UNIVERSITÀ
DEGLI STUDI
DI BRESCIA

DOTTORATO DI RICERCA IN INTELLIGENZA ARTIFICIALE IN MEDICINA E INNOVAZIONE NELLA RICERCA CLINICA E METODOLOGICA

CICLO
XXXVI

SCALING ARTIFICIAL INTELLIGENCE IN ENDOSCOPY: FROM MODEL DEVELOPMENT TO MACHINE LEARNING OPERATIONS FRAMEWORKS

Settore Scientifico Disciplinare: MED/31 OTORINOLARINGOIATRIA

Dottorando: Alberto Paderno

Supervisore: Prof. Davide Farina

Summary

Introduction

The burgeoning domain of artificial intelligence (AI) has witnessed an exponential growth in recent years, spurred by notable technological and methodological advancements. In a short period of time, the typical approaches in computer vision went from manual feature extraction and subsequent analysis by basic machine learning algorithms, to direct application of neural networks. Furthermore, a landmark transition was the application of the transformer architecture, such as GPT (Generative Pre-trained Transformers), in natural language processing. This advancement has not only revolutionized everyday applications but has also begun to significantly impact the medical sector. From this development, the vision transformer architecture emerged as a promising strategy to address challenges within computer vision, consistently achieving state-of-the-art results and in some cases surpassing the previous dominance of convolutional neural networks.

This investigation started with a comprehensive review of the current advancements in AI and computer vision within Otolaryngology – Head and Neck Surgery. This understanding facilitated the delineation of specific objectives, goals, and areas that have been relatively under-explored. The primary aim of this step was to ascertain the prerequisites for creating a system dedicated to computer vision analysis in endoscopy and surgical procedures. Once the primary goal was established, a rigorous technical assessment was undertaken to discern the most fitting methodological approach and architectural paradigm for the problem at hand.

Initially, the endeavor was to design a computer vision tool capable of detecting and segmenting upper aerodigestive tract (UADT) neoplasms, transitioning from a preliminary proof of concept to a robust multi-site algorithm. This primarily addressed the challenges associated with static images. Additionally, our research ventured into the dynamic aspect of vision: movement. We designed a deep learning model to pinpoint laryngeal pivot points and assess vocal fold motility, a crucial determinant in laryngeal cancer staging. Both methodologies underwent iterative refinement throughout the course of the research.

In the concluding phase, the research pivoted towards the latest innovations in AI: vision foundation models (or Large Vision Models). The advent of vision transformer-based foundation models exudes immense potential across various computer vision tasks, including the medical field. Notably, vision transformers, when trained via self-supervision, could potentially mitigate the long-standing challenge of requiring extensive expert annotations, thereby profoundly benefiting niche fields with limited cases, such as Otolaryngology – Head and Neck Surgery. Throughout this research, various experiments underscored the promise held by these generic foundation models in medical image classification, even within the constraints of few-shot learning.

Finally, to transition from a research paradigm to tangible clinical applications, a complex infrastructure to manage processes and policies is strictly required. This realm, recently coined as MLOps (Machine Learning Operations), was delved into by devising a bespoke web application aimed at enhancing and expediting the annotation process in upper aerodigestive tract (UADT) endoscopy.

The subsequent section will provide a broader overview of the research path that has been followed throughout this exploration, describing the rationale and context for each step.

1. **Exploration phase: What is there, what is needed, and how can we achieve that.**

In the first step, the aim of the research was to precisely define the conceptual framework and methodology of the study. In fact, this represents one of the first applications of artificial intelligence in this field. A broad and comprehensive review of the international literature on AI application in clinical endoscopy has been performed, providing the first direct definition of this branch of research, and proposing the term "Videomics". This term denotes a burgeoning field where several computer vision and deep learning methods are systematically used to interpret the unstructured data of video obtained during diagnostic video-endoscopy.

In the available literature, Narrow Band Imaging (NBI), an optical biopsy technique applied in the field of clinical endoscopy, provided the best results in terms of diagnostic performance when using AI-based algorithms. For this reason, we selected NBI as the primary imaging technique in some of the subsequent studies.

Furthermore, we identified the best performing analytic methods to be applied to video-endoscopies, initially selecting convolutional neural networks (CNNs) among the wide variety of AI-based algorithms. Different types of CNNs have also been tested to identify the ones most fitted for each computational task.

2. Setting a plan: Identifying objectives and future perspectives

After the first general review, we better classified the different objectives and technical approaches. Future efforts in the field of endoscopic computer vision can be framed around five broad tasks with increasing complexity and computational load, which can be summarized as follows:

- Quality assessment of endoscopic images;
- Classification of pathologic and non-pathologic frames;
- Detection of lesions inside frames;
- Segmentation of pathologic lesions;
- In-depth characterization of neoplastic lesions.

This provided an initial structure for the subsequent research, clarifying progressive steps and objectives.

3. Defining standards: Guidelines for adequate reporting in medical computer vision studies

The in-depth review of the literature revealed critical flaws in the current reporting of computer vision research in the medical domain. Here, we defined and proposed standards and guidelines to adequately frame and report studies on artificial intelligence on medical images. In particular, we cover the domains of study definition and objectives, technical specifications, and selection of adequate endpoints.

4. Proof of concept: Setting the acquisition pipeline and testing technical approaches.

In this step, we performed a preliminary study with a limited number of frames (N=226) to standardize each phase of the analysis: endoscopic evaluation, conversion of the video files, extraction of the frames, manual annotation, optimization of the images, and algorithm training. The study identified different critical aspects that allowed for optimization of the entire process for the subsequent cases. Homogeneous illumination turned out to be a critical factor impacting the diagnostic potential of the CNNs. Similarly, the application of the CNN-based algorithms to distinct subsites of the upper aerodigestive tract showed different diagnostic performances according to the complexity and the number of diverse structures in the region, this aspect has been better investigated in the subsequent section of this research process.

After this first experience, a specific application was introduced to speed up the annotation process (i.e., Labelme). This approach was progressively improved during the research process, going from better local applications (i.e., Label Studio), to dedicated online platforms (i.e., Roboflow and Hasty.ai).

5. From proof of concept to structured clinical study: Overall assessment of CNNs for the semantic segmentation of UADT lesions.

After the promising results of the proof-of-concept study. We performed a more extensive evaluation (N=1034 endoscopic frames) including various anatomical subsites and a more innovative technical approach (i.e., instance segmentation). According to the previously available data, we focused our evaluation on NBI frames. With this work, we proved the feasibility of multi-object segmentation and classification in the endoscopy field, and we highlighted the significant differences in performance when comparing various UADT subsites.

6. From static vision to motion: The role of vocal fold motility.

Vocal fold motility has critical importance in laryngeal cancer staging. This variable is one of the key determinants of the cT category in the TNM staging, differentiating between cT1, cT2, and cT3 lesions. For this reason, a purely morphologic automatic identification of tumor extension would not be sufficient to achieve complete tumor staging. However, preliminary research showed the possibility of distinguishing between normal vocal fold motility and vocal fold paralysis through AI-based methods. Our aim was to extend this analysis to oncologic cases, providing an objective evaluation of vocal fold motility and potentially predicting vocal muscle infiltration.

7. Artificial intelligence models as a commodity: Standardized workflows for out-of-the-box training of pivotal vision models – The nasal polyps case-study.

The rapid pace of development in artificial intelligence led to the introduction of multiple "general purpose" vision algorithms that are easily finetuned for each specific application. This is associated to the development of novel tools for image collection and annotation, allowing to significantly scale a previously lengthy and burdensome process.

Here, we approached a simple case-study (i.e., nasal polyps) to prove the commoditization of machine learning algorithm. In fact, in the current stage of development, it is possible to apply "pre-packaged" models (e.g., YOLOv8) to highly specific tasks with minimal training and still achieve optimal results in terms of diagnostic performance.

8. Foundation models in clinical endoscopy: Proof of concept classification task with minimal training data – The oropharyngeal cancer case-study.

A paradigmatic shift in the realm of medical computer vision has been the development and deployment of transformer-based vision foundation models (such as DINOv2 by Meta). These algorithms are trained with a self-supervised approach on vast datasets, enabling them to derive reliable embeddings from each image, which are vector representations that encapsulate and distill visual patterns and objects.

Our investigation was centered around a dual objective: Primarily, to evaluate the diagnostic prowess of embeddings derived from DINOv2 in differentiating between normal mucosa and neoplastic tissue within the oropharyngeal region. Secondarily, to situate this endeavor within the overarching discourse of foundation models and their applicability in analyzing complex medical images.

9. Foundation vision models for automatic feature extraction: New approaches to reduce subjectivity.

In this follow-up experiment, we evaluated the efficacy of large vision models, notably those based on the transformer architecture, in autonomously discerning image attributes, or embeddings, tailored for image classification. Specifically, our objective was to trace the trajectory of progress from earlier methodologies that hinged on manual feature extraction. By revisiting an oropharyngeal cancer dataset, previously assessed through manual extraction, we juxtaposed the diagnostic proficiencies offered by both NBI and white light imaging. Our findings underscored that our approach for automated extraction was not only simpler and quicker, but also consistently outperformed manual feature extraction in terms of classification performance, particularly with white light images. This showed that current approaches are becoming progressively independent of image pre-processing and enhancement.

10. Working at scale: MLOps platforms to allow large-scale training and deployment of computer vision AI models.

Working with the Poznan Supercomputing and Networking Center, we developed ENDO-CLOUD (Enhanced Network for Deep learning-Oriented Classification and Leveraging of Optimized Upper aerodigestive tract Data), an advanced cloud-based system tailored for videolaryngoscopy. Machine learning, especially deep learning, offers potential solutions for videoendoscopic image processing. However, a key challenge is data gathering. We developed ENDO-CLOUD to provide a web-based segmentation platform that performs automatic representative frame selection from endoscopic videos of the UADT. This platform allows for collaboration between medical experts, deep learning specialists, and system administrators, ensuring efficient data management, security, and user-friendly segmentation procedures.

1. Videomics: Bringing deep learning to diagnostic endoscopy

Exploration phase: What is there, what is needed, and how can we achieve that.

Introduction

Artificial intelligence is beginning to transform clinical medicine in specialties where large datasets of annotated images are an essential element of the clinical workflow. The use of machine learning (ML) algorithms has augmented human judgment by identifying adverse events in the operating room,[1] detect diabetic retinopathy,[2] and even identify skin cancer.[3] From radiology to pathology, deep learning[4] has promise in reaching a diagnostic accuracy comparable with that of human experts from automating the detection of pneumonia on chest roentgenograms[5] and CT scans,[6] to identifying clinically occult nodal metastasis in breast cancer.[7] This rapid pace of innovation suggests that the development of expert systems will soon be applicable in everyday clinical settings, providing real-time assistance to the physician in a variety of diagnostic tasks, using computer technology to improve the human vision and judgment.

Notwithstanding these promises, at present, little progress has been made in applying deep learning algorithms to video-endoscopy, which plays a prominent role in otorhinolaryngology, head and neck surgery, pulmonary medicine, and gastroenterology, as well as in thoracic and abdominal surgery. For the purpose of this section, we explore how automated analysis of unstructured data obtained by video-endoscopy can provide valuable information during initial diagnosis, measuring treatment-response, and assessing prognosis. Endoscopic evaluation has always been a crucial component of head and neck oncology (HNO), since tumor superficial spread assessment has a significant impact on treatment selection and may not be adequately quantified by conventional radiologic imaging. Often seen as inherently descriptive, deep learning has helped to convert subjective assessment into objective findings based on systematic evaluation of visual data seen on video, analogous to findings obtained by conventional techniques in genomics and proteomics.

Here, in fact, we propose the term ’Videomics’ as a burgeoning field wherein several methods of computer vision and deep learning are systematically used to organize the unstructured data of video obtained during diagnostic endoscopy. Indeed, in this review of the literature, we argue that based on a growing number of publications, a new discipline is emerging within the large field of computer vision and pattern recognition. Herein, we review this promising new field, assessing preliminary findings, potential and limitations, and consider future developments.

Machine learning in endoscopy

Because of the higher caseload compared with HNO, gastrointestinal endoscopy was the first field in which ML was effectively applied. For this reason, it is useful to first analyze the progress in this branch of endoscopy to identify potential advances that are applicable to evaluation of the upper aerodigestive tract (UADT). Even in this broader research field, reports assessing the role of ML are scarce. As of today, efforts have been directed toward different lines of analysis, in particular: (1) blind-spot detection and automatic quality control; (2) lesion detection; (3) lesion classification; and (4) lesion characterization. This approach is strictly related to the perceived needs in digestive system endoscopy and has given promising results and potential real-life applications.

Concerning blind-spot detection, Wu et al.[8] developed a convolutional neural network (CNN)-based system aimed at detecting early gastric cancer although avoiding blind-spots during esophagogastroduodenoscopy. The algorithm was trained to identify the different subsites of the esophagus and stomach to ensure the complete visualization of the entire gastroesophageal mucosa. Further- more, CNN was trained to distinguish between nor- mal mucosa and early gastric cancer. In both tasks, the accuracy was remarkable (>90%). The same authors[9] validated the efficacy of this blind-spot detection system in a randomized controlled trial, showing a significantly lower blind-spot rate in CNN-assisted endoscopy vs. a control group (5.9% vs. 22.5%).

In the same perspective, Su et al.[10] developed an automatic quality control system aimed at improving diagnostic accuracy during colonoscopy. The system was based on CNN models for timing the withdrawal phase, supervise withdrawal stability, evaluate bowel preparation, and detect colorectal polyps. A randomized controlled trial showed that this CNN-based quality control system significantly increased lesion detection (adenomas and polyps) during colonoscopy compared to that without CNN assistance.

Lesion detection and characterization remain the main objective of ML-based strategies in gastrointestinal endoscopy. Texture analysis has shown good preliminary results in detecting mucosal abnormalities (e.g., colon polyps),[11] and CNNs proved to be a key instrument in this field. In fact, the vast majority of recent reports on automatic lesion detection and classification have taken advantage of this algorithm architecture. Different authors have described its significant potential in the detection and diagnosis of gastric, esophageal, and small bowel cancers, as well as gastrointestinal polyps.[12-16] Furthermore, CNNs are also useful in classification tasks, distinguishing between normal and inflamed mucosa (gastritis), and identifying early gastric cancer using magnifying endoscopy.[17-19]

Interestingly, although some authors employed conventional white light (WL) endoscopy, most studies have applied ML evaluation to Narrow Band Imaging (NBI) pictures or videos. In this view, multispectral imaging may have the potential to further improve detection and characterization of mucosal lesions in the field of automatic analysis, adding more definition to tumor margins and highlighting features of submucosal vascularization that are not evident during WL endoscopy. In 2012, Takemura et al.[20] demonstrated the value of NBI in the classification of colonoscopy magnified images using support vector machines. This aspect was explicitly investigated by Horie et al.,[14] who reported that NBI had a higher sensitivity compared with conventional WL endoscopy (although not reaching a statistically significant difference).

Finally, ML has shown promise in the in-depth characterization of known lesions of the gastrointestinal tract and may also provide risk stratification for malignant transformation of nonneoplastic mucosa. Specifically, recent studies[21,22] have demonstrated that CNNs can differentiate between early and deeply infiltrating gastric cancer. This result shows the potential of Videomics approaches to go beyond simple diagnosis and extract more extensive information on the lesion itself. Nakahira et al.[22] further confirmed this potential by showing that CNN is able of correctly stratify the risk of gastric tumor development by analyzing the non- neoplastic mucosa at video-endoscopy.

Machine learning applications in upper aero-digestive tract endoscopy

Video-endoscopic evaluation of the UADT poses even more challenges than gastrointestinal endoscopy.[23] This anatomic region is, in fact, structurally more complex, composed of a wide variety of tissues,[24,25] and easily shaded. Furthermore, deglutition, gag, and cough reflexes often come into play, interrupting or limiting the observation. The oral cavity and oropharynx are the most accessible sites; however, their video-endoscopic evaluation is not standardized and may be performed using rigid or flexible endoscopes, or even external cameras. Therefore, this factor adds an adjunctive layer of complexity to image analysis since data collection should be ideally standardized and characterized by low variance.

