show the classification of a left vocal fold carcinoma and cyst, respectively.

Alessandro Ioppi[1*], Elisa Bellini[2,3*], Maria Sofia Salvetta[4,5], Filippo Marchi[2,3], Domenico di Maria[6], Giorgio Peretti[2,3], Pasquale D'Alessio[1], Pietro Perotti[1], Ottavio Piccin[1,7**], Claudio Sampieri[8-10**]

[1] Department of Otorhinolaryngology-Head and Neck Surgery, "S. Chiara" Hospital, Azienda Sanitaria Universitaria Integrata del Trentino (ASUIT), Trento, Italy; [2] Unit of Otorhinolaryngology-Head and Neck Surgery, IRCCS Ospedale Policlinico San Martino, Genoa, Italy; [3] Department of Surgical Sciences and Integrated Diagnostics (DISC), University of Genoa, Genoa, Italy; [4] Unit of Otorhinolaryngology, Head & Neck Department, Policlinico G.B. Rossi, University of Verona, Verona, Italy; [5] Department of Otorhinolaryngology, "S. Maria del Carmine" Hospital, Azienda Sanitaria Universitaria Integrata del Trentino (ASUIT), Rovereto, Italy; [6] Department of Otolaryngology, AORN "San Pio", Benevento, Italy;

Received: January 7, 2026
Accepted: January 29, 2026

Correspondence
Alessandro Ioppi
E-mail: alessandroioppi@gmail.com

How to cite this article: Ioppi A, Bellini E, Salvetta MS, et al. Videomics and artificial intelligence in endoscopic diagnosis of laryngeal lesions: mapping current evidence through a scoping review. Acta Otorhinolaryngol Ital 2026(Suppl. 1);46:S19-S33. https://doi.org/10.14639/0392-100X-suppl.1-46-2026-A1967

Summary

Laryngeal lesions are common and despite advances like high-definition videolaryngoscopy and enhanced imaging modalities such as narrow-band imaging, laryngoscopy remains operator-dependent. In this setting, artificial intelligence (AI) represents a promising tool to support clinical evaluation. This scoping review evaluated the current applications of AI in the endoscopic diagnosis of laryngeal lesions. A comprehensive search of MEDLINE and Scopus databases included 35 studies addressing AI-based detection, classification, or segmentation of laryngeal pathologies. Detection models frequently achieved real-time inference speeds and strong performance metrics,

although external validation was limited. Classification studies showed particularly robust results for binary tasks distinguishing high-risk from low-risk lesions, with some models achieving sensitivity and accuracy exceeding 90%. Segmentation models demonstrated the potential for precise delineation of cancer margins, a capability of notable relevance for surgical planning and intraoperative decision-making. Despite promising advances, heterogeneity in study design, limited external validation, and reliance on single-centre datasets currently restrict broad clinical implementation. Nonetheless, the emerging integration of AI into laryngeal endoscopy represents a significant step toward reproducible and accessible diagnostic assessment.

Keywords: artificial intelligence, deep learning, convolutional neural network, detection, larynx, segmentation

[7] Centro Interdipartimentale di Scienze Mediche (CISMED), University of Trento, Trento, Italy; [8] Department of Otorhinolaryngology, Hospital Clinic, Barcelona, Spain; [9] Head and Neck Cancer Unit, Hospital Clinic, Barcelona, Spain; [10] Department of Experimental Medicine (DIMES), University of Genoa, School of Medicine, Genoa, Italy.

**AI and EB contributed equally to this work. **OP and CS share the last authorship.*

Introduction

Approximately 6-22% of premalignant laryngeal lesions progress to malignancy and therefore early and accurate differentiation between benign and potentially malignant lesions is critical, as timely diagnosis and intervention significantly impact survival and quality of life [1].

Laryngoscopy represents the gold standard for evaluating laryngeal pathology. Moreover, diagnostic accuracy has markedly improved with the advent of white light (WL) high-definition digital videolaryngoscopy (HD-VLS) and imaging modalities that enhance submucosal vascular pattern, such as narrow-band imaging (NBI) [2-4]. Nevertheless, despite the advances of technology, laryngoscopy remains operator-dependent, necessitating a substantial learning curve and depending on inherent human limitations [5].

It is known that rural areas exhibit higher incidence and mortality rates of laryngeal cancer compared with urban regions, possibly due to a shortage of trained Otolaryngologists [6]. In such settings, the use of a diagnostic method that enables accurate and comprehensive evaluation of laryngeal lesions, while remaining cost-effective, would be of paramount importance [7].

In view of the foregoing, artificial intelligence (AI) represents a potentially powerful tool to assist clinicians in addressing the complexities of upper aerodigestive tract (UADT) examination. The integration of AI into endoscopy harnesses the power of computer vision (CV), a branch of AI that enables algorithms to extract meaningful information from visual data. This approach is commonly referred to as "videomics" [8,9].

Lately, the potentiality of AI in the diagnosis of laryngeal lesions has been investigated in several studies, through the development of different AI tasks, such as detection, classification, and segmentation. The aim of this paper is to focus on the application of AI to the endoscopic diagnosis of benign and malignant laryngeal lesions.

Materials and methods

A review of the literature was performed up to November 2025, following the Preferred Reporting Items for Systematic Reviews and MetaAnalyses (PRISMA) guidelines [10]. The following electronic databases were searched: MEDLINE and Scopus. The search strategy included MeSH terms comprising the site of examination (e.g., larynx), the endoscopic exam (e.g., laryngoscopy) and the application of AI (e.g., deep learning). The search strings are reported in Supplementary online material.

Inclusion criteria were: (1) Application of AI to laryngeal endoscopy (including both videos and frames as input data); (2) Use of AI for the analysis of laryngeal lesions (benign and/or malignant); (3) Properly reported outcomes; (4) Full text available. Exclusion criteria were: (1) AI applied to other diagnostic techniques (e.g., radiology, pathology, or other imaging modalities); (2) Review articles, systematic reviews, and letters to the editor; (3) Studies published in languages other than English. Selected papers went through screening and assessment of eligibility before being included.

Data collection and charting from the included study was performed independently by 3 authors (AI, EB, and MSS) and reviewed by all the authors. As presented in Figure 1, 35 articles were included in the review. Overall data were ultimately charted and analysed to describe the application of AI to diagnose laryngeal neoplasm, and 3 main fields of application were identified (lesion detection, classification and segmentation).

Results and discussion

Laryngeal lesions encompass a wide spectrum of conditions, from inflammatory to neoplastic diseases, and their variable presentation means that malignant lesions may be overlooked by less experienced clinicians or when suboptimal equipment is used. Accurate identification and diagno-

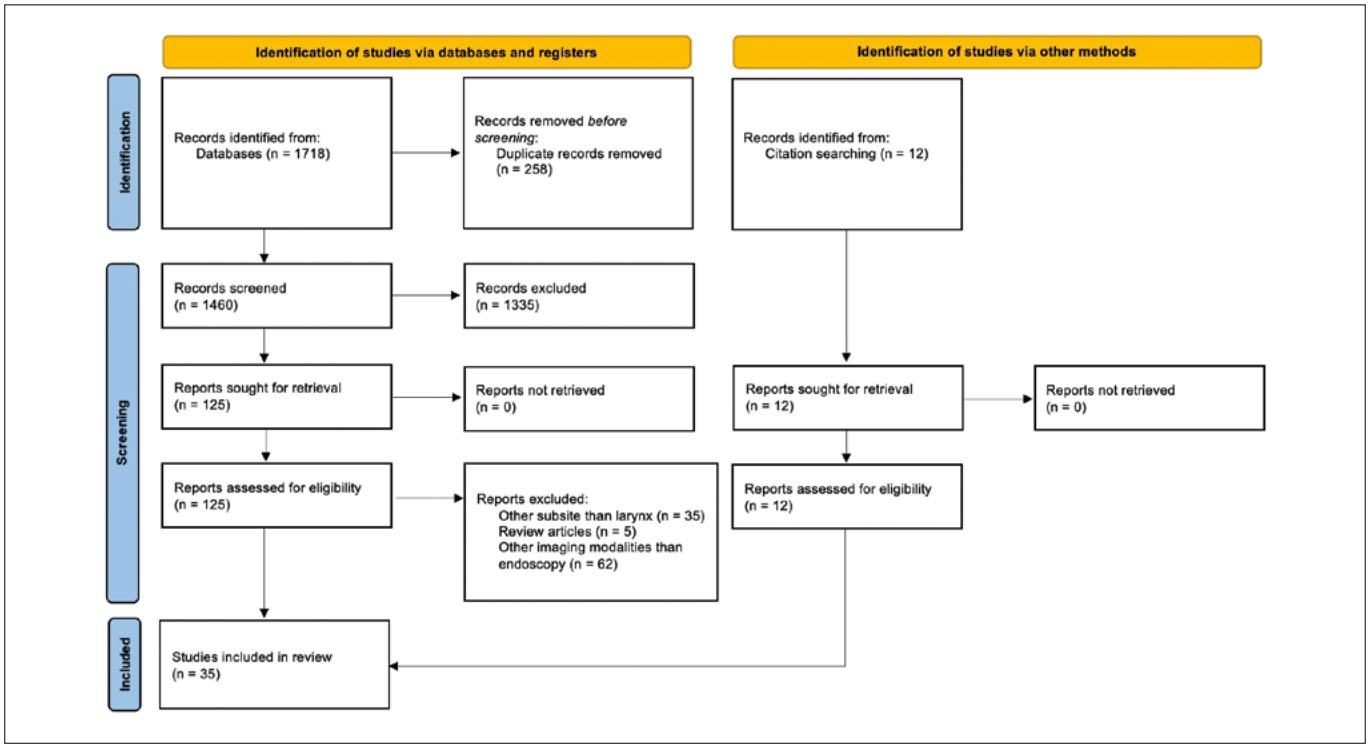


Figure 1. Preferred Reporting Items for Systematic Reviews and MetaAnalyses (PRISMA) flow diagram explaining the selection criteria of the included studies.

sis are essential for determining the appropriate treatment plan, and in case of malignancy, precise delineation of the lesion's superficial extent is critical for effective surgical management. In this context, AI has demonstrated substantial potential in videomics, achieving notable performance in detecting lesions during endoscopy, classifying them on frames or videos, and segmenting the superficial extent of malignant disease. The meta-analyses by Marrero-Gonzalez et al. [11] and Zurek et al. [12], involving 12 and 11 studies respectively, efficiently summarise the excellent outcomes of AI applied to laryngeal endoscopy, reporting a pooled sensitivity for the classification of benign from malignant lesions of 91%. Nevertheless, conducting a meta-analysis in this field is technically challenging, as the included studies exhibit substantial methodological heterogeneity. Each experiment uses a different algorithm, and thus the results depend on the model's implementation, context, and computational resources. Moreover, even if the diagnostic performances of numerous AI models are convincing, their applicability in clinical practice must first be tested and demonstrated on external validation cohorts, and this pivotal aspect is not adequately addressed in the existing reviews in the literature.

In a previous state-of-the-art review, we provided a guide through the complexities of AI applied to UADT endoscopy, analysing the outcome reporting systems, and explaining the main CV tasks in this field [8]. In this review, we guide readers through the complexities of the algorithmic architectures used in videomics focusing on the larynx, examining the most significant results of the included papers, analyzing the 3 main fields in which AI has been applied so far in the endoscopic diagnosis of laryngeal lesions: lesion detection, classification, and superficial delineation (Cover figure).

AI models in videomics

In recent years, the application of AI to laryngeal endoscopy has been the focus of increasingly extensive research, and its development has led to the validation of more complex and articulated algorithms. The vast majority of the included studies exploit deep learning (DL) models, while simpler models, such as the ones based on machine learning (ML) techniques (e.g., Support Vector Machine, Random Forest, or logistic regression model), are far less employed in CV. This is because they require manually engineered features during training (such as texture, colour, geom-

etry), they lose efficacy when analysing complex input such as endoscopic images or video, and are outperformed by DL models in almost all cases. However, as reported by Xiong et al. [13] and Kuo et al., [14] models like Support Vector Machine can play a role especially in case of small input datasets or when combined to convoluted neural network (CNN) in the role of classificators in a hybrid pipeline. Indeed, ML models, when trained well and on a high-quality dataset, are able to elaborate accurate predictions requiring minimal computational power. Nonetheless, these models are scarcely applicable in clinical practice, since they are unable to handle the vast amount of information contained in an endoscopic image or video. This problem is well faced by DL models, which can autonomously extrapolate significant information by input data and, even if require substantial training data and computational resources, they can extract and interpret information from images and videos at a level that may exceed the capabilities of the human eye. The DL algorithms structures employed in CV and in the papers included can be summarised in three main classes: (1) CNN; (2) encoder-decoder architecture; (3) models implementing more sophisticated designs (transformer-based models or hybrid models). The first architecture employs a large number of local convolutions that can extract patterns from visual input data (such as textures, edges, shades, structural components of the image) and finally provide a prediction. To handle the enormous amount of information, these models are generally based on deep convolutional layers pre-trained on large image datasets (*e.g.* ImageNet) and are fine-tuned on specific images such as laryngoscopic images. Moreover, they usually use pooling functions, which reduce the spatial dimensions of the feature maps (downsampling) while preserving the most relevant information. In other words, pooling "summarises" small regions of the image into more compact ones (such as squares of 2x2 or 3x3 pixels) and assign them a simple value, in order to be recognised by the algorithm, reducing computational load and easing the extraction of structural pattern (e.g., vocal fold edges, lesion margins, vascular anomalies). Since sequential convolutional layers can extract increasingly precise information from data and understand recognisable patterns that can be aggregated to generate predictions, CNNs are specifically designed for classification (e.g., AlexNet, VGG, GoogleNet, ResNet, DenseNet) or detection (e.g., YOLO, SSD).

On the other hand, segmentation models employed in the field of videomics have usually been structured through an encoder-decoder architecture (e.g., U Net, SegNet). This can be observed in selected studies, since all the segmentation papers that use this kind of architecture [15-19] report significant better results than those employing a CNN structure for segmentation purposes [20,21]. The encoder-decoder models are characterised by a double branch design: the first one is represented by the encoder part (which works like a classic CNN, extracting patterns from data and downsampling the input information) and the second one is the decoder. This latter is able to upsample the data, producing a pixel-wise feature map enriching the original data with semantic information. To achieve this, the key element is the use of the so-called skip connections. These consist of direct links between the encoder and decoder layers that allow the model to preserve precise spatial information that would be compressed and lost in the encoding process. By reintroducing this information into the decoder, skip connections prevent imprecise data reconstruction and allow a more precise localisation of the target.