Oral cavity and oropharynx

Different authors[26,27] have recognized the value of ML in the evaluation and screening of oral cancer and potentially malignant lesions. Song et al.[28] developed a smartphone-based automatic image classification system for oral dysplasia and malignancy employing CNNs. The system aimed to screen high-risk populations in middle- and low-income countries and took advantage of dual-modality images (WL and autofluorescence). The authors demonstrated the potential of dual-modal image analysis, which showed better diagnostic performance than single-modal images. The final model reached an accuracy of 87%, sensitivity of 85%, and specificity of 89%.

Mascharak et al.[29] were the first to use ML to better identify oropharyngeal tumor margins using a simple naıve Bayesian classifier (color and texture). Interestingly, the diagnostic performance was significantly enhanced by multispectral NBI compared with conventional WL video-endoscopy. Five-fold cross-validation yielded an area under the curve (AUC) above 80% for NBI models and below 55% for WL endoscopy models ($P < 0.001$).

Finally, Paderno et al.[30] published preliminary data showing that it is possible to obtain real- time oral and oropharyngeal tumors segmentation using different fully CNNs applied to NBI endoscopic images, identifying potential confounding factors and technical drawbacks.

Larynx and hypopharynx

In general, laryngo-pharyngeal lesions are those more frequently investigated when assessing the role of automatic analysis by ML. This is due to use of a standardized endoscopic approach through trans-nasal or transoral video-endoscopy and the relative similarity with gastrointestinal subsites. In 2014, Huang et al.[31] proposed an automatic system aimed at recognizing images of the glottis and classifying different vocal fold disorders. The technique was based on a support vector machine classifier and reached an accuracy of 99%. However, the patterns to be classified were limited to 'normal vocal fold,' 'vocal fold paralysis,' 'vocal fold polyp,' and 'vocal fold cyst,' and did not include dysplasia or malignancy.

A preliminary attempt at automatic detection and classification of laryngeal tumors has been described by Barbalata et al.[32] The authors used anisotropic filtering to analyze the submucosal vasculature of normal and neoplastic laryngeal mucosa during NBI video-endoscopic examination, obtaining an overall classification accuracy of 83%. Although not employing adaptive algorithms, the study confirmed the value of NBI in maximizing feature extraction in endoscopic images.

Subsequent studies, focusing on the diagnosis and classification of pharyngo-laryngeal lesions at video-endoscopy, extensively employed CNNs and demonstrated remarkable results. A work by Laves et al.[33] used CNNs to segment a novel 7-class (void, vocal folds, other tissue, glottal space, pathology, surgical tools, and tracheal tube) dataset of the human larynx during transoral laser microsurgery. The dataset, consisting of 536 manually segmented endoscopic images, was tested to monitor the morphological changes and autonomously detect pathologies. Different CNN architectures were investigated, and a weighted average ensemble net- work of UNet and ErfNet (two of the most used CNNs in the current literature on this topic) was the best suited for laryngeal segmentation with a mean Intersection-over-Union (IoU) evaluation metric of 84.7%.

Xiong et al.[34] developed a CNN-based diagnostic system trained using 13,721 laryngoscopic images of cancer, premalignant lesions, benign alterations, and normal tissue collecting exams across several centers in China. The CNN distinguished malignant/premalignant lesions from benign ones and normal tissues with an accuracy of 87% (sensitivity 73%, specificity 92%, and AUC 92%). Ren et al.[35] described a similar approach, training the CNN with a total of 24,667 laryngoscopy images (normal, vocal nodule, polyps, leukoplakia, and malignancy), and achieving an overall accuracy of 96%. Strikingly, the CNN-based classifier outperformed physicians in the evaluation of the abovementioned conditions.

Further detection and classification attempts have mainly taken advantage of NBI images, which yielded superior results in terms of diagnostic performance, as previously demonstrated by Mascharak et al.[29] in the oropharyngeal site, and confirmed by Tamashiro et al.[36] These studies[36-38] were performed in the setting of transoral esophagogastroduodenoscopy and were aimed at detecting incidental laryngo-pharyngeal cancer during the procedure. However, direct comparisons between the different studies may be misleading because of the heterogeneous definition of 'correct diagnosis.'

Tamashiro et al.[36] focused on pharyngeal cancer and reported an accuracy, sensitivity, and specificity of 67%, 80%, and 57%, respectively. These results were slightly improved when limiting the analysis to NBI frames only. The authors trained a 'Single Shot MultiBox Detector' with a total of 5,403 images. Adequate detection was considered as frames including less than 80% of the area with noncancerous sites. Kono et al.[37] showed similar results in pharyngeal cancer detection by using a mask region-based CNN trained with 4,559 images. Each frame was judged as cancer when its probability score was >0.60, and its dimensions overlapped with the cancer area by a factor of >0.20. Accuracy, sensitivity, and specificity were 66%, 92%, and 47%, respectively.

Finally, Inaba et al.[38] trained a CNN-based algorithm (RetinaNet) with sequential sets of images until reaching 400 frames of superficial laryngo- pharyngeal cancer and 800 frames of normal mucosa. The diagnostic accuracy gradually improved with the sequential addition of training images until reaching an accuracy, sensitivity, and specificity of 97%, 95%, and 98%, respectively. The definition of correct diagnosis was set with an IoU parameter > 0.4.

Future perspectives and conclusions

Videomics is an emerging discipline that has the potential to significantly improve human detection of clinically significant lesions during video-endos- copy across medical and surgical disciplines. Preliminary reports have shown promising diagnostic potential and demonstrated the ability of ML algorithms to provide adjunctive information on tumor characteristics, such as depth of infiltration and, hence, infer important tumor-related issues such as extra-visceral extension, submucosal spread, and risk of regional/distant metastases. However, as early 'proof-of-concept' studies are published, it is important to note that these efforts are not yet part of routine endoscopic examination. In this view, further advances may allow obtaining an ever-growing amount of data from video-endoscopic sequences, thus assisting in tumor staging, margin recognition, treatment planning, and prog- nostic assessment. Furthermore, features extracted from video-endoscopy may be integrated into broader '-omic' models (including radiomics, genomics, proteomics, salivaomics, etc.), thus creating a precise representation of a given tumor and/or fine- tune the assessment of a specific patient. This is a crucial step in the perspective of tailoring treatment and personalized medicine.

For Videomics to flourish and to deliver practical tools for clinicians in daily practice, it is imperative to create large-scale image and video repositories. Currently, many ongoing efforts are fragmented and highly variable in their approach: the anatomical regions investigated (upper or lower digestive tracts), quality of images (definition, focus, illumination, and color balance), type of spectral filters (WL, NBI, autofluorescence, others), and setting (office-based, intraoperative) vary widely. Nonetheless, the current bottleneck for the development of ML-based video-analysis techniques is represented by the need to manually annotate training images.

In this view, the development of self-supervised learning techniques using unlabeled data for CNN pretraining and training may significantly and progressively improve algorithms without the need for human intervention.[39]

Finally, study objectives (detection, classification, or segmentation) are often not clearly stated or distinguished in each study. Last but not least, the statistical definition of correct and incorrect diagnosis is subjectively determined by each author, leading to significant variation in diagnostic performance metrics (e.g., accuracy, sensitivity, specificity, and AUC). For this reason, at this early stage in the field, research teams should focus on standardization of data collection, identification of common targets, and optimal reporting. With such a collaborative stepwise approach, Videomics is likely soon augment human detection during endoscopy and improve cancer treatment and subsequent out- comes.

References

1. Gordon L, Austin P, Rudzicz F, Grantcharov T. MySurgeryRisk and Machine Learning: A Promising Start to Real-time Clinical Decision Support. Ann Surg 2019; 269:e14-e15.
2. Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA 2016; 316:2402-2410.

3. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542:115-118.
4. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015; 521:436-444.
5. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 2018; 15:e1002686.
6. Harmon SA, Sanford TH, Xu S, et al. Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. Nat Commun 2020; 11:4080.
7. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA 2017; 318:2199-2210.
8. Wu L, Zhou W, Wan X, et al. A deep neural network improves endoscopic detection of early gastric cancer without blind spots. Endoscopy 2019; 51:522-531.
9. Wu L, Zhang J, Zhou W, et al. Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. Gut 2019; 68:2161-2169.
10. Su JR, Li Z, Shao XJ, et al. Impact of a real-time automatic quality control system on colorectal polyp and adenoma detection: a prospective randomized controlled study (with videos). Gastrointest Endosc 2019.
11. Geetha K, Rajan C. Automatic Colorectal Polyp Detection in Colonoscopy Video Frames. Asian Pac J Cancer Prev 2016; 17:4869-4873.

12. Barbosa DC, Roupar DB, Ramos JC, et al. Automatic small bowel tumor diagnosis by using multi-scale wavelet-based analysis in wireless capsule endoscopy images. Biomed Eng Online 2012; 11:3.

13. Billah M, Waheed S, Rahman MM. An Automatic Gastrointestinal Polyp Detection System in Video Endoscopy Using Fusion of Color Wavelet and Convolutional Neural Network Features. Int J Biomed Imaging 2017; 2017:9545920.

14. Horie Y, Yoshio T, Aoyama K, et al. Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. Gastrointest Endosc 2019; 89:25-32.

15. Hirasawa T, Aoyama K, Tanimoto T, et al. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. Gastric Cancer 2018; 21:653-660.

16. Yoon HJ, Kim JH. Lesion-Based Convolutional Neural Network in Diagnosis of Early Gastric Cancer. Clin Endosc 2020; 53:127-131.

17. Ueyama H, Kato Y, Akazawa Y, et al. Application of artificial intelligence using a convolutional neural network for diagnosis of early gastric cancer based on magnifying endoscopy with narrow-band imaging. J Gastroenterol Hepatol 2020.

18. Li L, Chen Y, Shen Z, et al. Convolutional neural network for the diagnosis of early gastric cancer based on magnifying narrow band imaging. Gastric Cancer 2020; 23:126-132.

19. Horiuchi Y, Aoyama K, Tokai Y, et al. Convolutional Neural Network for Differentiating Gastric Cancer from Gastritis Using Magnified Endoscopy with Narrow Band Imaging. Dig Dis Sci 2020; 65:1355-1363.

20. Takemura Y, Yoshida S, Tanaka S, et al. Computer-aided system for predicting the histology of colorectal tumors by using narrow-band imaging magnifying colonoscopy (with video). Gastrointest Endosc 2012; 75:179-185.

21. Yoon HJ, Kim S, Kim JH, et al. A Lesion-Based Convolutional Neural Network Improves Endoscopic Detection and Depth Prediction of Early Gastric Cancer. J Clin Med 2019; 8.

22. Nakahira H, Ishihara R, Aoyama K, et al. Stratification of gastric cancer risk using a deep neural network. JGH Open 2020; 4:466-471.

23. Abe S, Oda I. Real-time pharyngeal cancer detection utilizing artificial intelligence: Journey from the proof of concept to the clinical use. Dig Endosc 2020.

24. Lin YC, Wang WH, Lee KF, et al. Value of narrow band imaging endoscopy in early mucosal head and neck cancer. Head Neck 2012; 34:1574-1579.

25. 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.

26. Yoshida K. Future Prospective of Light-Based Detection System for Oral Cancer and Oral Potentially Malignant Disorders by Artificial Intelligence Using Convolutional Neural Networks. Photobiomodul Photomed Laser Surg 2019; 37:195-196.

27. Kar A, Wreesmann VB, Shwetha V, et al. Improvement of oral cancer screening quality and reach: The promise of artificial intelligence. J Oral Pathol Med 2020.

28. Song B, Sunny S, Uthoff RD, et al. Automatic classification of dual-modalilty, smartphone-based oral dysplasia and malignancy images using deep learning. Biomed Opt Express 2018; 9:5318-5329.

29. Mascharak S, Baird BJ, Holsinger FC. Detecting oropharyngeal carcinoma using multispectral, narrow-band imaging and machine learning. Laryngoscope 2018; 128:2514-2520.

30. Paderno P, 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 2020 (in press).

31. Huang CC, Leu YS, Kuo CF, et al. Automatic recognizing of vocal fold disorders from glottis images. Proc Inst Mech Eng H 2014; 228:952-961.

32. Barbalata C, Mattos LS. Laryngeal Tumor Detection and Classification in Endoscopic Video. IEEE J Biomed Health Inform 2016; 20:322-332.

33. 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; 14:483-492.

34. 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.

35. Ren J, Jing X, Wang J, et al. Automatic Recognition of Laryngoscopic Images Using a Deep-Learning Technique. Laryngoscope 2020.

36. Tamashiro A, Yoshio T, Ishiyama A, et al. Artificial intelligence-based detection of pharyngeal cancer using convolutional neural networks. Dig Endosc 2020.

37. Kono M, Ishihara R, Kato Y, et al. Diagnosis of pharyngeal cancer on endoscopic video images by Mask region-based convolutional neural network. Dig Endosc 2020.

38. 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; 42:2581-2592.

39. Ross T, Zimmerer D, Vemuri A, et al. Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. Int J Comput Assist Radiol Surg 2018; 13:925-933.

2. Artificial intelligence in clinical endoscopy: Insights in the field of Videomics

Setting a plan: Identifying objectives and future perspectives

Introduction

As previously mentioned, the use of artificial intelligence (AI) is currently increasing in every field of medicine, progressively encompassing the entire patient care process, from embryo selection to survival prediction.[1] Machine learning (ML), a branch of AI, has the objective to automatically extract actionable insights from complex and large datasets that cannot (or are hard to) be effectively analyzed by conventional statistical methods or human intuition. ML algorithms have been designed to tackle the variability of "-omics" datasets (e.g., genomics, epigenomics, transcriptomics, proteomics, etc.), and unstructured data as radiologic images (radiomics), histopathology (pathomics), and surgical or videoendoscopic images (videomics).[2] In this area, ML is being applied to identify disease patterns and predict specific characteristics that may assist clinicians in diagnosis, therapeutic management, and follow-up.

While promising results have been obtained for -omics datasets, radiological and histopathologic images, analysis of videoendoscopic frames still represents a challenge. In this context, videomics represents a field wherein several methods of computer vision are systematically used to organize the unstructured data of frames obtained during diagnostic videoendoscopy. These applications are still in the early stage of development, especially in the field of otolaryngology, head and neck surgery, where the term videomics was first introduced.[2] In particular, one of the main limits in developing robust and automatic ML algorithms that can be translated in the clinical practice is represented by the paucity of annotated datasets to train algorithms.

Aims of Videomics

Diagnostic endoscopy is an essential component in assessment of the upper aerodigestive tract (UADT) and is a cornerstone as first-line diagnostic tool, especially after the introduction of the "bioendoscopy" concept.[3] The introduction of videoendoscopy significantly improved this field by the development of high-quality videorecording, image magnification, high-definition visualization, and advanced optical filters such as Narrow Bang Imaging (NBI), Storz Professional Image Enhancement System (SPIES or Image 1S), and I-Scan. These nuances, together with the constant advancement in ML, have opened new possibilities for image analysis in a computer vision-oriented approach. Here, deep learning (DL), a branch of ML, is playing a paramount role. In the field of supervised learning, when provided with both the "problem" (i.e., unlabeled videoendoscopic frame) and the "solution" (i.e., annotated frame or "ground truth"), DL algorithms iteratively learn their internal parameters (i.e., weights) to progressively improve diagnostic performance and specialize on a given objective. As described in this section, recent studies have focused on five broad tasks with increasing complexity and computational load, which can be summarized as follows:

- Quality assessment of endoscopic images (Figure 2.1);
- Classification of pathologic and non-pathologic frames (Figure 2.2);
- Detection of lesions inside frames (Figure 2.3);
- Segmentation of pathologic lesions (Figure 2.4);
- In-depth characterization of neoplastic lesions (Figure 2.5).