The third class of CV models, which has been explored only in recent years, is based on transformer and vision transformer architectures. Contrarily to CNNs, where the image is downsampled and analysed through convolutions, these models divide the data in numerous patches (producing a grid in the image) which are interpreted as numerical vectors. In this way, the algorithm is able to analyse the whole image at the same time, and therefore to catch relevant relationships between different patches of the image, even if located in distant areas. This is possible due to specific algorithm units called self-attention modules that allow the model to weigh the relative significance of every patch created by the transformer. Ultimately, this architecture enables the model to construct a global representation of the image, capturing relationships even between pixels that are far apart.

Detection task

This assignment refers to the localisation and labelling of a laryngeal lesion within an endoscopic frame. As reported in Table I, this task is usually carried out by CNN-based models and this is confirmed by our review, since 10 of 11 detection papers employed this architecture [5,7,17,22-28], while only one [29] published in 2025, exploited a vision transformer. Kang et al. [7] integrated a transformer module into a CNN to enhance diagnostic performance. The evaluation of such detection models typically relies on standard metrics, as previously described [8]. A prediction is considered a true positive when the region identified by the model mostly overlaps the ground-truth area (usually at least 50%), a criterion known as Intersection over Union (IoU). The most commonly used performance indicators are recall (sensitiv-

Table I. Detection task: summary of the studies included for mucosal lesion detection.

First author, year	Algorithm architecture	Main task	No. of samples	Main outcomes	External validation	Details
Zhang, 2025 [29]	Other (vision transformer)	LSCC detection and classification	3140 images (617 pts)	Sensitivity = 88%, specificity = 98%, accuracy = 95%, precision = 89%, F1 score = 88%.	No	Images obtained with rigid endoscope Only WL images
Kang, 2024 [7]	CNN + transformer	LSCC detection and classification	2023 images (613 pts)	Accuracy = 98.07%, recall = 98.31%, F1 score = 98.18%	No	Only WL images
Fang, 2023 [23]	CNN	LSCC detection	279 images (279 pts)	Sensitivity = 73%, specificity = 93%, accuracy = 73%. Cyst AUC = 0.86, nodules AUC = 0.78; SCC AUC = 0.89; pre-SCC AUC = 0.84	No	Detection of 5 classes (pre-SCC, SCC, cyst, nodule, healthy) Only WL images
Kim, 2023 [37]	CNN	Laryngeal benign lesions detection and classification	2183 images	F1 score = 0.85, accuracy = 0.94, precision = 0.88, recall = 0.82, specificity = 0.97	No	Images obtained with rigid endoscope Only WL images
Bhattacharjee, 2023 [25]	CNN + ensemble model	LSCC detection and classification	3000 images (30 pts)	Healthy: precision = 0.98, recall = 0.98, F1 score = 0.98 LSCC: precision = 0.98, recall = 0.98, F1 score = 0.98	No	Only WL images
Azam, 2022 [17]	CNN	LSCC detection	624 images (219 pts)	Precision = 0.66, Recall = 0.62 Recall, mAP50 = 0.63	No	Average computation time of 36 FPS--> real time use Trained and tested on WL and NBI frames
Cen, 2019 [26]	CNN	LSCC detection	400 images	Faster R-CNN: precision = 0.85, recall = 0.90, AP = 0.89, FPS = 9; SSD: precision = 0.83, recall = 0.90, AP = 0.89, FPS = 47; YOLOv3: precision = 0.77, recall = 0.80, AP = 0.79, FPS = 50	No	Images obtained with rigid endoscope Only WL images
Wellenstein, 2023 [27]	CNN	LSCC detection	4488 images	TP% = 78%, precision = 0.68, recall = 0.77, F1 score = 0.72	No	Real time detection (62 FPS) for the model with most parameters not suitable to run on commercialised laptop Only WL images
Bur, 2023 [28]	CNN + feature pyramid vector map	LSCC detection and classification	8172 images (147 pts)	Accuracy = 88.5%, mAP50 = 0.51	No	Only WL images
Baldini, 2025 [5]	CNN + super resolution branch	LSCC detection	3892 images (1593 pts)	mAP50 = 0.82	Yes	Trained on heterogeneous data (from 3 different datasets) Trained and tested on WL and NBI frames Real-time use (almost 60 FPS)
Nie, 2025 [22]	CNN	LSCC detection and classification	1353 images	Precision = 0.94, Recall = 0.79, mAP50 = 0.89	No	11 classes for classification, unbalanced dataset. Only WL images

LSCC: laryngeal squamous cell carcinoma; SCC: squamous cell carcinoma; WL: white light; NBI: narrow band imaging; CNN: convoluted neural network; AUC: area under the curve; mAP50: mean average precision at 50% of intersection over union; FPS: frames per second; SSD: single shot detector; AP: average precision; TP: true positive; pts: patients.

ity), precision (positive predictive value), average precision (AP), and the F1 score. AP is generally the most informative metric, as it represents the area under the precision–recall curve and thus integrates these 2 measures over all confidence thresholds. AP refers to performance on a single class, whereas mean AP (mAP) extends this evaluation across multiple classes. The F1 score, by contrast, is the harmonic mean of precision and recall computed at a single confidence threshold (i.e., 50%), making it a less complete measure than mAP. Only 4 of 11 detection papers included mAP in the computed results of their models [5,17,22,28], with values ranging from 0.51 to 0.89. It is important to note that mAP is typically reported at an IoU threshold of 50% (mAP50). Some studies, however, calculate the metric across a range of thresholds (e.g., mAP50-95), averaging performance over all IoU levels. For instance, Nie et al. reported an mAP50 of 0.89 and an mAP50-95 of 0.67 [22], indicating that while the model performs strongly in general lesion detection, its precision declines at higher IoU thresholds (e.g., 80-95%). This may be due to the imbalance of the training dataset (e.g., 804 malignant frames versus only 6 lipoma frames), the small size of the lesion within the image, or the presence of irregularly shaped lesions. From a clinical standpoint, a model with a high mAP50-95 would be essential for detecting even small or rare lesions.

The clinical usefulness of a detection model strongly depends on its ability to operate at real-time speed (~25 frames per second [FPS]), and this topic has been addressed by various authors [5,17,26,27]. Azam et al. developed a model able to detect squamous cell carcinoma (SCC) in videolaryngoscopies, achieving a true positive rate of 82% and a mAP of 0.63 with a computation speed of 38.5 FPS [17]. Cen and colleagues evaluated several architectures on a dataset of only 400 images and reported promising results (precision 0.77, recall 0.80, AP 0.789, FPS 50) [26]. It is important to note, however, that the training images were obtained exclusively during procedures under general anaesthesia using rigid endoscopes, resulting in close-up views of the lesions and high-quality images. Notably, Baldini et al. [5] and Wellenstein et al. [27] also reported performance suitable for real-time use, achieving 60 and 62 FPS, respectively. Nevertheless, only the study of Baldini et al. focused on improving the outcome of these models in detecting small lesions. Interestingly, most studies achieving real-time performances used a version of the YOLO architecture, which is specifically designed for detection tasks and is well known for its fast computational performance even beyond the medical field [30]. Adjunctively, even a well-designed model with strong performance and fast inference time must be validated on external datasets to demonstrate its generalisation capability before clinical implementation. Notably, only one of the 11 studies evaluated the proposed model on an external cohort, reporting good results (mAP50 = 0.82 and 0.84 for internal and external validation, respectively), despite relying on a relatively small training dataset [5]. Lastly, the complexity of a model directly influences the computational resources required and, consequently, the hardware needed to run it. This consideration is essential when evaluating the potential clinical application of an algorithm, particularly in terms of cost-effectiveness.

Classification task

This task pertains to the capability of predicting the categorical class of an object present within an image. The output consists of a data label that assigns a specific class to the image based on the model's prediction. For laryngeal endoscopy, this task usually consists of discriminating the histopathological nature of a lesion starting from its endoscopic appearance, the so-called "optical biopsy" [31-33]. Twenty-six articles included in this review reported data on classification (Tab. II) [7,13,14,19-23,25,28,29,34-48]. Among these, 21 studies employed CNN-based models [7,13,19-22,25,28,35-44,46-48], whereas the remaining adopted alternative approaches. Furthermore, only 6 of the studies included incorporated external validation to substantiate their findings [7,20-22,35,36]. In the studies included in the review, the classification of neoplastic lesions – or, alternatively, the distinction between benign and malignant lesions – was evaluated. Among the reported metrics, accuracy is the most commonly cited, though its values vary considerably. Accuracy provides an overall measure of model reliability but may offer a misleading assessment when the dataset contains imbalanced classes. For this reason, other metrics such as precision, recall and F1 score should also be evaluated. Zhao et al. [38] reported an accuracy of 80.2% when classifying 4 categories of conditions (polyp, keratinisation, SCC, and normal mucosa), which increased to 93.9% when the classification was simplified to 2 categories (high-risk *vs* low-risk). This latter observation is also supported by Dunham and colleagues [41], with an accuracy of 80.8% achieved across 5 categories, rising up to 93% when analysing only 2 categories. The highest value of accuracy (97.8%) was reported by Nobel et al. [19], who employed a hybrid model (CNN + encoder-decoder) on a 24,000 images dataset to segment the vocal fold and therefore provide a classification between 5 classes.

It can be inferred that binary classification tends to produce more reliable outcomes. Intuitively, increasing the complexity of the prediction task typically results in reduced accura-

Table II. Classification task: summary of the studies included for mucosal lesion classification.

First author, year	Algorithm architecture	Main task	No. of samples	Main outcomes	External validation	Details
Zhang, 2025 [29]	Other (vision transformer)	Laryngeal lesion classification	3140 images (617 pts)	Sensitivity = 88%, specificity = 98%, accuracy = 95%, precision = 89%, F1 score = 88%	No	Employs transfer learning to increase accuracy with low input parameters. Images obtained with rigid endoscope under general anaesthesia Only WL images
Kang, 2024 [7]	CNN + transformer	LSCC detection and classification	2023 images (613 pts)	Classification in 3 classes: normal - abnormal - carcinoma. accuracy = 98.07%, recall = 98.31%, F1 score = 98.18%	No	Only WL images Build an algorithm that helps prioritize SCC patients to third level care
Qiu, 2024 [34]	Other	LSCC detection and classification	1109 videos (35488 frames)	Accuracy = 92.4%, sensitivity = 95.6%, precision = 94.1%, F1 score = 94.8%	No	Medical video classification in the context of laryngoscopic videos (555 normal cases, 240 benign cases, and 314 malignant cases)
Xu, 2023 [35]	CNN	LSCC classification	2254 images	Internal validation: accuracy = 92%, AUC = 97.4%, sensitivity= 91.6%, specificity = 92.4% External validation: ACC = 86.3%, AUC = 92.6%, sensitivity = 86%, specificity = 86.5%.	Yes	Only WL images
Li, 2023 [36]	Encoder-decoder + CNN	LSCC classification	31543 images	Internal validation: accuracy = 95.6%, sensitivity = 94.8%, specificity = 96.4%, AUC = 97.4%	Yes	External test: tested on 5 different external datasets with comparable outcomes Video test: comparable outcomes. AUC = 97.4% Comparison with human experts: model achieved 0.940 accuracy, performing comparable to expert laryngologists
Kim, 2023 [37]	CNN	Laryngeal benign lesions detection and classification	2183 images	F1 score = 0.85, accuracy = 0.94, precision = 0.88, recall = 0.82, specificity = 0.97	No	Images obtained with rigid endoscope Only WL images
Bhattacharjee, 2023 [25]	CNN + ensemble model	LSCC detection and classification	3000 images (30 pts)	Healthy: precision = 0.98, recall = 0.98, F1 score = 0.98 LSCC: precision = 0.98, recall = 0.98, F1 score = 0.98	No	Only WL images
Zhao, 2022 [38]	CNN	Laryngeal lesion classification	456 images	Classification of 4 classes (polyp, keratinisation, SCC, healthy) - overall accuracy = 80.23%, F1 score = 0.78, AUC = 0.95. Classification of 2 classes (keratinisation+SCC *vs* polyp+healthy) - overall accuracy = 0.94, sensitivity = 0.89, specificity = 0.99, and AUC = 0.98	No	Only WL images
Cho, 2021 [39]	CNN	Laryngeal lesion classification	4106 images	F1 score Reinke's edema: 0.72 ± 0.06, Nodules: 0.73 ± 0.09, Cyst: 0.78 ± 0.08, Granuloma: 0.94 ± 0.02, Leukoplakia: 0.89 ± 0.02, Normal: 0.92 ± 0.02, Palsy: 0.96 ± 0.01, Papilloma: 0.87 ± 0.04, Polyp: 0.86 ± 0.02	No	Comparing classification power of CNN vs trainees. AI is overall better in discriminating different classes of laryngeal lesions

Table II. *continues.*

First author, Year	Algorithm architecture	Main task	No. of samples	Main outcomes	External validation	Details
Ren, 2020 [40]	CNN	Laryngeal lesion classification	4667 images	Overall accuracy = 96.24%	No	Only WL images The model outperformed physicians for most laryngeal conditions
Xiong, 2019 [13]	CNN + support vector machine	Laryngeal lesion classification	13721 images	Classification of 2 classes (SCC + pre-SCC *vs* benign and norm): Sensitivity = 0.73, specificity = 0.92, AUC = 0.92, accuracy = 0.87	No	Only WL images When compared to human experts the model was comparable to an expert with 10-20 years of experience
Bur, 2023 [28]	CNN + feature pyramid vector map	LSCC detection and classification	8172 images (147 pts)	Accuracy = 88.5%, mAP50 = 0.51	No	Only WL images
Dunham, 2022 [41]	CNN	Laryngeal lesion classification	19353 images	Classification of 5 classes (healthy, nodule, polyp, papilloma, web) - accuracy = 80.8%, precision = 71.7%-89%, recall = 70%-88%. Classifier for 2 classes (SCC+premalignant *vs* healthy) - accuracy = 93%, recall = 92%	No	Only WL images Important class imbalance between SCC (1005) and benign pathology (papilloma 3633, polyps 4577)
Kuo, 2021 [14]	Other (support vector machine)	Laryngeal lesion classification	284 images	Accuracy = 93.3%	No	Apply a complex algorithm able to 1. Reduce illumination problem; 2. Automatically screen clear images; 3. Segment automatically the ROI; 4. Support vector machine to classify lesions. Only WL images
Wang, 2024 [21]	CNN	Laryngeal leukoplakia segmentation and classification	5362 images + 50 videos (551 pts)	AUC = 73.1-86.9; accuracy = 68.7-82.0; sensitivity = 58-73.3, specificity = 79.4-100	Yes	Complete algorithm, studied to analyze the image, segment the ROI and provide classification of leukoplakias. Only WL images With AI, the AUC improved from 0.72 to 0.80 for senior clinicians and from 0.65 to 0.80 for junior clinicians
Yao, 2024 [42]	CNN	Laryngeal polyp classification	37024 total frames	Accuracy 85% and AUC 0.84	No	Only WL images Training dataset made by machine-labeled frames instead of human-labeled
Yan, 2023 [43]	CNN	LSCC classification	2179 images	Specificity = 78.59%, sensitivity = 74.16%, accuracy = 78.05%, NPV = 95.63%, PPV = 32.51%.	No	Dataset acquired from 6 centres with 5 video acquiring systems Unbalanced dataset (288 malignant and 1891 benign) Only WL images