Figure 2.1. *Depiction of the potential input and output of a quality assessment algorithm.*

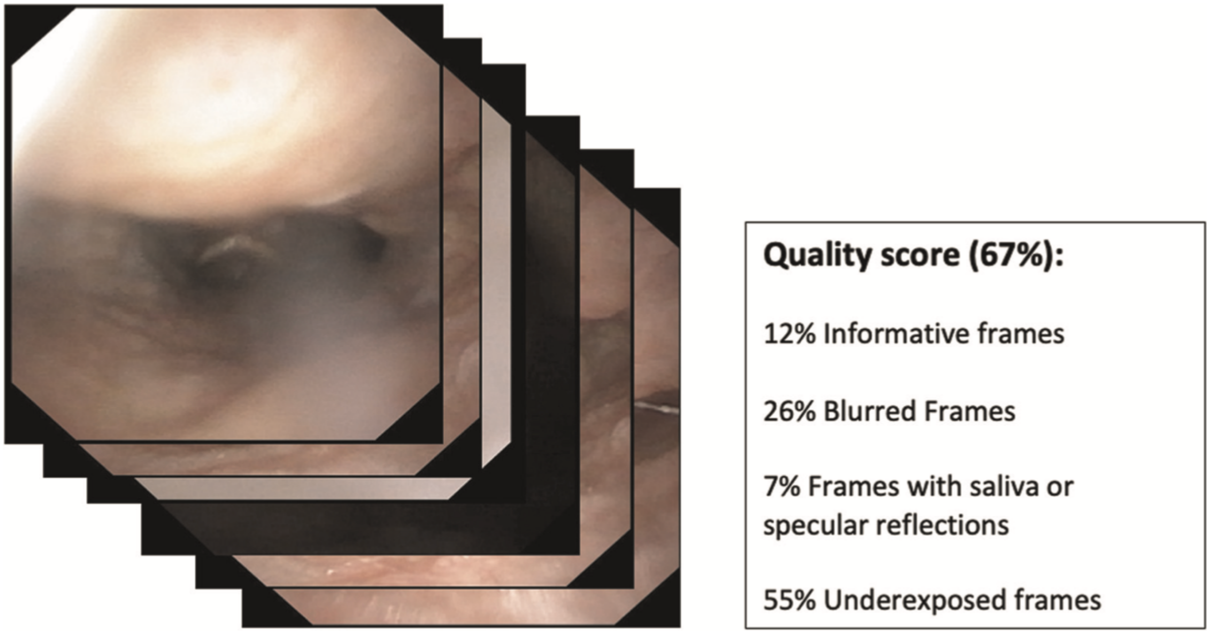


Figure 2.2. *Example of a classification task. The algorithm distinguishes between normal and pathologic frames without identifying the area involved by the disease.*

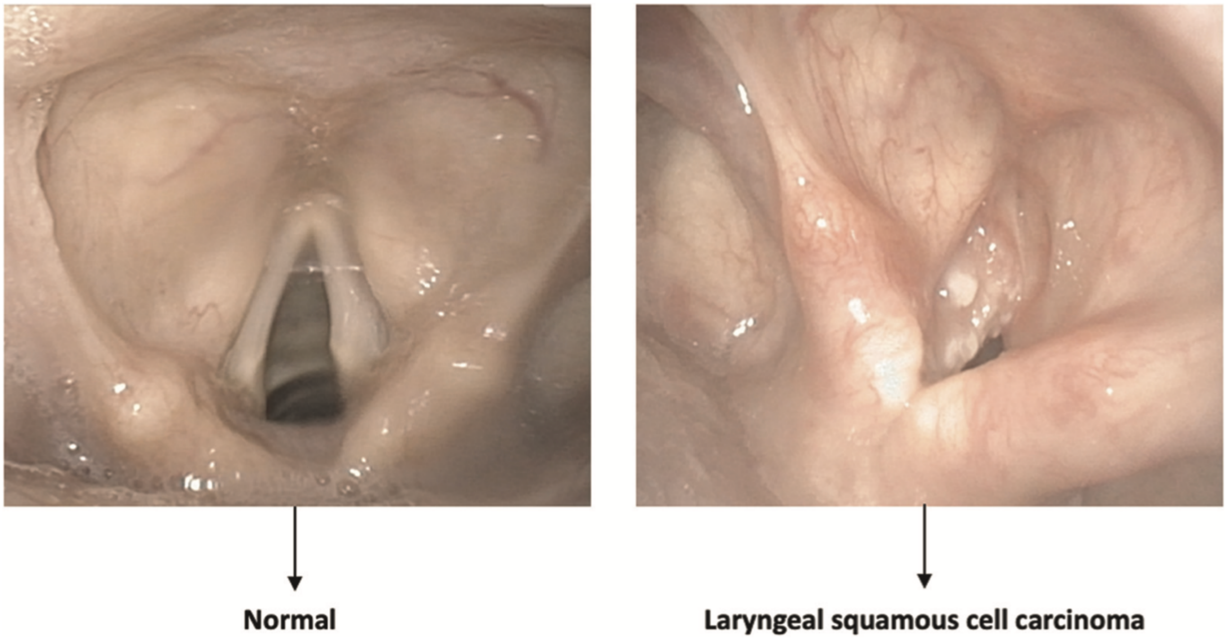


Figure 2.3. *Image showing a bounding box localizing a laryngeal lesion. This is the typical output of detection algorithms.*

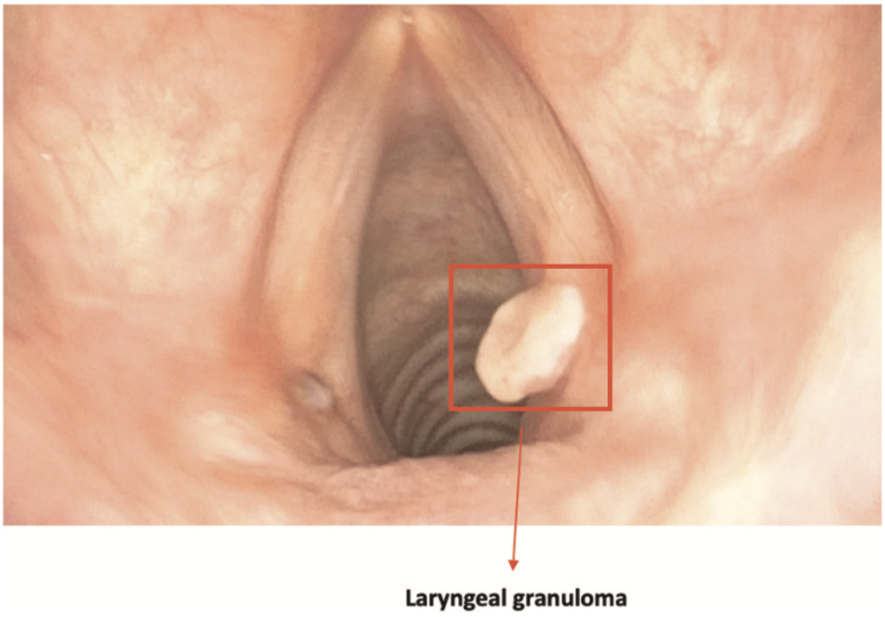


Figure 2.4. *Automatic segmentation of a laryngeal lesion provided by a convolutional neural network after adequate training and optimization.*

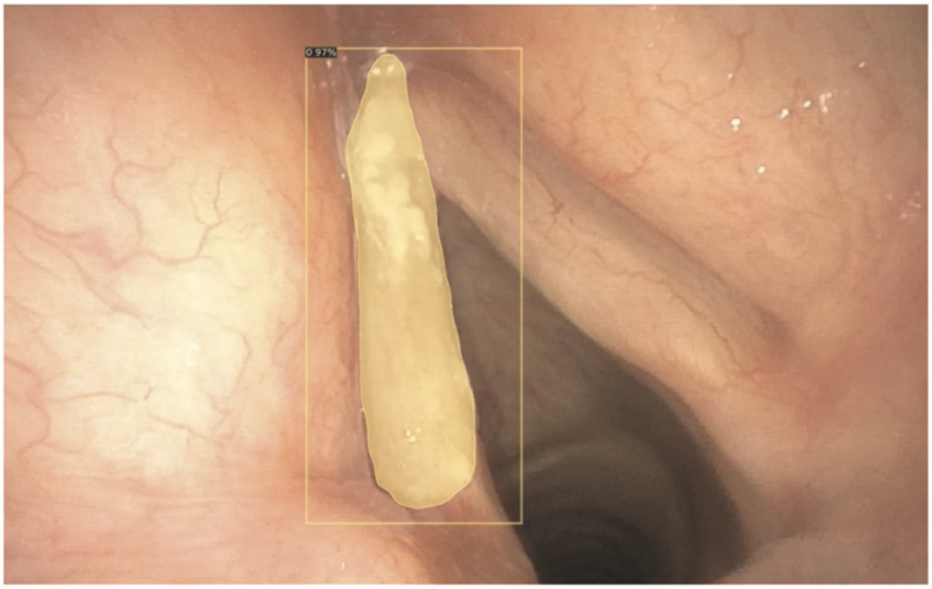
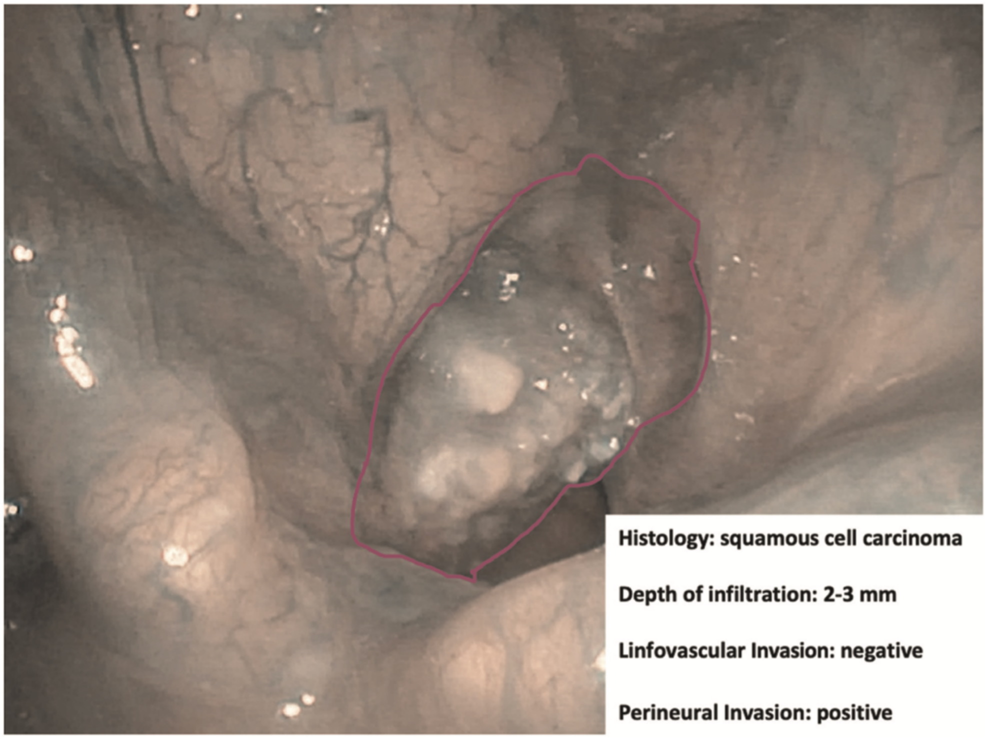


Figure 2.5. *Endoscopic NBI frame showing an example of adjunctive data drawn from in-depth characterization by hypothetical machine learning algorithms.*



Histology: squamous cell carcinoma

Depth of infiltration: 2-3 mm

Linfovascular Invasion: negative

Perineural Invasion: positive

In this area, a stepwise approach has the potential to make use of incremental refinements of algorithms and develop functional “minimal viable products” that can be introduced in clinical practice as early as possible, even without the full suite of the abovementioned applications. This is especially true considering that, as mentioned, the main limiting factor in this field is the paucity of large dedicated datasets that are usable for training. Gómez et al.[4] initially addressed this issue in the field of high-speed laryngeal videoendoscopy by collecting and publishing the BAGLS multihospital glottis segmentation dataset. However, with the progressive expansion of the available training images, it will be possible to tackle increasingly complex challenges.

Furthermore, the application of transfer learning techniques may significantly improve algorithm training and reduce the number of images needed to reach optimal performance. Transfer learning consists in pre-training the algorithm with images that are not directly correlated to the task to be explored, but which have certain similarities with the target dataset. For example, the weights of convolutional neural networks (CNNs) for endoscopic image analysis are often initialized with weights obtained by training CNNs on natural images from everyday objects (i.e., ImageNet dataset). This allows CNNs to detect generic low-level features (e.g., corners, edges). Pre-training with endoscopic images from a different anatomic site may provide an adjunctive advantage, especially in small datasets. CNNs are then fine-tuned to slightly adjust their parameters using endoscopic images.

A potential approach to address the low number of manually annotated images is offered by unsupervised and self-supervised learning. Unlike supervised learning, which is biased towards how it is being supervised, unsupervised learning derives insights directly from the data itself, groups the data, and helps to make data-driven decisions without external biases.[5] This approach may be particularly useful to cluster endoscopic frames into different categories (e.g., low visibility vs. good visibility) to help the clinician's assessment. On the other hand, self-supervised learning takes advantage of unlabeled images of the same pathology but captured from different views to significantly enhance the performance of pretraining. However, these options still need to be fully explored in the field of UADT endoscopy.[6]

Quality assessment

The first step in which AI can be effectively applied to diagnostic videoendoscopies is their quality control. In fact, in every examination, the majority of videoendoscopic frames are not diagnostic due to the presence of technical or patient-related factors that limit visualization. These factors, in the field of UADT evaluation, are mainly represented by repeated swallowing, gag reflex, secretions, blurring of the camera, specular reflections, and over- and underexposure. Automatic identification and classification of these issues can be of help in real time determination of the quality of an endoscopic examination, and may allow to automatically detect the most significant frames in a given recording.

In this field, Patrini et al.[7] developed a ML-based strategy for automatic selection of informative videolaryngoscopic frames. This approach resulted in a recall (i.e., true positive rate = true positives over true positives and false negatives) of 0.97 when classifying informative vs. uninformative frames (i.e., blurred, with saliva or specular reflections, and underexposed) with support vector machines (SVM) (i.e., conventional ML algorithms), and 0.98 with a CNN-based classification. Furthermore, their work demonstrated the potential of transfer learning in medical image analysis.

As a proof of concept, recent advances in the field of gastrointestinal endoscopy have led to the development of a fully automatic framework that can detect and classify different artifacts, segment artifact instances, provide a quality score for each frame, and restore partially corrupted frames.[8]

Classification

Classification is a typical task in the field of DL, distinguishing between normal and pathological mucosa. Here, the objective is not to localize or finely characterize a particular lesion, but rather to distinguish entire frames into different classes, usually pathologic vs. non-pathologic.

In this field, He et al.[9] applied CNN to interpret images of laryngeal squamous cell carcinoma using static NBI frames to determine whether a lesion was benign or malignant. The model reached an accuracy of 90.6%, a sensitivity of 88.8%, and a specificity of 92.2%. Furthermore, the authors demonstrated that the accuracy of the CNN model was higher than that of human experts. A similar approach was described by Esmaeili et al.,[10] training a CNN for the automatic classification of

NBI images into benign and malignant. A pre-trained ResNet50 architecture was adopted, and three experiments with several models were generated and validated. The model showed a striking diagnostic performance and achieved a testing accuracy of 0.83.

Considering multiple classification groups, Zhao et al.[11] proposed a four class-system of vocal cord targets (i.e., normal mucosa, polyp, keratinization, and carcinoma), and a laryngoscopy dataset was divided into "urgent" (keratinization, carcinoma) and "non-urgent" (normal mucosa, polyp) cases. An overall accuracy of 80.2%, a F1 score (i.e., the harmonic mean of the precision and recall, a measure of accuracy) of 0.78, and an area under the curve (AUC) of 0.96 were achieved. The proposed method delivered high classification performance of normal mucosa, polyps, and carcinoma in an extremely rapid time.

Other studies[12,13] have employed ML to classify pharyngo-laryngeal benign lesions during videoendoscopy, demonstrating notable results. A preliminary attempt was described in 2014 by Huang et al.,[12] who proposed an automatic system aimed at recognizing the dynamic image of the glottis and classifying different vocal fold disorders ("normal vocal fold," "vocal fold paralysis," "vocal fold polyp," and "vocal fold cyst"). This study used an SVM classifier and reached an accuracy of 98.7%. However, the patterns to be classified did not include dysplasia or malignancy. Dunham et al.[13] proposed the concept of "optical biopsy" (already introduced by the Brescia group before implementing AI applications in videomics)[14] using CNN technology. The first objective was to classify endoscopic images into one of five benign classes (normal mucosa, nodules, papilloma, polyps, and webs). The second was, using a binary classifier, to distinguish malignant/premalignant from benign lesions. The overall accuracy for the multiclass benign vocal fold lesion classifier was 80.8%, while the binary test achieved an overall accuracy of 93%.