Table II. *continues.*

First author, Year	Algorithm architecture	Main task	No. of samples	Main outcomes	External validation	Details
Xiong, 2024 [44]	CNN	Laryngeal leukoplakia segmentation and classification	6180 (WL = 3080, NBI = 3100).	Classification outcomes (Low grade *vs* High grade + SCC): - WL: sensitivity = 93%, specificity = 94%, PPV = 93%, NPV = 94% - NBI: sensitivity = 99%, specificity = 97%, PPV = 99%, NPV = 97%	No	Trained and tested on WL and NBI frames
Kang, 2024 [45]	Other (swin transformer)	Laryngeal lesion classification	5008 images (1230 pts)	4-class classifier (healthy, benign, pre-SCC, SCC): Accuracy = 92.78%, AUC = 0.97, F1 score = 76.18%, specificity = 97.61%	Yes	Only WL images The model demonstrated superior performance to the 3 human experts Model inference speed = 339.76 FPS Human inference speed = 0.0556 FPS (18 s per picture).
You, 2025 [47]	CNN	Laryngeal leukoplakia classification	666 images	Accuracy = 96.12%	No	Only NBI images Six categories: normal tissue, inflammatory keratosis, mild dysplasia, moderate dysplasia, severe dysplasia, SCC
You, 2023 [46]	CNN + other (visual transformer)	Laryngeal leukoplakia classification	932 images	Overall accuracy WL = 0.96 Overall accuracy NBI = 0.95	No	Trained and tested on WL and NBI frames Dataset unbalanced with NBI images x3
Nie, 2025 [22]	CNN	LSCC detection and classification	1353 images	Precision = 0.94, Recall = 0.79, mAP50 = 0.89	No	11 classes for classification, unbalanced dataset Only WL images
Yin, 2021 [48]	CNN	Laryngeal lesion classification	3057 images (1950 pts)	Average AUC = 0.89, malign AUC = 0.94, average accuracy = 73%	No	Only WL images DL workflow: detect the lesion and locate the critical area, then classify the lesion based on that area and not on the original image. Accuracy on original image 71%, on critical area 77% (AUC 84.3 *vs* 91.2)
Nobel, 2024 [19]	CNN + encoder-decoder	Vocal folds segmentation and laryngeal lesion classification	24000 images	Classification 5 classes (SCC, dysphonia, paresis, polyp, and healthy) accuracy = 97.88%.	No	Only WL images
Tie, 2024 [20]	CNN	Laryngeal leukoplakia segmentation and classification	7057 images (426 pts)	Classification of 6 classes (hyperplasia, inflammation, mild dysplasia, moderate dysplasia, severe dysplasia/CIS, and SCC) - AUC = 0.868 in the internal dataset and 0.884 in the external set - On prospective video: AUC = 0.82, accuracy = 0.840, sensitivity = 1.0, specificity = 0.73, PPV = 0.71, NPV = 1.0	Yes	Trained and tested on WL and NBI frames External validation + video validation for real time use + human-machine comparison AI significantly improved AUC and accuracy for all laryngologists ($p < 0.05$)

WL: white light; LSCC: laryngeal squamous cell carcinoma; SCC: squamous cell carcinoma; CNN: convoluted neural network; AUC: area under the curve; ACC: accuracy; mAP50: mean average precision at 50% of IoU; ROI: region of interest; NPV: negative predictive value; PPV: positive predictive value; FPS: frames per second; NBI: narrow band imaging; DL: deep learning; CIS: carcinoma in situ; pts: patients.

cy. Due to this, the most promising models to be introduced into clinical practice are likely the ones for differentiating high-risk from non-high-risk lesions. This is underlined in the publication of Kang et al. [7], where the model obtained a 98% accuracy and was specifically developed to screen patients and redirect high-risk ones to third-level care. Considering that all these studies are preclinical trials, future clinical validations will undoubtedly be necessary to obtain more reliable data for real-world use. Within the studies considered, only 5 [13,20,36,46,47] included NBI images in the training dataset, reporting interesting results. Specifically, Li et al. [36] developed a model that extracts peculiar features from WL and NBI images separately, and further fuses the 2 features to obtain a comprehensive prediction. With this strategy, the authors observed more accurate predictions on the overall dataset, compared with other models simply using mixed images during training. On the other hand, Xiong and colleagues reported outstanding results when classifying low-grade *vs* high-grade lesions and SCC [44], achieving a sensitivity of 93% and a specificity of 94% for WL, and a sensitivity of 99% and a specificity of 97% for NBI. These findings highlight the value of NBI in supporting AI models for diagnosing high-risk vocal fold lesions and suggest that it should be incorporated into future algorithmic architectures and training datasets.

Several studies have examined and compared the classification performance of AI models with that of specialist physicians. Ren et al. [40] reported that the CNN-based classifier outperformed laryngologists for most laryngeal conditions. The same results are described in the paper of Cho and colleagues [39], where the model achieved an overall F1 score of 88% and outperformed trainees in classifying all the lesions provided (cysts, granulomas, nodules, palsies, papillomas, and polyps). Similarly, Li et al. reported that the classifier achieved an overall accuracy of 94%, performing comparably to expert laryngologists but outclassing others with less experience [36]. Overall, the current literature appears consistent in showing that well-trained DL models can achieve comparable performance to that of expert physicians in classifying laryngeal lesions. However, these findings are not yet sufficient to support the clinical implementation of such algorithms, largely because most models lack reproducibility and robustness.

Segmentation task

The segmentation task refers to the identification of targets by assigning a label to every pixel in the image, producing a pixel-wise mask that outlines the precise contours of each object. Algorithms developed for this task can assist in delineating the superficial extent of laryngeal lesions or identifying the boundaries of regions of interest, such as the vocal folds or ventricular bands. Since this scoping review focuses on the diagnosis of laryngeal lesions, we excluded studies evaluating AI models aimed solely at segmenting anatomical structures without relevance to lesion identification. In this light, only 9 articles were included in our review regarding the segmentation task (Tab. III) [16-21,44,49,50]. To evaluate the performance of a segmentation algorithm, precision, recall, AP, and accuracy are considered alongside the IoU and the Dice Similarity Coefficient (DSC). The DSC quantifies the overlap between the predicted mask and ground-truth segmentation, and it is the most informative metric for assessing semantic segmentation models.

Segmentation may have meaningful clinical applications, particularly in the characterisation and assessment of malignancies. In our review, only 4 studies focused specifically on the segmentation of laryngeal SCC [16,17,49,50], whereas the remaining articles examined laryngeal leukoplakia [18,20,21,44] or incorporated segmentation to identify the region of interest prior to the classification task [19]. Among the studies reporting data on SCC segmentation, Azam et al. [17] reported a DSC of 81% using an encoder-decoder architecture. The model performed equally well on both WL and NBI images, suggesting that it was able to interpret complex vascular patterns independently of the optical filter employed. Therefore, AI might enhance the use of NBI even in less experienced centres by improving the accuracy of lesion detection and margin identification, regardless of the operator's familiarity with the technique. In a subsequent study Sampieri et al. reported the results of SegMENT-Plus, a segmentation model, which demonstrated excellent performance across different external validation cohorts [50]. The model accurately delineated laryngeal SCC boundaries in endoscopic images, achieving results comparable to those of 2 residents in Otolaryngology. Findings from 2 external datasets confirmed the model's robust generalisation capabilities. Moreover, its computational efficiency supported seamless application to videolaryngoscopies, effectively simulating real-time deployment. The model achieved the following median performance metrics: DSC = 0.83 (0.70-0.90), IoU = 0.83 (0.73-0.90), and accuracy = 0.97 (0.95-0.99). Finally, Paderno et al. [16] confirmed the feasibility of applying instance segmentation to the UADT using DL algorithms, although they reported lower diagnostic performance in the oral cavity compared with other anatomical subsites. The model achieved a DSC of 0.90 ± 0.05 when applied to laryngeal SCC, while it was 0.60 ± 0.26 when implemented on oral cavity lesions ($p < 0.001$). This dis-

Table III. Segmentation task: summary of the studies included for mucosal lesion segmentation.

First author, year	Algorithm architecture	Main task	No. of samples	Main outcomes	External validation	Details
Azam, 2024 [49]	Encoder-decoder	LSCC segmentation	4289 (766 pts)	DSC = 81.4-84.9% IoU = 81.8-85.7%	Yes	Trained and tested on WL and NBI frames
Sampieri, 2024 [50]	Encoder-decoder	LSCC segmentation	3933 (557 pts)	DSC = 0.83 (0.70-0.90) IoU = 0.83 (0.73-0.90)	Yes	Trained and tested on WL and NBI frames Tested on real intraoperative laryngoscopy videos Model inference speed 25.6 FPS Outcome comparable to junior and senior residents
Paderno, 2023 [16]	Encoder-decoder	LSCC segmentation	1034 images	DSC = 0.90 ± 0.05	Yes	Only NBI images Inferior diagnostic results in the oral cavity compared with the larynx ($p < 0.001$)
Azam, 2022 [17]	Encoder-decoder	LSCC segmentation	683 images	IoU = 0.68, DSC = 0.81, recall = 0.95, precision = 0.78, accuracy = 0.97	Yes	Trained and tested on WL and NBI frames Transfer learning from ImageNet Model tested even on oral cavity and oropharynx images, improving state-of-the-art model result
Ji, 2020 [18]	Encoder-decoder	Leukoplakia segmentation	649 images	DSC = 0.79	No	Only WL images Inference speed: 0.205 s per image (≈5 FPS)
Wang, 2024 [21]	CNN	Leukoplakia segmentation and classification	5362 images + 50 videos (551 pts)	DSC = 66.4%, precision = 54.9%, recall = 51.2% (internal validation) DSC = 58.2, precision = 49.6%, recall = 46.5% (external validation)	Yes	Complete algorithm, studied to analyze the image, segment the ROI and provide classification of leukoplakias. Only WL images with AI, the AUC improved from 0.72 to 0.80 for senior clinicians and from 0.65 to 0.80 for junior clinicians
Nobel, 2024 [19]	Encoder-decoder + CNN	Vocal folds segmentation and disease classification	24,000 images	Segmentation of vocal folds: accuracy = 91.47%, DSC = 71.52%, Precision = 91.9%, IoU = 87.46%	No	No NBI Segmentation of vocal folds and not lesions
Tie, 2024 [20]	CNN	Laryngeal leukoplakia segmentation and classification	7057 images (426 pts)	DSC=0.61 on internal cohort, 0.42 on external dataset	Yes	Trained and tested on WL and NBI frames External validation + video validation for real time use + human-machine comparison AI significantly improved AUC and accuracy for all laryngologists ($p < 0.05$)
Xiong, 2024 [44]	CNN	Laryngeal leukoplakia segmentation and classification	6180 (WL = 3080, NBI = 3100)	WL: mAP50 = 81% NBI: mAP50 = 92%	No	Trained and tested on WL and NBI frames

WL: white light; NBI: narrow band imaging; LSCC: laryngeal squamous cell carcinoma; CNN: convoluted neural network; mAP50: mean average precision; ROI: region of interest; FPS: frames per second; DSC; dice similarity coefficient; IoU: intersection over union; pts: patients.

crepancy may be attributed to the anatomical complexity of the oral cavity, where teeth, mucosal folds, and taste buds can interfere with the model's ability to accurately delineate lesions.

It is noteworthy that all 3 studies were based on mixed training datasets containing both WL and NBI frames, and that their validation procedures included both internal datasets and external cohorts. This may partly account for the promising results reported by the authors, although the relatively small training datasets raise the possibility of selection bias (3393 images [50], 1034 images [16], and 683 images [17]).

Among the studies focusing on laryngeal leukoplakia, DSC values ranged from 61 to 78.7% [18,20,21]. The lower performance observed in these works may be attributed to the limited use of NBI – neither Ji et al. [18] nor Wang et al. [21] included this enhanced modality – which could have aided the model in more accurately defining lesion borders, particularly in hyperkeratotic conditions such as leukoplakia.

Focusing on clinical practice, the combination of NBI with automated assessment of cancer boundaries is of particular interest in transoral surgery of laryngeal suspicious lesions or carcinomas. Indeed, it is well established – and has been reported by numerous authors – that NBI can improve the detection rate of laryngeal SCC and reduce the rate of positive margins after surgery, thus improving oncological outcomes [51,52]. From a clinical standpoint, these findings suggest that automatic segmentation may serve multiple roles in the management of laryngeal cancer during both in office and intraoperative evaluation. On the other hand, the development of models capable of accurately delineating regions of interest during laryngoscopy could represent an important future advancement, particularly for assessing the completeness of laryngoscopic examinations or characterising laryngeal mobility alterations. This task is known to be particularly challenging for clinicians, as highlighted in a recent consensus paper on laryngeal cancer management [53], and has been explored in CV primarily through high-speed videoendoscopy or stroboscopy. To assess vocal fold hypomobility or fixation, several studies have investigated the segmentation of the glottal area or the vocal folds themselves [54,55]. Although this approach is computationally demanding and has yielded satisfactory results for analysing functional voice disorders, it still remains unsuitable for evaluating vocal fold fixation caused by laryngeal lesions. More recently, Koivu et al. [56] validated an alternative approach based on the automated detection of laryngeal key points, originally introduced by Villani et al. [57]. This strategy enabled the assessment of laryngeal mobility in non-neoplastic conditions such as muscle tension dysphonia, presbyphonia, laryngeal dystonia, and vocal fold palsy. However, these studies do not address the evaluation of laryngeal carcinoma-related hypomobility, as the lesion itself often obscures the key points required for accurate model prediction.

Conclusions

The integration of AI into laryngeal endoscopy represents a major advance in the diagnostic evaluation of laryngeal lesions. Evidence from the studies reviewed shows that AI can reach high diagnostic accuracy, often comparable to human performance. This is a significant advantage given the operator dependence and variability inherent to conventional endoscopy.