Different authors[15,16] also demonstrated the feasibility of classifying oropharyngeal and oral cavity lesions using ML technology. For the oropharynx, the Stanford group[15] used a naïve Bayesian classifier (color and texture) to demonstrate the value of NBI imaging instead of white light (WL) videoendoscopy, which added more definition to tumor margins and highlighted submucosal vascularization. Five-fold cross-validation provided an AUC over 80% for NBI and under 55% for WL endoscopy models ($p<0.001$).

In the oral cavity, in 2018 Song et al.[16], employing CNNs, proposed a low-cost, smartphone-based, automatic image classification system. The authors collected data from 190 patients across several centers in India to detect oral dysplasia and malignancy using a dual-mode image analysis with WL and autofluorescence (AF). The study compared accuracy of the single- (WL or AF) and dual-mode (WL and AF) image analysis, demonstrating that the latter had a better diagnostic performance. The final model reached an accuracy of 87%, sensitivity of 85%, and specificity of 89%.

Detection

Lesions detection remains the main objective of DL-based strategies in contemporary clinical videoendoscopy. Different authors have described the potential of CNN in detection of cancer, premalignant lesions, benign lesions, and normal tissue. In this setting, algorithms are constantly being improved that better conform to specific tasks or subsites.

As mentioned in the previous section, Inaba et al.[17] trained a CNN-based algorithm (RetinaNet) to detect superficial laryngo-pharyngeal cancer. To evaluate diagnostic accuracy, 400 pathologic images and 800 of normal mucosa were collected, reaching an accuracy, sensitivity, and specificity of 97%, 95%, and 98%, respectively. The definition of correct diagnosis was set with an intersection over union (IoU) (i.e., the measure of overlap between prediction and ground truth) >0.4. Interestingly, the authors showed a direct correlation between the algorithm diagnostic performance and the number of images used for training. This is a not surprising outcome and clearly highlights the importance of training data, both in quantitative as well as in qualitative terms, during the training phase of an algorithm. In fact, to date, the low number and small size of the available medically-oriented datasets are the real bottleneck that limit the development of clinically relevant computer vision algorithms. A similar approach was described by Xiong et al.[18] who developed a CNN-based diagnostic system using videoendoscopic images of laryngeal cancer, premalignant lesions, benign lesions, and normal tissue. The results were comparable to those obtained by a human expert with 20 years of experience.

Concerning real-time detection, Matava et al.[19] and Azam et al. (in collaboration with our group)[20] developed CNN algorithms that were applied in real time during videoendoscopy and which aimed at identifying, on one side, normal airway anatomy and, on the other, UADT lesions. Using this type of approach, DL may be a useful complementary tool for clinicians in endoscopic examinations, progressively implementing the concept of human-computer collaboration. In detail, Matava et al.[19] compared the predictive performance of three models (ResNet, Inception, and MobileNet) in the identification of normal components of laryngeal and tracheal airway anatomy. ResNet and Inception achieved a specificity of 0.98 and 0.97, and a sensitivity of 0.89 and 0.86, respectively. Finally, Azam et al.[20] identified a CNN model for real-time laryngeal cancer detection in WL and NBI videoendoscopies. The dataset, consisting of 219 patients, was tested with an algorithm that achieved 0.66 precision (i.e., positive predictive value = true positives over true and false positives), 0.62 recall, and 0.63 mean average precision with an IoU>0.5. In addition, the model ran with an average computation time per videoframe of 0.026 seconds.

Segmentation

Automated segmentation of anatomical structures in medical image analysis is a prerequisite for autonomous diagnosis and represents one of the most complex tasks in the field of computer vision. In this case, the algorithm does not only need to detect lesions, but also to automatically delineate their margins. Recent CNN-based methods have demonstrated remarkable results and are well-suited for such a complex task.

During transoral laser microsurgery, a 7-class (void, vocal folds, other tissue, glottic space, pathology, surgical tools, and tracheal tube) dataset was trained by Laves et al.[21] using a CNN-based algorithm, as previously mentioned, this represents the first application of segmentation algorithms in this anatomical area. Different CNN architectures were investigated, and a weighted average ensemble network of UNet and ErfNet (two of the most commonly used CNNs) turned out to be the best suited for laryngeal segmentation, with a mean IoU of 84.7%. Advances in ML and computer vision have led to the development of methods for accurate and efficient real-time segmentation. As presented in section 4, our group performed a preliminary exploration on the use of fully CNNs for real-time segmentation of squamous cell cancer in videoendoscopies of the oral cavity (OC) and oropharynx (OP).[22] In this work, we compared different architectures and detailed their diagnostic performance and inference time, demonstrating their significant potential and the possibility to achieve real-time segmentation. However, for the first time, we suggested that highly heterogeneous subsites such as those encountered in the OC may have inferior results when compared with more structurally homogeneous areas such as the OP. This is in line with what we previously observed when applying bioendoscopic tools alone in a non-AI environment,[14] and is possibly related to the larger epithelial differentiation within the OC vs. the OP, and to specific limits related to oral examination (presence of light artifacts and confounders such as tongue blade, teeth, or dentures).

When dealing with laryngeal lesions, Fehling et al.[23] explored the possibility to achieve a fully automated segmentation of the glottic area and vocal fold tissue using a CNN in high-speed laryngeal videos. The algorithm obtained a Dice similarity coefficient (i.e., the measure that evaluates the intersection of the two regions as a ratio to the total area of them both) of 0.85 for the glottis, 0.91 for the right, and 0.90 for the left vocal fold. Furthermore, the results revealed that, in both pathologic and healthy subjects, the automatic segmentation accuracy obtained was comparable or even superior to manual segmentation.

Generally, laryngo-pharyngeal lesions are those more frequently examined when measuring the role of automatic analysis by ML. In fact, only limited studies on nasopharyngeal disease differentiation have been performed based on endoscopic images. For example, Li et al.[24] proposed a method to segment nasopharyngeal malignancies in endoscopic images based on DL. The final model reached an accuracy of 88.0%.

Finally, DL proved to be a promising addition to the field of endoscopic laryngeal high-speed videos. In clinical practice, the previous lack of dedicated software to analyze the data obtained resulted in a purely subjective assessment of the symmetry of vocal fold movement and oscillation. The development of easy-to-use DL-based systems that are capable of automatic glottal detection and midline segmentation allowed obtaining objective functional data without the need for manual or semi-automatic annotation as previously described, among others, by Piazza et al.[25] thus significantly simplifying the process. These results were obtained through an organized and stepwise approach headed by the Erlangen research group that achieves high-fidelity automatic segmentation of the glottis[23] and glottal midline[26] as well as extraction of relevant functional parameters.[27] Thanks to these preliminary data, a DL- enhanced software tool for laryngeal dynamics analysis was developed.[28] This software provides 79 unique quantitative analysis parameters for video- and audio-based signals, and most of these have already been shown to reflect voice disorders, highlighting its clinical importance.

In-depth characterization

All the previously described tasks aim to provide an accurate definition of a given lesion, classifying it according to its nature, defining its location in the frame, and delineating its margins (with possible future roles in real-time definition of resection margins during a surgical procedure). However, all these objectives only reproduce what is generally achieved by an expert clinician, and do not try to overcome the limits of human perception, even though their future implementation within a telemedicine environment would represent a large step towards more homogeneous diagnostic opportunities.

However, there is already indirect evidence that pattern recognition capabilities of novel AI systems may allow finding a correlation between the endoscopic appearance of a given lesion and its finer characteristics. Among these, depth of infiltration (DOI), so far investigable only by radiologic imaging or histopathologic evaluation,[29] plays a remarkable role in prognostication of OC cancer and has fueled great interest in the possibility of speeding up its definition by AI tools applied to videomics. Identification through videomics of other tumor characteristics, such as histopathological risk factors (e.g., perineural and lymphovascular invasion), viral status (Human Papilloma and Epstein-Barr viruses), and genomic markers is definitively more ambitious but already within the reach of similar approaches like radiomics and pathomics. Bridges connecting all these sources of information would be of great help in the near future to build up sharable profiling of tumors and their microenvironment.

Recent studies in the gastrointestinal tract, for example, have provided the proof of concept of this hypothesis and demonstrated that CNNs can differentiate between early and deeply infiltrating gastric cancer.[30] Nakahira et al.[31] further confirmed the potential of this approach by showing that CNN was able to correctly stratify the risk of gastric tumor development by analyzing the non-neoplastic mucosa at videoendoscopy.

Future perspectives

The introduction of computer vision in UADT endoscopy is still in its infancy and further steps will need to be taken before reaching widespread application. In this view, the first step outside of purely research-driven applications will be the use of ML algorithms for human-computer collaboration. Dedicated algorithms can assist in every step of the endoscopic diagnostic approach, from quality assurance, effective storage and video classification to risk determination, histologic definition, margins evaluation, and in-depth lesion profiling. As previously stated, this will be a stepwise approach that will start from easier tasks (i.e., quality assurance) and will progress towards more complex and more clinically relevant objectives. The ideal outcome will be to achieve accurate lesion characterization in terms of histologic nature, margins, and biologic characteristics, and to be able to integrate these insights fully and objectively with data from other types of examinations (e.g., radiology, molecular biology, and histopathology).

Morphologic image analysis is the main field in which videomics is evolving in the context of clinical endoscopy. However, other more innovative aspects can be assessed by taking advantage of current computer-vision technologies. A particularly interesting feature in otolaryngology is vocal fold motility; in fact, objective evaluation of this variable can be extremely helpful in both assessment of functional deficits and in the precise staging of neoplastic disease of the glottis. This is especially true when considering that the AJCC/UICC TNM classification[32] of laryngeal cancer relies on purely subjective definitions of "normal vocal cord mobility", "impaired vocal cord mobility", and "vocal cord fixation" for categorization of T1, T2, and T3 glottic tumors, respectively.

In this field, Adamian et al.[33] recently developed an open-source computer vision tool for automated vocal fold tracking from videoendoscopies that is capable of estimating the anterior angle between vocal folds of subjects with normal mobility and those with unilateral vocal fold paralysis. The authors demonstrated the possibility to identify patients with vocal fold palsy by assessing the angle of maximal glottic opening (49° vs. 69°; $p<0.001$). In particular, an angle of maximum opening <58.6° was predictive of paralysis with a sensitivity and specificity of 0.85. Notwithstanding, this approach has significant limits in evaluation of reduced mobility due to neoplastic involvement since it relies on identification of the free margin of vocal folds, which is often altered by glottic tumors. However, the development of alternative strategies is providing valuable outcomes in such a task, as described in the following sections.

Finally, novel surgical technologies such as transoral robotic[34] and exoscopic surgery[35] rely on digital video acquisition of a large amount of data and will potentially extend the applications of videomics to the intraoperative setting of quality and safety control as well as didactic proficiency. This is especially interesting considering the urgent need for more extensive training and collaborative datasets that will enable better refinement of ML algorithms, coming not only from diagnostic instrumentation, but also from surgical robots and exoscopic tools.

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. gennaio 2019;25(1):44–56.

2. Paderno A, Holsinger FC, Piazza C. Videomics: bringing deep learning to diagnostic endoscopy. Curr Opin Otolaryngol Head Neck Surg. aprile 2021;29(2):143–8.

3. Piazza C, D Bon F, Peretti G, Nicolai P. «Biologic endoscopy»: optimization of upper aerodigestive tract cancer evaluation. Curr Opin Otolaryngol Head Neck Surg. aprile 2011;19(2):67–76.

4. Gómez P, Kist AM, Schlegel P, Berry DA, Chhetri DK, Dürr S, et al. BAGLS, a multihospital benchmark for automatic glottis segmentation. Sci Data 2020; 7(1):1-12.

5. Raza K, Singh NK. A Tour of Unsupervised Deep Learning for Medical Image Analysis. Curr Med Imaging Former Curr Med Imaging Rev 2021;17(9):1059–77.

6. Azizi S, Mustafa B, Ryan F, Beaver Z, Freyberg J, Deaton J, et al. Big self-supervised models advance medical image classification. Proceedings of the IEEE/CVF International Conference on Computer Vision 2021;3478-3488.

7. 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):1225–38.

8. Esteva A, Chou K, Yeung S, Naik N, Madani A, Mottaghi A, et al. Deep learning-enabled medical computer vision. NPJ Digit Med 2021;4(1):5.

9. He Y, Cheng Y, Huang Z, Xu W, Hu R, Cheng L, et al. A deep convolutional neural network-based method for laryngeal squamous cell carcinoma diagnosis. Ann Transl Med 2021;9(24):1797.

10. Esmaeili N, Sharaf E, Gomes Ataide EJ, Illanes A, Boese A, Davaris N, et al. Deep Convolution Neural Network for Laryngeal Cancer Classification on Contact Endoscopy-Narrow Band Imaging. Sensors 2021;21(23):8157.

11. Zhao Q, He Y, Wu Y, Huang D, Wang Y, Sun C, et al. Vocal cord lesions classification based on deep convolutional neural network and transfer learning. Med Phys 2022;49(1):432–42.

12. Huang CC, Leu YS, Kuo CFJ, Chu WL, Chu YH, Wu HC. Automatic recognizing of vocal fold disorders from glottis images. Proc Inst Mech Eng 2014;228(9):952–61.

13. Dunham ME, Kong KA, McWhorter AJ, Adkins LK. Optical Biopsy: Automated Classification of Airway Endoscopic Findings Using a Convolutional Neural Network. The Laryngoscope 2022;132(S4):S1–8.

14. Piazza C, Del Bon F, Paderno A, Grazioli P, Perotti P, Barbieri D, et al. The diagnostic value of narrow band imaging in different oral and oropharyngeal subsites. Eur Arch Otorhinolaryngol 2016;273(10):3347–53.

15. Mascharak S, Baird BJ, Holsinger FC. Detecting oropharyngeal carcinoma using multispectral, narrow-band imaging and machine learning: Multispectral Imaging of Oropharynx Cancer. The Laryngoscope 2018;128(11):2514–20.

16. Song B, Sunny S, Uthoff RD, Patrick S, Suresh A, Kolur T, et al. Automatic classification of dual-modalilty, smartphone-based oral dysplasia and malignancy images using deep learning. Biomed Opt Express 2018;9(11):5318.

17. Inaba A, Hori K, Yoda Y, Ikematsu H, Takano H, Matsuzaki H, et al. Artificial intelligence system for detecting superficial laryngopharyngeal cancer with high efficiency of deep learning. Head Neck 2020;42(9):2581–92.

18. Xiong H, Lin P, Yu JG, Ye J, Xiao L, Tao Y, et al. Computer-aided diagnosis of laryngeal cancer via deep learning based on laryngoscopic images. EBioMedicine 2019;48:92–9.

19. 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):44.

20. Azam MA, Sampieri C, Ioppi A, Africano S, Vallin A, Mocellin D, et al. Deep Learning Applied to White Light and Narrow Band Imaging Videolaryngoscopy: Toward Real-Time Laryngeal Cancer Detection. The Laryngoscope 2021 (epub ahead of print)

21. 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):483–92.

22. Paderno A, Piazza C, Del Bon F, Lancini D, Tanagli S, Deganello A, 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;11:626602.

23. Fehling MK, Grosch F, Schuster ME, Schick B, Lohscheller J. Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep Convolutional LSTM Network. PloS One 2020;15(2):e0227791.

24. Li C, Jing B, Ke L, Li B, Xia W, He C, et al. Development and validation of an endoscopic images-based deep learning model for detection with nasopharyngeal malignancies. Cancer Commun Lond Engl 2018;38(1):59.

25. Piazza C, Mangili S, Del Bon F, Gritti F, Manfredi C, Nicolai P, et al. Quantitative analysis of videokymography in normal and pathological vocal folds: a preliminary study. Eur Arch Otorhinolaryngol 2012;269(1):207–12.

26. Kist AM, Zilker J, Gómez P, Schützenberger A, Döllinger M. Rethinking glottal midline detection. Sci Rep 2020;10(1):20723.

27. Schlegel P, Kniesburges S, Dürr S, Schützenberger A, Döllinger M. (2020). Machine learning based identification of relevant parameters for functional voice disorders derived from endoscopic high-speed recordings. Sci Rep 2020;10(1), 1-14.

28. Kist AM, Gómez P, Dubrovskiy D, Schlegel P, Kunduk M, Echternach M, et al. A Deep Learning Enhanced Novel Software Tool for Laryngeal Dynamics Analysis. J Speech Lang Hear Res 2021;64(6):1889-1903.