By outlining current algorithmic approaches and assessing their performance, this review highlights both the promise and the limitations of AI-assisted endoscopic assessment. Although existing data support their feasibility and potential value, methodological heterogeneity, scarce external validation, and the prevalence of single-centre datasets underscore the need for more rigorous and multicentric research.

Looking ahead, the growing incorporation of AI into clinical practice is expected to refine the characterisation of laryngeal lesions and support precision diagnostics. The emerging field of videomics, in particular, offers opportunities for non-invasive tumour profiling and integration of imaging biomarkers into clinical decision-making. As validation expands, AI-enhanced endoscopy could reshape diagnostic pathways in laryngeal oncology and contribute to more equitable, high-quality care.

Conflict of interest statement

The authors declare no conflict of interest.

Funding

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

Author contributions

AI: conceptualization; AI, EB, MSS, CS: methodology, investigation; AI, EB: writing – original draft preparation; AI, EB, FM, CS: writing – review and editing; AI, FM, DM, GP, PD, PP, OP, CS: supervision.

All authors have read and agreed to the published version of the manuscript.

Ethical consideration

Not applicable.

References

1 Wilmes C, Goril A, Marres H, et al. A systematic review of the clinical impact of implementing artificial intelligence in upper aerodigestive tract endoscopy. Head Neck 2025;47:2998-3018. https://doi.org/10.1002/hed.28213

2 Sawashima M, Hirose H. New laryngoscopic technique by use of fiber optics. J Acoust Soc Am 1968;43:168-169. https://doi.org/10.1121/1.1910752

3 Piazza C, Cocco D, De Benedetto L, et al. Narrow band imaging and high definition television in the assessment of laryngeal cancer: a prospective study on 279 patients. Eur Arch Otorhinolaryngol 2010;267:409-414. https://doi.org/10.1007/S00405-009-1121-6

4 Vilaseca I, Valls-Mateus M, Nogués A, et al. Usefulness of office examination with narrow band imaging for the diagnosis of head and neck squamous cell carcinoma and follow-up of premalignant lesions. Head Neck 2017;39:1854-1863. https://doi.org/10.1002/hed.24849

5 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;260:108539. https://doi.org/10.1016/j.cmpb.2024.108539

6 Sampieri C, Peretti G. Democratizing cancer detection: artificial intelligence-enhanced endoscopy could address global disparities in head and neck cancer outcomes. Eur Arch Otorhinolaryngol 2025;282:2739-2743. https://doi.org/10.1007/s00405-025-09257-4

7 Kang YF, Yang L, Xu K, et al. A lightweight intelligent laryngeal cancer detection system for rural areas. Am J Otolaryngol 2024;45:104474. https://doi.org/10.1016/j.amjoto.2024.104474

8 Sampieri C, Baldini C, Azam MA, et al. Artificial intelligence for upper aerodigestive tract endoscopy and laryngoscopy: a guide for physicians and state-of-the-art review. Otolaryngol Head Neck Surg 2023;169:811-829. https://doi.org/10.1002/ohn.343

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

10 Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:N71. https://doi.org/10.1136/bmj.n71

11 Marrero-Gonzalez AR, Diemer TJ, Nguyen SA, et al. Application of artificial intelligence in laryngeal lesions: a systematic review and meta-analysis. Eur Arch Otorhinolaryngol 2025;282:1543-1555. https://doi.org/10.1007/s00405-024-09075-0

12 Żurek M, Jasak K, Niemczyk K, et al. Artificial intelligence in laryngeal endoscopy: systematic review and meta-analysis. J Clin Med 2022;11:2752. https://doi.org/10.3390/jcm11102752

13 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. https://doi.org/10.1016/j.ebiom.2019.08.075

14 Kuo CFJ, Lai WS, Barman J, et al. Quantitative laryngoscopy with computer-aided diagnostic system for laryngeal lesions. Sci Rep 2021;11:10147. https://doi.org/10.1038/s41598-021-89680-9

15 Li Y, Gu W, Yue H, et al. Real-time detection of laryngopharyngeal cancer using an artificial intelligence-assisted system with multimodal data. J Transl Med 2023;21:698. https://doi.org/10.1186/s12967-023-04572-y

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

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

18 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. https://doi.org/10.1016/j.bspc.2020.101913

19 Nobel SMN, Swapno SMMR, Islam MR, et al. A machine learning approach for vocal fold segmentation and disorder classification based on ensemble method. Sci Rep 2024;14:14435. https://doi.org/10.1038/s41598-024-64987-5

20 Tie C, Li D, Zhu J, et al. Multi-instance learning for vocal fold leukoplakia diagnosis using white light and narrow-band imaging: a multicenter study. Laryngoscope 2024;134:4321-4328. https://doi.org/10.1002/lary.31537

21 Wang ML, Tie CW, Wang JH, et al. Multi-instance learning based artificial intelligence model to assist vocal fold leukoplakia diagnosis: a multicentre diagnostic study. Am J Otolaryngol 2024;45:104342. https://doi.org/10.1016/j.amjoto.2024.104342

22 Nie X, Zhang X, Wang D, et al. Laryngeal cancer diagnosis based on improved YOLOv8 algorithm. Mach Learn Sci Technol 2025;6:015011. https://doi.org/10.1088/2632-2153/ada2d9

23 Fang S, Fu J, Du C, et al. Identifying laryngeal neoplasms in laryngoscope images via deep learning based object detection: a case study on an extremely small data set. IRBM 2023;44:100799. https://doi.org/10.1016/j.irbm.2023.100799

24 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;20:1-10. https://doi.org/10.1186/s12938-021-00886-4

25 Bhattacharjee R, Suganya Devi K, Vijaykanth S. Detecting laryngeal cancer lesions from endoscopy images using deep ensemble model. In: 2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT). IEEE 2023:1-6. https://doi.org/10.1109/IConSCEPT57958.2023.10170113

26 Cen Q, Pan Z, Li Y, et al. Laryngeal tumor detection in endoscopic images based on convolutional neural network. In: 2019 IEEE 2nd International Conference on Electronic Information and Communication Technology (ICEICT). IEEE 2019:604-608. https://doi.org/10.1109/ICEICT.2019.8846399

27 Wellenstein DJ, Woodburn J, Marres HAM, et al. Detection of laryngeal carcinoma during endoscopy using artificial intelligence. Head Neck 2023;45:2217-2226. https://doi.org/10.1002/hed.27441

28 Bur AM, Zhang T, Chen X, et al. Interpretable computer vision to detect and classify structural laryngeal lesions in digital flexible laryngoscopic images. Otolaryngol Head Neck Surg 2023;169:1564-1572. https://doi.org/10.1002/ohn.411

29 Zhang X, Zhao J, Zong D, et al. Taming vision transformers for clinical laryngoscopy assessment. J Biomed Inform 2025;162:104766. https://doi.org/10.1016/j.jbi.2024.104766

30 Ramos LT, Sappa AD. A decade of You Only Look Once (YOLO) for object detection: a review. IEEE Access 2025;13:192747-192794. https://doi.org/10.1109/ACCESS.2025.3630988

31 Missale F, Taboni S, Carobbio ALC, et al. Validation of the European Laryngological Society classification of glottic vascular changes as seen by narrow band imaging in the optical biopsy setting. Eur Arch

Otorhinolaryngol 2021;278:2397-2409. https://doi.org/10.1007/s00405-021-06723-7

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

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

[34] Qiu M, Li Y, Huang W, et al. 3D-LSPTM: an automatic framework with 3D-large-scale pretrained model for laryngeal cancer detection using laryngoscopic videos. In: 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE 2024:1-4. https://doi.org/10.1109/EMBC53108.2024.10781710

[35] Xu ZH, Fan DG, Huang JQ, et al. Computer-aided diagnosis of laryngeal cancer based on deep learning with laryngoscopic images. Diagnostics 2023;13:3669. https://doi.org/10.3390/diagnostics13243669

[36] Li Y, Gu W, Yue H, et al. Real-time detection of laryngopharyngeal cancer using an artificial intelligence-assisted system with multimodal data. J Transl Med 2023;21:698. https://doi.org/10.1186/s12967-023-04572-y

[37] Kim GH, Hwang YJ, Lee H, et al. Convolutional neural network-based vocal cord tumor classification technique for home-based self-prescreening purpose. Biomed Eng Online 2023;22:81. https://doi.org/10.1186/s12938-023-01139-2

[38] Zhao Q, He Y, Wu Y, et al. Vocal cord lesions classification based on deep convolutional neural network and transfer learning. Med Phys 2022;49:432-442. https://doi.org/10.1002/mp.15371

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

[40] Ren J, Jing X, Wang J, et al. Automatic recognition of laryngoscopic images using a deep-learning technique. Laryngoscope 2020;130:E686-E693. https://doi.org/10.1002/lary.28539

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

[42] Yao P, Witte D, German A, et al. A deep learning pipeline for automated classification of vocal fold polyps in flexible laryngoscopy. Eur Arch Otorhinolaryngol 2024;281:2055-2062. https://doi.org/10.1007/s00405-023-08190-8

[43] Yan P, Li S, Zhou Z, et al. Automated detection of glottic laryngeal carcinoma in laryngoscopic images from a multicentre database using a convolutional neural network. Clin Otolaryngol 2023;48:436-441. https://doi.org/10.1111/coa.14029

[44] Xiong M, Luo JW, Ren J, et al. Applying deep learning with convolutional neural networks to laryngoscopic imaging for automated segmentation and classification of vocal cord leukoplakia. Ear Nose Throat J 2024;20:1455613241275341. https://doi.org/10.1177/01455613241275341

[45] Kang Y, Yang L, Hu Y, et al. Self-attention mechanisms-based laryngoscopy image classification technique for laryngeal cancer detection. Head Neck 2025;47:944-955. https://doi.org/10.1002/hed.27999

[46] You Z, Han B, Shi Z, et al. Vocal cord leukoplakia classification using deep learning models in white light and narrow band imaging endoscopy images. Head Neck 2023;45:3129-3145. https://doi.org/10.1002/hed.27543

[47] You Z, Han B, Shi Z, et al. Single-view contrastive learning for laryngeal leukoplakia classification with NBI laryngoscopy images. Head Neck 2025;47:2584-2593. https://doi.org/10.1002/hed.28157

[48] Yin L, Liu Y, Pei M, et al. Laryngeal image dataset and classification of laryngeal disease based on attention mechanism. Pattern Recognit Lett 2021;150:207-213. https://doi.org/https://doi.org/10.1016/j.patrec.2021.06.034.

[49] Azam MA, Sampieri C, Ioppi A, et al. Automatic delineation of laryngeal squamous cell carcinoma during endoscopy. Biomed Signal Process Control 2024;88:105666. https://doi.org/10.1016/j.bspc.2023.105666

[50] Sampieri C, Azam MA, Ioppi A, et al. Real-time laryngeal cancer boundaries delineation on white light and narrow-band imaging laryngoscopy with deep learning. Laryngoscope 2024;134:2826-2834. https://doi.org/10.1002/lary.31255

[51] 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;124:294-298. https://doi.org/10.1177/0003489414556082

[52] Bertino G, Cacciola S, Fernandes WB, et al. Effectiveness of narrow band imaging in the detection of premalignant and malignant lesions of the larynx: validation of a new endoscopic clinical classification. Head Neck 2015;37:215-222. https://doi.org/10.1002/HED.23582

[53] 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. https://doi.org/10.1016/j.oraloncology.2024.106744

[54] Yousef AM, Deliyski DD, Zacharias SRC, et al. Spatial segmentation for laryngeal high-speed videoendoscopy in connected speech. J Voice 2023;37:26-36. https://doi.org/10.1016/j.jvoice.2020.10.017

[55] Yan Y, Chen X, Bless D. Automatic tracing of vocal-fold motion from high-speed digital images. IEEE Trans Biomed Eng 2006;53:1394-1400. https://doi.org/10.1109/TBME.2006.873751

[56] Koivu A, Nwosu OI, Ota M, et al. Feasibility of real-time automated vocal fold motion tracking for in-office laryngoscopy. Laryngoscope 2026;136:596-604. https://doi.org/10.1002/lary.70104

[57] Villani FP, Fiorentino MC, Federici L, et al. A deep-learning approach for vocal fold pose estimation in videoendoscopy. J Imaging Inform Med 2026;39:842-852. https://doi.org/10.1007/s10278-025-01431-8

Supplementary online material

Search strings

MEDLINE

(((“Endoscopy”[Mesh] OR “Laryngoscopy”[Mesh] OR “Image*”[tw] OR “Photo*”[tw] OR “Fibroscopy”[All fields] OR “Videolaryngoscop*”[tw] OR “Videoendoscope*”[tw] OR “Video*”[tw]) AND (“Larynx”[All fields] OR “Upper Airway”[All fields])) AND (“Machine Learning”[Mesh] OR “Deep Learning”[tw] OR “Convolutional Neural Networks”[tw] OR “Artificial Intelligence”[Mesh] OR “Neural Network*”[tw] OR “Detection”[tw] OR “Segmentation”[tw] OR “Classification”[tw] OR “Computer Vision”[All fields])) NOT (“Pathology”[Mesh] OR “Histology”[Mesh] OR “Genomics”[Mesh] OR “Radiomics”[All fields] OR “Radiology”[Mesh] OR “Tomography”[All fields] OR “Magnetic Resonance Imaging”[All fields] OR “Sleep disorder*”[All fields] OR “Drug-Induced Sleep Endoscopy”[tw] OR “DISE”[tw] OR “Positron-Emission Tomography”[Mesh] OR “Radiomic”[tw] OR “Genetic*”[tw])

SCOPUS

(TITLE-ABS-KEY(“Endoscopy” OR “Laryngoscopy” OR “Image*” OR “Photo*” OR “Fibroscopy” OR “Videolaryngoscop*” OR “Videoendoscope*” OR “Video*”) AND TITLE-ABS-KEY(“Larynx” OR “Upper Airway”)) AND (TITLE-ABS-KEY(“Machine Learning” OR “Deep Learning” OR “Convolutional Neural Networks” OR “Artificial Intelligence” OR “Neural Network*” OR “Detection” OR “Segmentation” OR “Classification” OR “Computer Vision”)) AND NOT (TITLE-ABS-KEY(“Pathology” OR “Histology” OR “Genomics” OR “Radiomics” OR “Radiology” OR “Tomography” OR “Magnetic Resonance Imaging” OR “Sleep disorder*” OR “Drug-Induced Sleep Endoscopy” OR “DISE” OR “Positron-Emission Tomography” OR “Radiomic” OR “Genetic*”))