29. Piazza C, Montalto N, Paderno A, Taglietti V, Nicolai P. Is it time to incorporate «depth of infiltration» in the T staging of oral tongue and floor of mouth cancer? Curr Opin Otolaryngol Head Neck Surg 2014;22(2):81–9.

30. Yoon HJ, Kim S, Kim JH, Keum JS, Oh SI, Jo J, et al. A Lesion-Based Convolutional Neural Network Improves Endoscopic Detection and Depth Prediction of Early Gastric Cancer. J Clin Med 2019;8(9):E1310.

31. Nakahira H, Ishihara R, Aoyama K, Kono M, Fukuda H, Shimamoto Y, et al. Stratification of gastric cancer risk using a deep neural network. JGH Open Open Access J Gastroenterol Hepatol 2020;4(3):466–71.

32. Amin MB, Edge S, Greene F, Byrd DR, Brookland RK, Washington MK, Gershenwald JE, Compton CC, Hess KR, et al. (Eds.). AJCC Cancer Staging Manual (8th edition). Springer International Publishing: American Joint Commission on Cancer; 2017.

33. Adamian N, Naunheim MR, Jowett N. An Open-Source Computer Vision Tool for Automated Vocal Fold Tracking From Videoendoscopy. The Laryngoscope 2021;131(1):E219-25.

34. Curry M, Malpani A, Li R, Tantillo T, Jog A, Blanco R, et al. Objective assessment in residency-based training for transoral robotic surgery. The Laryngoscope 2012;122(10):2184–92.

35. Paderno A, Deganello A, Lancini D, Piazza C. Is the exoscope ready to replace the operative microscope in transoral surgery? Curr Opin Otolaryngol Head Neck Surg 2022;30(2):79–86.

3. Deep learning in endoscopy: The importance of standardization

Defining standards: Guidelines for adequate reporting in medical computer vision studies

Introduction

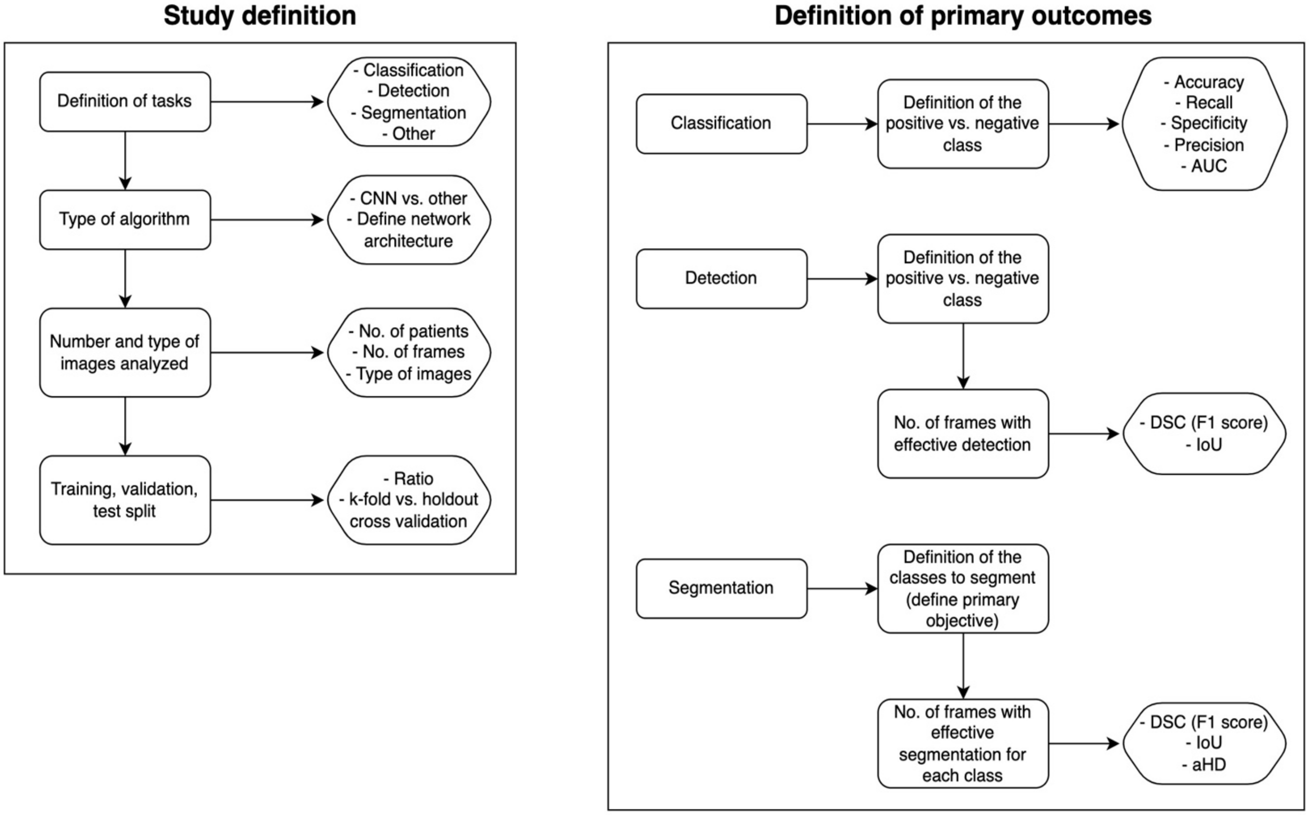
As previously described, different DL architectures have been proposed to address a variety of tasks, including image classification, object detection, segmentation and characterisation.[1] These advancements have led to significant progress in assisting clinicians in the evaluation of endoscopic frames.[2] However, a major challenge is that the performance of these methods is often not directly comparable among different institutions/series because of the lack of standardised evaluation methodology.

The importance of standardising outcomes in DL for medical imaging cannot be overemphasised and its lack is a major barrier to translation into actual clinical practice of such algorithms for videomics. To enable fair algorithm comparisons, it is essential to have common evaluation metrics that are agreed upon by the research community from both clinical and technical perspectives.[3]

The dataset used for algorithm training, validation and testing should be representative of the real-world clinical scenario, possibly including data from different centres and annotated by different clinicians.

Herein, we propose a general guideline to standardise reporting in studies focused on the automatic analysis of endoscopic images (i.e., videomics). Figure 3.1 shows a schematic framework aimed at addressing study definition and outcome metrics.

Figure 3.1. *Schematic framework detailing a proposal to standardize study definitions and outcome measures in studies assessing deep learning in videoendoscopy. Legend: aHD, average Hausdorff distance; AUC, area under the curve; CNN, convolutional neural network; DSC, Dice Similarity Coefficient; IoU, Intersection-over-Union.*


Study definition
Definition of tasks
- Classification
- Detection
- Segmentation
- Other
Type of algorithm
- CNN vs. other
- Define network architecture
Number and type of images analyzed
- No. of patients
- No. of frames
- Type of images
Training, validation, test split
- Ratio
- k-fold vs. holdout cross validation
Definition of primary outcomes
Classification
Definition of the positive vs. negative class
- Accuracy
- Recall
- Specificity
- Precision
- AUC
Detection
Definition of the positive vs. negative class
No. of frames with effective detection
- DSC (F1 score)
- IoU
Segmentation
Definition of the classes to segment (define primary objective)
No. of frames with effective segmentation for each class
- DSC (F1 score)
- IoU
- aHD

Study definition

The first step is to clearly describe the setting and methodology:

1. Objective: clearly state the type of task to be assessed.
2. Algorithm: describe the DL algorithm, its architecture, and technical features (e.g., loss function, optimizer, learning rate, batch size, validation metric, strategy to stop training, training curves).

3. Describe the number of patients and number of frames extracted from each (with selection criteria). Technical information regarding endoscopic and recording equipment, as well as the use of optical filters (e.g., narrow band imaging), should also be available.
4. The training-validation-test set split ratio should be described, specifying the type of cross-validation. Patients should be clustered in different sets.

Definition of primary outcomes

A further step is to clearly define the study outcomes by selecting measures that are actually representative of diagnostic performance. Figure 3.1 shows the suggested metrics for each task, selecting those with a lower risk of resulting in an illegitimate high score. Adjunctive metrics can be added, but should not be considered as primary outcomes unless there is a clear rationale.

Regarding classification, thanks to the clear dichotomic distinction between positive and negative results, the same rationale used for standard diagnostic tests should be applied. A wide range of metrics (see Figure 3.1) should be provided to allow a comprehensive assessment of the algorithm's diagnostic characteristics. In addition, in case of class unbalance, appropriate metrics should be considered (e.g., F1-score instead of accuracy, precision-recall curve).

Conversely, special considerations should be applied to detection and segmentation tasks. It is essential to define the percentage of frames in which a region of interest (ROI) has been detected. Here, the most commonly used parameter is accuracy (Acc), which measures the percentage of correctly classified pixels. However, endoscopic frames are commonly highly class-imbalanced. Frames usually contain a single region of interest involving only a small portion of pixels, whereas the remaining image is labelled as background. Because of the positive impact of true negatives (i.e., background), Acc will always result in high scores. Similarly, specificity (Spec) indicates the model capability to detect the background in an image. Due to the large fraction of pixels annotated as background compared to the ROI, specificity values close to 1 are typical. For this reason, Acc and Spec should not be used as primary outcomes.

Sensitivity (also defined as Recall) is another popular metric, but in detection and segmentation tasks it is less representative than overlap metrics such as the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU). In particular: the DSC is calculated by taking the number of data points that are shared by both sets and dividing it by the total number of data points; IoU is calculated by taking the intersection of the predicted segmentation and the ground truth segmentation, and dividing it by the union of the two.

Finally, the Hausdorff Distance (HD) measures the distance between two sets of points (i.e., ground truth and predicted segmentation), and allows scoring localisation similarity by focusing on the delineation of margins. Since the HD is sensitive to outliers, the average HD may be better suited for most applications.

This is a first proposal to standardise reporting in videomics studies. Some technical concepts should be taken into consideration to improve collaboration and reliably assess the performance of novel DL algorithms before introducing them into a clinical setting.

References

1. Paderno A, Gennarini F, Sordi A, et al. Artificial intelligence in clinical endoscopy: Insights in the field of videomics. Front Surg 2022;9:933297.
2. Paderno A, Holsinger FC, Piazza C. Videomics: bringing deep learning to diagnostic endoscopy. Curr Opin Otolaryngol Head Neck Surg 2021;29:143-148.

3. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 2015;15:29.

4. Deep learning for automatic segmentation of oral and oropharyngeal cancer using Narrow Band Imaging: Preliminary experience in a clinical perspective

Proof of concept: Setting the acquisition pipeline and testing technical approaches.

Introduction

Surgical data science (SDS)[1] is an emerging field of medicine aimed at extracting knowledge from medical data and providing objective measures to assist in diagnosis, clinical decision making, and prediction of treatment outcomes. In this context, image segmentation, an essential step in computer vision, can be defined as the task of partitioning an image into several non-intersecting coherent parts.[2] It is also well known that segmentation is a prerequisite for autonomous diagnosis, as well as for various computer- and robot-aided interventions. Many methodologies have been proposed for image segmentation,[3] but recent and successful approaches are based on fully convolutional neural networks (FCNNs), applying convolutional filters that learn hierarchical features from data (i.e., input images), and then collecting them in maps. In general, a high number of filters will give better results up to a certain point, when a further increase in their number either does not improve the segmentation performance, or deteriorates it.[4]

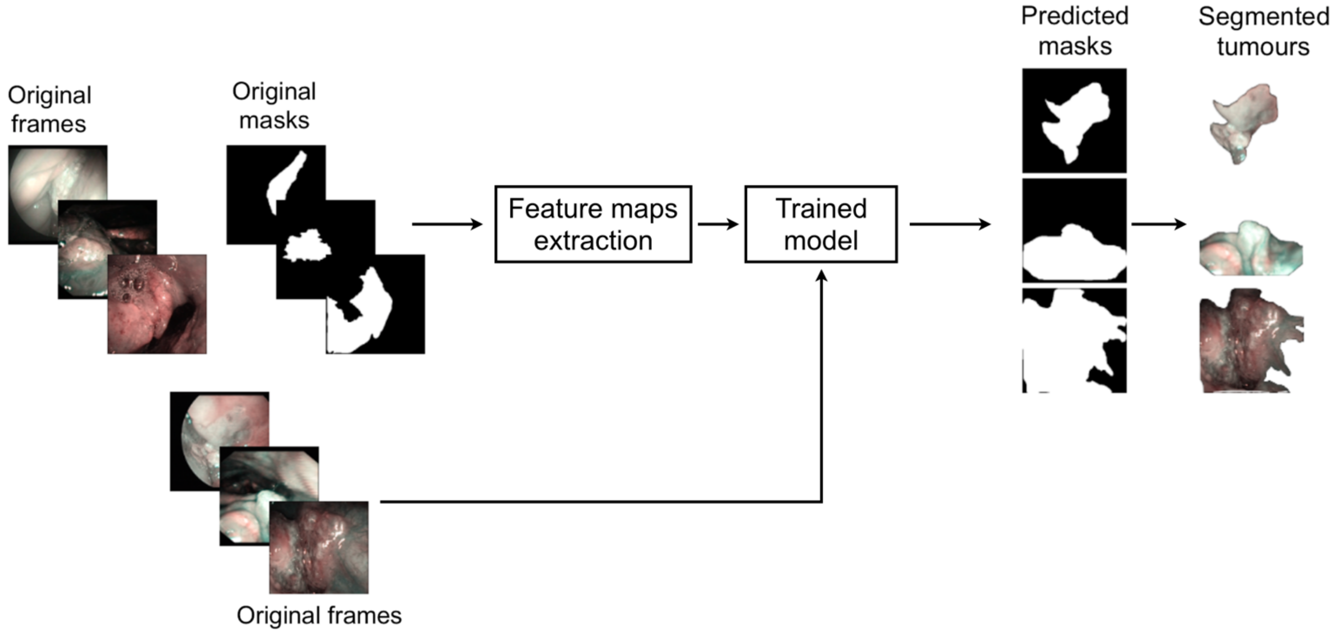
FCNNs applied to video-analysis are of particular interest in the field of head and neck oncology, since endoscopic examination (and its storage in different ways and media) has always represented a crucial step in diagnosis, staging, and follow-up of patients affected by upper aero-digestive tract cancers. In this view, Narrow Band Imaging (NBI) represents an already consolidated improvement over conventional white light endoscopy, allowing for better and earlier identification of dysplastic/neoplastic mucosal alterations.[5-8] However, so far, NBI-based endoscopy remains a highly operator-dependent procedure, and its standardization remains particularly challenging, even when employing simplified pattern classification schemes.[9-14] In fact, even though such bioendoscopic tools are aimed at identifying pathognomonic superficial vascular changes in forms of abnormal intrapapillary capillary loops, a relatively long learning curve and intrinsic subjectivity of subtle visual evaluations still hamper their general and widespread adoption in daily clinical practice. Furthermore, subtle differences in NBI patterns according to the head and neck subsite to be analyzed have also been described. This is especially true considering the oral cavity (OC) in comparison with other upper aerodigestive tract sites.[9]

The aim of this study was to test FCNN-based methods for semantic segmentation of early squamous cell carcinoma (SCC) in video-endoscopic images belonging to OC and oropharyngeal (OP) subsites to pave the way towards development of intelligent systems for automatic NBI video-endoscopic evaluations.

Materials and Methods

This study was performed following the principles of the Declaration of Helsinki and approved by the Institutional Review Board, Ethics Committee of our academic hospital (Spedali Civili of Brescia, University of Brescia, Brescia, Italy). The workflow of the approach used is shown in Figure 4.1. In particular, informative NBI frames were selected from videos of OC and OP SCC through a case by case evaluation (“Original frames”, Figure 4.1). Each frame was manually annotated by an expert clinician contouring the lesion margins, thus creating a mask referring to every frame ("Original mask", Figure 4.1). The original frames and original masks were employed to train the FCNNs in order to obtain an automatic tumor segmentation.

Figure 4.1. *Workflow of the approach used for detection of mucosal SCC in videoendoscopic frames by NBI.*



Mucosal cancer segmentation

Two datasets were retrieved from the institutional registry analyzing 34 and 45 NBI endoscopic videos of OC and OP, respectively. Each video was from a different patient affected by SCC, clinically presenting as a leuko- or erythroplastic lesion. Image acquisition was performed at the Department of Otorhinolaryngology – Head and Neck Surgery, ASST Spedali Civili, University of Brescia, Brescia, Italy between January 2010 and December 2018. Only video-endoscopies of biopsy-proven OC and OP SCC were included in the study. Patients with previous surgical and/or non-surgical treatments for tumors of these anatomical sites and frankly ulcerated neoplasms with significant loss of substance were excluded from the analysis.

All videos were acquired under white light and NBI by a rigid telescope coupled to an Evis Exera II HDTV camera connected to an Evis Exera II CLV-180B light source (Olympus Medical Systems Corporation, Tokyo, Japan). From the total amount of frames constituting the NBI videos, non-informative frames (i.e., blurred, out of focus, dark, or with signs of bleeding) were discarded through a case by case evaluation. After this selection process, the dataset referring to the OC was composed of 110 frames, while a total of 116 frames composed the OP dataset. Table **4.**1 shows the number of frames tested per patient for each dataset and the relative total amount of frames and patients involved. Each frame in these databases was manually annotated by an expert clinician contouring the lesion margins. The correspondent mean lesion size in percentage of pixels with respect to the entire frame size for each dataset is reported in Table **4.**2.

Table 4.1. *Investigated datasets for mucosal SCC segmentation task and corresponding number of NBI videoframes per patient.*

Oral cavity			
	No. patients	**No. frames per patient**	**No. frames**
	6	1	6
	26	2	52
	8	4	32
	5	4	20
Total	45		110
Oropharynx			
	No. patients	**No. frames per patient**	**No. frames**
	10	2	20
	8	3	24
	12	4	48
	4	6	24
Total	34		116

Table 4.2. *Investigated datasets for mucosal SCC segmentation task and corresponding amount of mean percentages of lesion pixels per frame and relative standard deviations.*

Dataset	**Mean of lesion pixels in %**	**Standard deviation of lesion pixels in %**
Oral cavity	22.84	11.68
Oropharynx	38.04	18.54

Before segmenting the tumor area with FCNNs, the images underwent a cropping procedure to remove black borders. Given the different dimensions and shapes of extracted NBI video-frames, the cropping was customized for each of them. For memory constraints, frames were down-sampled to dimensions of 256×256 pixels to prevent exceeding the available GPU memory (∼14858 MB). Prior to FCNN-based segmentation, images were standardized sample-wise, namely the image mean was removed from each image. Given the small size of the two datasets, data augmentation was performed to avoid overfitting and to increase the ability of the model to better generalize the results. Hence, the training set was augmented by ∼10 times at each cross-validation, imposing the following random transformations to the frames (and corresponding gold-standard masks obtained with manual segmentation): image rotation (random rotation degree in range 0°-90°), shift (random shift in range 0-10% of the frame side length for both width and height), zoom (with zoom values in range 0 and 1), and horizontal and vertical flip.

Three FCNNs were investigated to segment neoplastic images in OC and OP. The architectures tested were:

U-Net, a fully convolutional U-shaped network architecture for biomedical image segmentation;[15]

U-Net 3, consisting of the previous deep network improved by Liciotti et al.[16] to work with very few training images and yield more precise segmentations;

ResNet, composed of a sequence of residual units.[17]

Technical definitions

FCNNs are a type of artificial neural networks that have wide application in visual computing. Their deep hierarchical model roughly mimics the nature of mammalian visual cortex, making FCNNs the most promising architectures for image analysis. FCNNs present an input layer, an output layer, and a variable number of hidden layers, that transform the input image thorough the convolution with small filters, whose weights and biases are learned during a training procedure.

U-Net is a fully convolutional U-shaped FCNN that is especially suitable for biomedical images. The descending path U-Net is made of repeated 3x3 convolutions and max-pooling, for down-sampling the input image. This path acts as an encoder for feature extraction. The ascending path consists of 3x3 convolutions and up-sampling, for restoring the original input image size. This path acts as decoder for feature processing to achieve the segmentation. The encoder and decoder are linked to each other via long skip connections. U-Net3 is inspired by U-Net but introduces batch normalization, which makes the training process faster. ResNet is also divided in two parts: the descending and ascending paths, each consisting of 5 blocks. Each block of the descending path is made of a convolutional sub-block and two identity sub-blocks, whereas in the ascending path there is one up-convolutional sub-block and two identity sub-blocks. The convolutional and identity sub-blocks follow the implementation of He et al.[17] and are made of convolution filters. In order to study the complexity of ResNet, in this work we also tested the performance of ResNet considering 1 block, 3 blocks, and 4 blocks per path. In each block, we kept the number of filters for each convolution equal to 16. We also investigated the ResNet with 1 block per path and 8 filters per convolution instead of 16: this represents the simplest model.

Data analysis

FCNN performance was evaluated for each network tested and compared to the gold standard represented by manual annotation performed by expert clinicians. A contingency table considering true positive (TP), true negative (TN), false negative (FN), and false positive (FP) results was used. The overall accuracy (Acc) was calculated and defined as the ratio of the correctly segmented area by the algorithm over the annotated area by the expert examiner. The positive and negative samples refer to pixels within and outside the segmented region, respectively. Precision (Prec) was defined as the fraction of relevant instances among the retrieved ones (i.e., positive prediction value; true positives over true and false positives). Recall (Rec) was defined as the fraction of the total amount of relevant instances that were actually retrieved (i.e., true positive rate; true positives over true positives and false negatives). The Dice Similarity Coefficient (Dsc) was evaluated as overlapping measure. The Dsc is a statistical validation metric based on the spatial overlap between two sets of segmentations of the same anatomy. The value of Dsc ranges from 0, indicating no spatial overlap between two sets of binary segmentation results, to 1, indicating complete overlap. Tumor detection performance was evaluated by measuring the computational time required by each of the FCNN architectures investigated to perform automatic segmentation per frame.

Analysis of variance (Anova test) with a significance level of 0.05 was performed to check whether the averages of the computed metrics significantly differed from each other. When significant differences were found, a pairwise T-test for multiple comparisons of independent groups was performed.

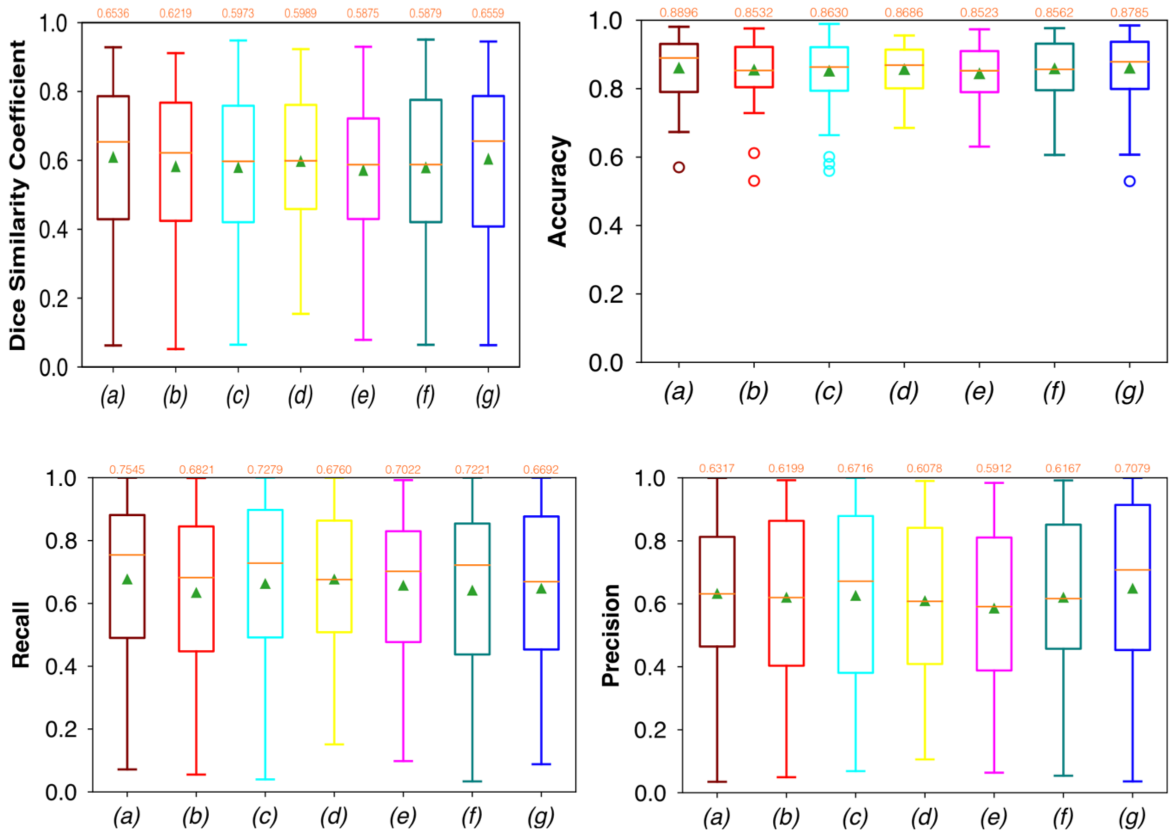
Results

Oral cavity dataset

For FCNN-based segmentation of the OC dataset, the best results in terms of Dsc were achieved by ResNet with 5(×2) blocks (5 for the descending and 5 for the ascending path) and 16 filters, with a median value of 0.6559, as reported in Figure 4.2. The comparison in terms of Acc in Figure 4.3 for the tested FCNN architectures showed the best results in terms of median for the U-Net, with a value of 0.8896. Both the abovementioned architectures showed the best results in terms of other metrics. Specifically, ResNet with 5(×2) blocks and 16 filters appeared to be the best in terms of Rec, with a median value of 0.7545, as reported in Figure 4.4. U-Net, in contrast, showed the best result in terms of Prec, with a median value of 0.7079, as reported in Figure 4.5. All FCNNs tested presented very high values of variance, leading to very low values of minima for all metrics evaluated, except for Acc which showed better results than in the OP dataset.

No significant difference was found when analyzing variance with the Anova test (p>0.05) to the Dsc, Acc, Rec, and Prec vectors constituted by the metrics of each architecture.

***Figure 4.2-4.5**. Boxplots of performance metrics for the OC dataset, obtained for (a) U-Net architecture, (b) U-Net 3, (c) ResNet with 4(×2) blocks and 16 filters, (d) ResNet with 1(×2) blocks and 8 filters, (e) ResNet with 1(×2) blocks and 16 filters, (f) ResNet with 3(×2) blocks and 16 filters, and (g) ResNet with 5(×2) blocks and 16 filters. Green triangles indicate the mean values, while the orange numbers at the top of each boxplot are the corresponding median values.*



The computational times required by the FCNNs for the automated segmentation task for one image are reported in Table 4.3. It is worth noticing that less deep networks, such as ResNet with 1(×2) blocks and 8 filters, and ResNet with 1(×2) blocks and 16 filters, achieved automated segmentation in shorter times than the others. In particular, ResNet with 1(×2) blocks and 8 filters took only 14 ms to predict a single frame.

Table 4.3: *Tested FCNNs and the corresponding times of inference for each single frame, expressed in milliseconds (ms).*

FCNNs	**Inference time per frame (ms)**
U-Net	~115
U-Net 3	~96
ResNet with 4x2 blocks, 16 filters	~66

ResNet with 1x2 blocks, 8 filters	~14
ResNet with 1x2 blocks, 16 filters	~23
ResNet with 3x2 blocks, 16 filters	~59
ResNet with 5x2 blocks, 16 filters	~59

Oropharyngeal dataset

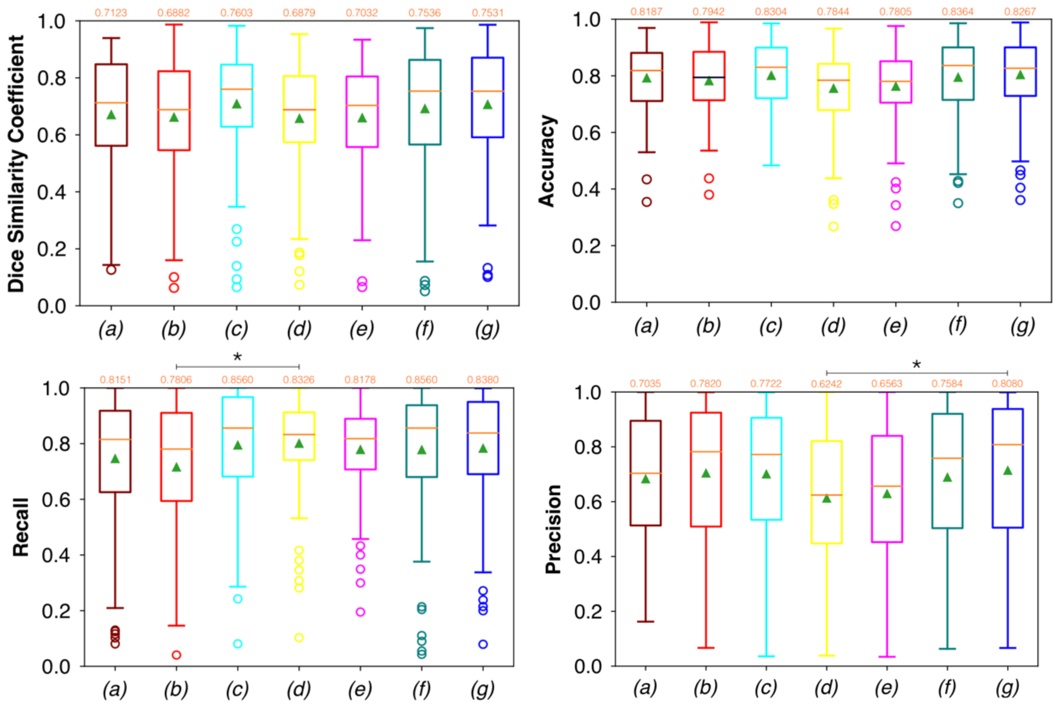
Considering FCNN-based segmentation for the OP dataset, the best results in terms of Dsc were achieved by ResNet with 4(×2) blocks and 16 filters, with a median value of 0.7603, as reported in Figure 4.6. The comparison in terms of Acc in Figure 4.7 for the FCNN architectures showed the best results in terms of median for the ResNet with 3(×2) blocks and 16 filters, with a median value of 0.8364. Both the abovementioned architectures also showed the best results in terms of Rec, with a median value of 0.8560 for both, as reported in Figure 4.8.

Conversely, considering the comparison in terms of Prec in Figure 4.9, the best result was achieved with a deeper network, the ResNet with 5(×2) blocks and 16 filters. However, no significant difference was found when analyzing variance with the Anova test ($p>0.05$) to the Dsc vectors constituted by the Dsc of each architecture.

A significant difference among the different FCNNs was found applying the same test to the Rec vectors (Figure 4.8). A further investigation was performed using a pairwise T-test for multiple comparisons of independent groups that demonstrated a p value of 0.043 between U-Net 3 and ResNet with 1(×2) blocks and 8 filters, demonstrating that architectures with skip connections (i.e., all the architectures tested except for U-Net and U-Net 3) had greater performances in detecting mucosal sites affected by SCC.

Moreover, a significant difference among the different FCNNs was also found by applying the Anova test to the Prec vectors (Figure 4.9). A further investigation using a pairwise T-test for multiple comparisons of independent groups showed a p value of 0.0454 between ResNet with 5(×2) blocks and 16 filters, and ResNet with 1(×2) blocks and 8 filters, demonstrating that deeper architectures were more precise in detecting SCC.

Figure 4.6-4.9. *Boxplots of performance metrics for the OP dataset, obtained for (a) U-Net architecture, (b) U-Net 3, (c) ResNet with 4(×2) blocks and 16 filters, (d) ResNet with 1(×2) blocks and 8 filters, (e) ResNet with 1(×2) blocks and 16 filters, (f) ResNet with 3(×2) blocks and 16 filters, and (g) ResNet with 5(×2) blocks and 16 filters. Green triangles indicate the mean values, while the orange numbers at the top of each boxplot are the corresponding median values.*



Samples of original frames, manual masks, and relative predicted masks for the OC and OP are shown in Figure 4.10-4.12 in order to provide a visual input on the characteristics of correctly and incorrectly segmented tumors, and non-diagnostic cases.

Figure 4.10: *Sample of original OC frames, manual masks, and relative predicted masks for ResNet with 5 (x2) blocks and 16 filters. The red and green boxes correspond to values of Dsc less than 45% (Dsc <0.45) and Dsc greater than 85% (Dsc >0.85), respectively.*

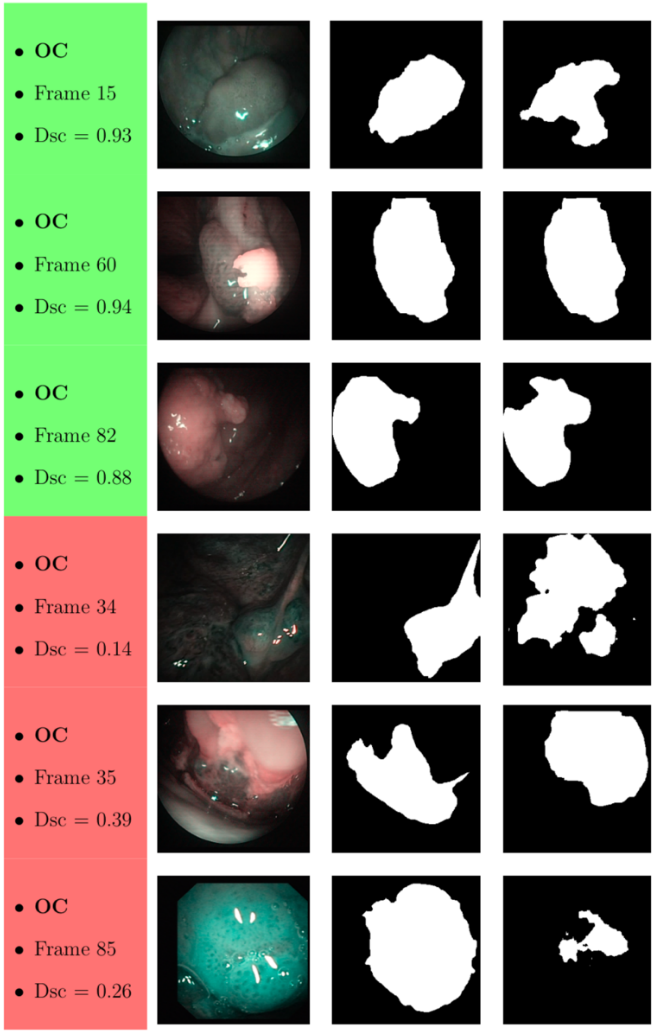


Figure 4.11: *Sample of original OP frames, manual masks, and relative predicted masks for ResNet with 4 (x2) blocks and 16 filters. The red and green boxes correspond to values of Dsc less than 45% (Dsc <0.45) and Dsc greater than 85% (Dsc >0.85), respectively.*

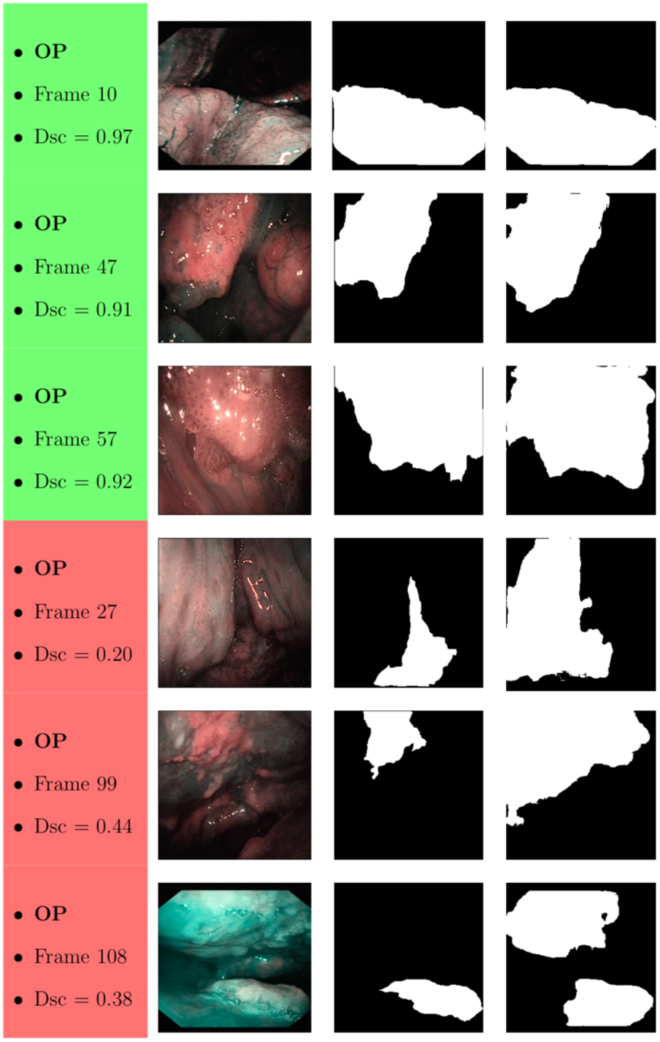
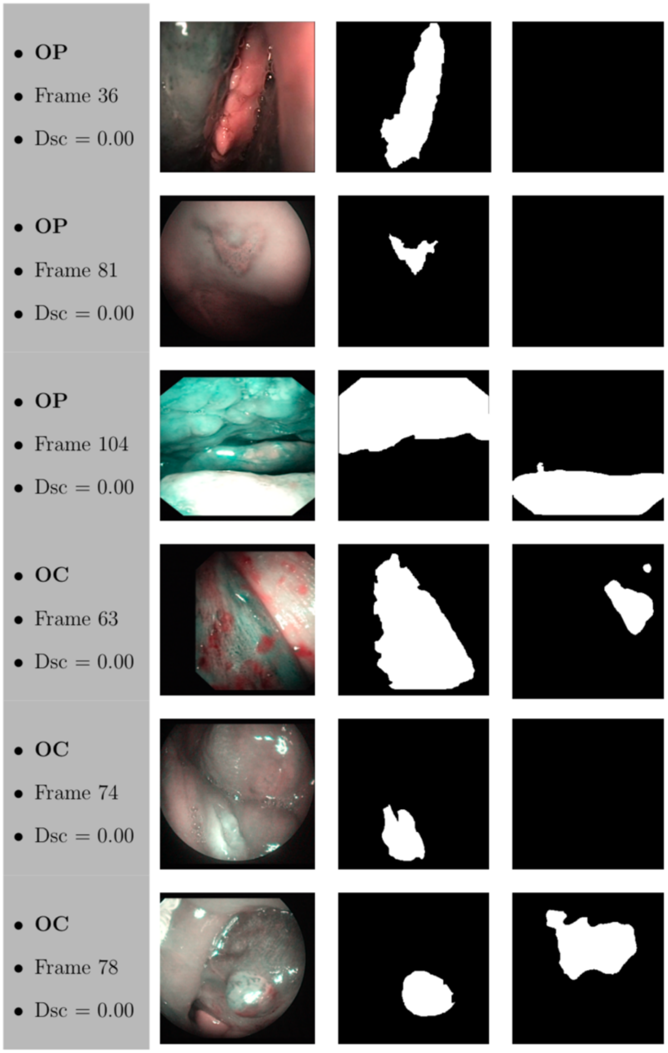


Figure 4.12: *Sample of original frames excluded from the boxplot comparisons due to their Dsc less than 5 % (Dsc <0.05) assessed by ResNet with 4 (x2) blocks and 16 filters for OP frames, and ResNet with 5 (x2) blocks and 16 filters for OC frames. The manual masks and relative predicted masks are also reported.*



Discussion

This study presents a computer-aided method for segmentation of SCC through FCNN-based evaluation of NBI video-endoscopic frames afferent to two frequently involved upper aero-digestive tract sites (OC and OP), and evaluates its performance in distinguishing between neoplastic and healthy areas. The overall median Dsc for OC and OP frames of the best performing FCNN [ResNet with 1(×2) blocks and 8 filters] with the shortest time of inference (14 ms) were 0.5989 and 0.6879, respectively.

Of note, this was the first attempt to automatically segment SCC in complex anatomical regions from NBI video-frames. Considering the absence of deep-learning methods in the head and neck literature from which to draw inspiration, this early experience can be considered as a practical approach for segmentation of pathological areas in endoscopic videos, applicable in real time during routine clinical activities, given the short time of inference needed per frame.

Moreover, this approach demonstrates the value of SDS in OC/OP examination and could motivate more structured and regular data storage in the clinic. Indeed, large amounts of data would lead to the possibility of further exploring deep-learning-based algorithms for semantic segmentation, covering a more substantial variability of tissues classification scenarios. In addition, associating such diagnostic videos to subsequently obtained radiologic imaging, pathological specimens, and prognostic characteristics, could pave the way to data mining aimed at understanding adjunctive tumor features (e.g., HPV status, depth of infiltration, risk of regional/distant metastasis) by simple video-endoscopy.

In this field, few methodologies have been presented for automatic diagnosis of tumors of the upper aero-digestive tract. As in the present series, most were focused on optimizing the analysis by providing adjunctive features (e.g., NBI, autofluorescence) complementing those obtained by conventional white light endoscopy. Taking advantage of the value of autofluorescence in the OC, Song et al.[18] developed an automatic image classification using a smartphone-based system for OC lesions employing CNNs that evaluated dual-modality images (white light and autofluorescence). The final model reached an accuracy of 87%, sensitivity of 85%, and specificity of 89%.

Conversely, different approaches aimed at maximizing extraction of features focusing on tissue vascularization. Specifically, Barbalata et al.[19] proposed a method for automated laryngeal tumor detection based on post-processing of images. Laryngeal tumors were detected and subsequently classified, focusing on their abnormal intrapapillary capillary loops through anisotropic filtering and matched filter. This further reinforces the rationale of using NBI data for our analysis, since this light-filtering system better highlights blood vessels, thus increasing the quality and quantity of data to be analyzed in each image. This concept was also confirmed by Mascharak et al.[20] who took advantage of naıve Bayesian classifiers trained with low-level image features to automatically detect and quantitatively analyze OP SCC using NBI multispectral imaging. The authors showed a significant increase in diagnostic accuracy using NBI compared to conventional white light video-endoscopy.

Recent studies confirmed the potential of FCNNs in the automatic diagnosis of benign and malignant diseases of the upper aero-digestive tract,[21-25] demonstrating an outstanding Acc, comparable with that of experienced physicians. However, these studies were only focused on tumor detection and did not include OC and anterior OP tumors since the examination was only based on transnasal/transoral flexible video-endoscopy. Furthermore, no attempt at segmentation of the precise tumor margins was made.

Considering segmentation and margin recognition tasks, a study by Laves et al.[3] put effort into using FCNN to segment a dataset of the human larynx. The dataset, consisting of 536 manually segmented endoscopic images obtained during transoral laser microsurgery, was tested in order to monitor the morphological changes and autonomously detect pathologies. The intersection-over-Union metric reached 84.7%. To date, no attempt to a precise visual segmentation by FCNN has been described in the pertinent OC/OP cancer literature.

It should be underlined that our investigation is only a preliminary assessment of feasibility and future potential that could encourage collection of additional evidence and support more extensive studies. The datasets, in fact, were relatively small and partially patient-unbalanced, denoting their high variability. In particular, the OC dataset was composed of a relatively low number of frames in relation to its considerable anatomical complexity and variability of epithelia with different histological and NBI-associated features.[26] Moreover, the mean percentage of lesion pixels in each frame was only 22.82 ± 11.68% (with respect to the 38.04 ± 18.54% of the OP dataset). Hence, the high values of Acc might be partially due to the small size of the OC lesions with respect to the entire size of each frame presented in the dataset. Finally, it is worth noticing that all FCNNs tested presented very high values of variance, leading to low values of minima. This is probably due to the difficult task related to the small size of datasets and the significant tissue variability in the regions analyzed. Additionally, OC and OP are characterized by very different endoscopic superficial appearances, with the richness in lymphoid tissue of the latter being one of the most prominent diagnostic obstacles when searching for small tumors of this site even by NBI.[26] The lower overall accuracy FCNNs in the OP in the present study may be a sign of this potential confounding factor.

In general, the type of FCNN did not lead to radical differences in the diagnostic performance in both subsites (while some minor differences may be observed in the OP). The same holds true considering inference times, that were always in the range of "real-time detection" (between 14 and 115 ms). However, in the OP it was possible to observe a higher precision in deeper architectures, demonstrating that an added layer of complexity may improve diagnostic results. Still, deeper architectures were also those needing higher inference times, thus requiring more processing power and potentially impacting on the aim of real-time segmentation. In this view, when dealing with automatic detection and segmentation of mucosal neoplastic lesions, it will be essential to find a balance between the depth of the FCNN and the time needed to detect lesions and delineate their margins.

At a subjective evaluation, all FCNNs tended to detect malignant areas where illumination was more prominent, usually in the middle of the picture (Figures 4.10-4.12). This factor hints at the importance of optimal and homogeneous illumination, which should be equally distributed throughout the visual field and not directed only on its central portion. In fact, the operator usually centers the image on the lesion to be identified, leading to a significant bias in automatic segmentation by FCNNs. The key role of illumination has also been emphasized by others,[21,22] even showing different diagnostic performances in relation to the types of endoscopic device employed.[22] In this view, novel advances in the field of image analysis should be supported by a parallel technical evolution of endoscopes, especially in terms of homogeneous illumination, high definition, colors, and optimization of image clarity.

An adjunctive limitation of this type of studies is that the "ground truth" (i.e., the image segmentation defining true tumor margins) is defined through a single expert opinion. This issue is related to the current impossibility in creating a histopathologic image to be superimposed to the endoscopic view, defining tumor margins at a microscopic level. However, independent evaluations by multiple experts may lead to a more accurate definition of endoscopic tumor margins.

Conclusions

SDS has promising potential in the analysis and segmentation of OC and OP video-endoscopic images. All tested FCNN architectures demonstrated satisfying outcomes in terms of Dsc, Acc, Rec, and Prec. However, further advances are needed to reach a diagnostic performance useful for clinical applicability. On the other hand, the inference time of the processing networks were particularly short, ranging between 14 and 115 ms, thus showing the possibility for real-time application. Future prospective studies, however, should take into account the number and quality of training images, optimizing these variables through accurate planning and data collection.

References

1. Maier-Hein L, Vedula SS, Speidel S, Navab N, Kikinis R, Park A, et al. Surgical data science for next-generation interventions. Nat BioMed Eng (2017) 1:691–6. doi: 10.1038/s41551-017-0132-7
2. Pal NR, Pal SK. A review on image segmentation techniques. Pattern Recognit (1993) 26:1277–94. doi: 10.1016/0031-3203(93)90135-J
3. 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) 14:483–92. doi: 10.1007/s11548-018-01910-0
4. Cernazanu-Glavan C, Holban S. Segmentation of bone structure in X-ray images using convolutional neural network. Adv Electr Comput Eng (2013) 13:87–94. doi: 10.4316/AECE.2013.01015
5. Watanabe A, Taniguchi M, Tsujie H, Hosokawa M, Fujita M, Sasaki S. The value of narrow band imaging endoscope for early head and neck cancers. Otolaryngol Head Neck Surg (2008) 138:446–51. doi: 10.1016/j.otohns.2007.12.034
6. Piazza C, Dessouky O, Peretti G, Cocco D, De Benedetto L, Nicolai P. Narrow-band imaging: a new tool for evaluation of head and neck squamous cell carcinomas. Review of the literature. Acta Otorhinolaryngol Ital (2008) 28:49–54.

7. Piazza C, Bon FD, Peretti G, Nicolai P. 'Biologic endoscopy': optimization of upper aerodigestive tract cancer evaluation. Curr Opin Otolaryngol Head Neck Surg (2011) 19:67–76. doi: 10.1097/MOO.0b013e328344b3ed

8. Deganello A, Paderno A, Morello R, Fior M, Berretti G, Del Bon F, et al. Diagnostic Accuracy of Narrow Band Imaging in Patients with Oral Lichen Planus: A Prospective Study. Laryngoscope (2021) 131:E1156–61. doi: 10.1002/lary.29035

9. Takano JH, Yakushiji T, Kamiyama I, Nomura T, Katakura A, Takano N, et al. Detecting early oral cancer: narrowband imaging system observation of the oral mucosa microvasculature. Int J Oral Maxillofac Surg (2010) 39:208–13. doi: 10.1016/j.ijom.2010.01.007

10. Ni XG, He S, Xu ZG, Gao L, Lu N, Yuan Z, et al. Endoscopic diagnosis of laryngeal cancer and precancerous lesions by narrow band imaging. J Laryngol Otol (2011) 125:288–96. doi: 10.1017/S0022215110002033

11. Arens C, Piazza C, Andrea M, Dikkers FG, Tjon Pian Gi RE, Voigt- Zimmermann S, 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) 273:1207–14. doi: 10.1007/s00405-015-3851-y

12. Bertino G, Cacciola S, Fernandes WBJr., Fernandes CM, Occhini A, Tinelli C, 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) 37:215–22. doi: 10.1002/hed.23582

13. Ni XG, Wang GQ, Hu FY, Xu XM, Xu L, Liu XQ, et al. Clinical utility and effectiveness of a training programme in the application of a new classification of narrow-band imaging for vocal

cord leukoplakia: A multicentre study. Clin Otolaryngol (2019) 44:729–35. doi: 10.1111/coa.13361

14. Ni XG, Zhu JQ, Zhang QQ, Zhang BG, Wang GQ. Diagnosis of vocal cord leukoplakia: The role of a novel narrow band imaging endoscopic classification. Laryngoscope (2019) 129:429–34. doi: 10.1002/lary.27346
15. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Berlin: Springer (2015). p. 234–41. doi: 10.1007/978-3-319-24574-4_28
16. Liciotti D, Paolanti M, Pietrini R, Frontoni E, Zingaretti P. Convolutional networks for semantic heads segmentation using top-view depth data in crowded environment. In: 2018 24th international conference on pattern recognition (ICPR). IEEE: New York (2018). p. 1384–9. doi: 10.1109/ ICPR.2018.8545397
17. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition. New York: IEEE (2016) pp. 770–8. doi: 10.1109/CVPR.2016.90
18. Paderno A, Holsinger FC, Piazza C. Videomics: bringing deep learning to diagnostic endoscopy. Curr Opin Otolaryngol Head Neck Surg (2021). doi: 10.1097/MOO.0000000000000697
19. Song B, Sunny S, Uthoff RD, Patrick S, Suresh A, Kolur T, et al. Automatic classification of dual-modalilty, smartphone-based oral dysplasia and malignancy images using deep learning. BioMed Opt Express (2018) 9:5318–29. doi: 10.1364/BOE.9.005318
20. Barbalata C, Mattos LS. Laryngeal Tumor Detection and Classification in Endoscopic Video. IEEE J BioMed Health Inform (2016) 20:322–32. doi: 10.1109/JBHI.2014.2374975

21. Mascharak S, Baird BJ, Holsinger FC. Detecting oropharyngeal carcinoma using multispectral, narrow-band imaging and machine learning. Laryngoscope (2018) 128:2514–20. doi: 10.1002/lary.27159

22. Ren J, Jing X, Wang J, Ren X, Xu Y, Yang Q, et al. Automatic Recognition of Laryngoscopic Images Using a Deep-Learning Technique. Laryngoscope (2020) 130:E686–93. doi: 10.1002/lary.28539

23. Inaba A, Hori K, Yoda Y, Ikematsu H, Takano H, Matsuzaki H, et al. Artificial intelligence system for detecting superficial laryngopharyngeal cancer with high efficiency of deep learning. Head Neck (2020) 42:2581–92. doi: 10.1002/ hed.26313

24. Kono M, Ishihara R, Kato Y, Miyake M, Shoji A, Inoue T, et al. Diagnosis of pharyngeal cancer on endoscopic video images by Mask region-based convolutional neural network. Dig Endosc (2020). doi: 10.1111/den.13800

25. Abe S, Oda I. Real-time pharyngeal cancer detection utilizing artificial intelligence: Journey from the proof of concept to the clinical use. Dig Endosc (2020). doi: 10.1111/den.13833

26. Tamashiro A, Yoshio T, Ishiyama A, Tsuchida T, Hijikata K, Yoshimizu S, et al. Artificial intelligence-based detection of pharyngeal cancer using convolutional neural networks. Dig Endosc (2020). doi: 10.1111/den.13653

27. Piazza C, Del Bon F, Paderno A, Grazioli P, Perotti P, Barbieri D, et al. The diagnostic value of narrow band imaging in different oral and oropharyngeal subsites. Eur Arch Otorhinolaryngol (2016) 273:3347–53. doi: 10.1007/s00405- 016-3925-5

5. Instance segmentation of upper aerodigestive tract cancer: Site-specific outcomes

From proof of concept to structured clinical study: Overall assessment of CNNs for the semantic segmentation of UADT lesions

Introduction

The application of computer vision techniques in diagnostic videoendoscopies (i.e. Videomics)[1,2] is a promising research field that is currently showing a fast rate of growth in many medical specialties. The recent refinement of deep learning (DL) algorithms for image processing and their application in the medical field opened novel possibilities in the management of endoscopic exams that, in the past, had only subjective value. In particular, videoendoscopy is a key component in the management of upper aerodigestive tract (UADT) tumors, influencing their entire diagnostic process, treatment, and follow-up. Still, it remains a purely operator-dependent and time-consuming procedure, which is substantially limited by the variables of human experience and perception. This is especially true when endoscopy is applied in conjunction with optical biopsy techniques such as Narrow Band Imaging (NBI), requiring an even more specialized training and adding a further layer of complexity and subjectivity. Finally, no easily classifiable and structured data can be drawn from these examinations, significantly limiting their integration with other technologies (e.g., cross sectional imaging, ultrasound, genomic markers, and so on).

Our study aimed to explore the potential of a novel DL algorithm, Mask R-CNN[3], in the diagnostic approach to UADT squamous cell carcinoma (SCC). The primary goal was to detect and classify neoplastic lesions and, at the same time, precisely define their margins, a task overall defined as “instance segmentation”. In fact, Mask R-CNN provides a flexible and general framework for object instance segmentation that can also be potentially applied to medical images. This approach combines elements from the tasks of object detection (where the goal is to localize the lesion using a bounding box), object classification (where the purpose is to classify each pixel into a set of categories – e.g., tumor vs. normal mucosa), and semantic segmentation (where the aim is to automatically delineate the lesion’s margins). Finally, we included in our analysis three different areas of the UADT (oral cavity, oropharynx, larynx/hypopharynx) in order to identify potential site-related differences in the diagnostic capability of this DL algorithm, an information that is still lacking in the current specialized literature. In fact, studies assessing the value of artificial intelligence in endoscopy are generally focused on a single site and are difficult to generalize in the context of UADT SCC, which can arise from a wide variety of anatomical structures, as well as epithelial and mucosal types.

Materials and methods

A retrospective study was performed including videoendoscopies performed between September 2009 and January 2021 in patients treated at the Unit of Otorhinolaryngology – Head and Neck Surgery, University of Brescia, Italy for SCC of the UADT. A total of 7.567 videoendoscopies were collected from a dedicated archive. All recordings were anonymized and associated with the corresponding histopathologic report.

The study primary endpoint was the definition of the diagnostic accuracy (in terms of Dice Similarity Coefficient [Dsc]) of the Mask R-CNN algorithm when applied to NBI UADT videoendoscopic frames. The secondary endpoint was the comparison of the algorithm’s Dsc in the three different anatomical areas herein considered.

Inclusion criteria were as follows:

- Primary or recurrent SCC of the UADT (distinguished between those occurring in the oral cavity, oropharynx, and larynx/hypopharynx);
- NBI evaluation with adequate quality (without pooling of saliva, blood spots, swallowing reflex, coughing or other technical issues);
- available histological examination obtained at the time of videoendoscopy or subsequent surgery.

All patients were examined both under white light (WL) and NBI through transnasal videolaryngoscopy (HD Video Rhino-laryngoscope Olympus ENF-VH, ENF-VQ, or ENF-V2, Olympus Medical System Corporation, Tokyo, Japan) or through transoral endoscopy by 0° rigid telescope coupled to an Evis Exera II HD camera connected to an Evis Exera II CLV-180B/III CV-190 light source (Olympus Medical Systems Corporation, Tokyo, Japan). Endoscopic videos were selected independently by two otolaryngologists with extensive experience (at least 4 years) in endoscopic assessment of UADT lesions by NBI and independently reviewed by an adjunctive expert. Images were then manually quality-controlled, with exclusion of those that were blurred, obscured by blood or secretions, or without adequate NBI evaluation.

Image processing

Three representative frames per video were selected for every lesion and saved in .jpeg format. The most representative NBI videoframe was chosen and subsequent frames at 0.3 seconds time intervals were then automatically selected. Frame annotation was performed manually using the LabelMe application[4]. Annotations consisted of a variable number of key points marking the lesion margins in the videoendoscopic frame taking into account positive NBI patterns. The resulting masks were then saved in .json format and stored in a dedicated folder. Two clinical experts concomitantly annotated the images and a further review was performed by a senior staff member. When an agreement regarding lesion margins was not reached, the frame was excluded from the analysis.

After this selection process, a total of 1034 endoscopic images were obtained. Three different sub-datasets were generated according to the lesion primary site: oral cavity, oropharynx, and larynx/hypopharynx. In this way, the total frames analyzed were 653 for the larynx/hypopharynx, 246 for the oral cavity, and 135 for the oropharynx.

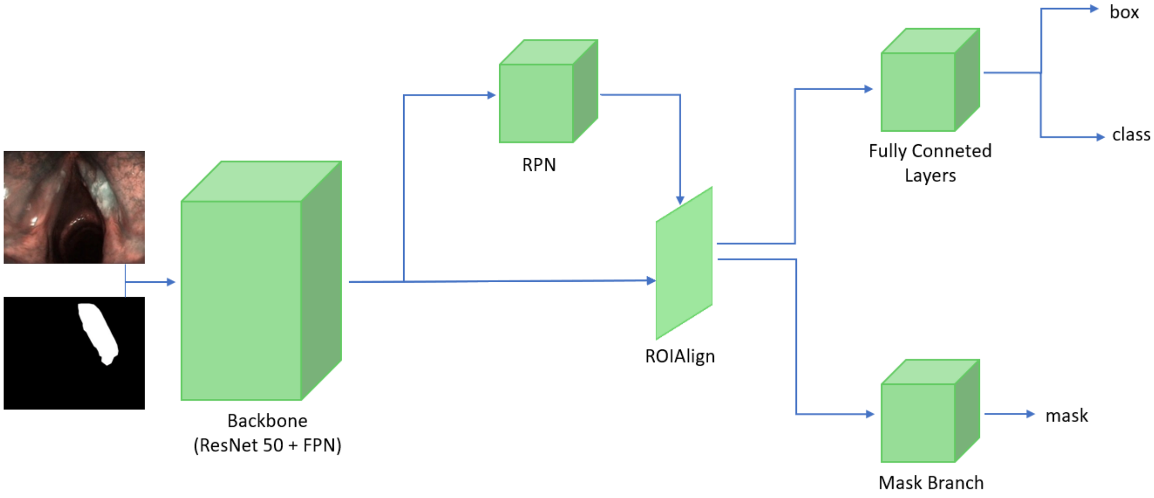
Dataset

The dataset included 1034 images from 323 patients. For algorithm training and testing the dataset was split over patients and balancing the three classes into three sets: 935 images from 290 subjects for training, 48 images from 16 subjects for validation, and 51 images from 17 subjects for testing. All images were resized to the same dimension of 480x640 pixels.

DL analysis

In this work, Mask R-CNN[5] was used to segment the tumor in endoscopic frames. This convolutional neural network (CNN) consists of backbone, Region Proposal Network (RPN), and three heads for classification, bounding-box regression, and segmentation (Figure 5.1).

Figure 5.1. *Schematic representation of the proposed architecture. The Mask R-CNN is made of a backbone (composed by a ResNet50 and a feature pyramid network), a region proposal network (RPN), ROIAlign, and three heads, for classification, bounding-box regression, and segmentation.*



As backbone, we used the ResNet50[6] combined with the Feature Pyramid Network (FPN),[7] to extract features from the input frame at multiple scales. Starting from the features computed with the backbone, the RPN identifies candidate regions containing the tumor. For each of the proposed regions, the final bounding box containing the tumor and the tumor segmentation are obtained from the three heads.

To cope with the relatively limited size of the dataset, we used the weights computed on the COCO dataset[8] to initialize the layers of Mask R-CNN. To reduce the risk of overfitting, we performed on-the-fly data augmentation during training by applying: random brightness changes in the range (0.5, 1.1), random contrast changes in the range (0.8, 3), and random rotation in the range (-20, 20).

The model was trained for 100 epochs, using the Stochastic Gradient Descent (SGD) as optimizer with an initial learning rate of 0.001 and momentum of 0.9. We used a loss which is the combination of different contributions: $\boxed{L = L_{cls} + L_{box\_reg} + L_{rpn\_cls} + L_{rpn\_loc} + L_{mask}}$ where $\boxed{L_{cls}}$ is the loss in the classification head, $\boxed{L_{box\_reg}}$ is the loss in bounding-box regression head, $\boxed{L_{rpn\_cls}}$ is the classification loss in the RPN, $\boxed{L_{rpn\_loc}}$ is the localization loss in the RPN, and

L_{mask} is the loss in segmentation head. The loss equations can be found in the original Mask R-CNN paper.[5]

Performance metrics and statistical analysis

As a primary endpoint the segmentation performance was evaluated using the Dsc, which is a statistical validation metric based on the spatial overlap between the predicted (A_{mask}) and ground-truth (A_{gt}) segmentation:

$$DSC = \frac{2 \text{ x } |A_{gt} \cap \ A_{mask}|}{|A_{gt}| + |A_{mask}|}$$

Dsc can assume values in a range from 0, indicating no overlap, to 1, indicating complete overlap. Furthermore, outcomes were also evaluated using the following spatial overlap-based metrics:

Pixel accuracy (Acc) represents the percent of pixels in the image which are correctly classified.

It is defined as: $Acc = \frac{TP + TN}{TP + TN + FP + FN}$

where TP, TN, FP, FN denote the true positives, true negatives, false positives, and false negatives, respectively.

Recall (Rec), also known as Sensitivity or True Positive Rate, defines the portion of positive pixels in the ground-truth which are also identified as positive in the predicted segmentation.

It is defined as: $Rec = \frac{TP}{TP + FN}$

Specificity (Spec), or True Negative Rate, measures the portion of negative pixels (background) in the ground-truth that are also identified as negative in the predicted segmentation.

It is defined as: $Spec = \frac{TN}{TN + FP}$

Precision (Prec), or Positive Predictive Value, measures how accurate the predictions are, i.e. the percentage of correct predictions.

It is defined as: $\boxed{Prec = \frac{TP}{TP + FP}}$

F1-score is a balance between precision and recall, also known as harmonic mean.

It is defined as: $\boxed{F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}}$

Intersection over Union (IoU), also referred to as Jaccard index, represents the area of overlap between the predicted segmentation and the ground truth divided by the area of union between the predicted segmentation and the ground truth.

It is defined as: $\boxed{\mathrm{IoU} = \frac{TP}{TP + FP + FN}}$

Mean Average Precision (mAP), which represents the average of the area under the Recall-Precision curve, was also computed.

Outcomes were compared between the different subsites analyzed using non-parametric statistics. The Kruskal-Wallis H-test was used for the overall comparison and the Mann-Whitney U rank test for pair comparisons. Statistical analysis was performed using Jupiter Notebook 6.4.5 with pandas 1.4.1 and ScyPy 1.8.0 libraries.

Results

Overall performance

The total number of images used for the test set was 51. The algorithm demonstrated the ability to correctly predict 39 images out of 51 images (76.5%). The average Dsc score was 0.79 (range, 0.26-0.97; standard deviation (SD), 0.22). Overall and site-specific performance metrics are summarized in Table 5.1 and Figure 5.2. Samples of the segmentation results are presented in Figure 5.3.

Table 5.1. *Summary of the diagnostic performance according to different metrics.* ***Dsc****: Dice similarity coefficient;* ***IoU****: Intersection over Union;* ***SD****: standard deviation.*

Mean value (SD)	**Overall**	**Larynx/hypophaynx**	**Oral cavity**	**Oropharynx**
Dsc	0.79 ±0.23	0.90 ± 0.05	0.60± 0.26	0.80 ± 0.30
Accuracy	0.91 ± 0.12	0.98 ± 0.01	0.79 ± 0.13	0.92 ± 0.14
Specificity	0.93 ± 0.12	0.98 ± 0.01	0.86 ± 0.16	0.92 ± 0.15
Precision	0.85 ± 0.24	0.94 ± 0.06	0.73 ± 0.32	0.79 ± 0.36
Recall	0.86 ± 0.22	0.91 ± 0.08	0.73 ± 0.33	0.95 ± 0.04
IoU	0.73 ± 0.27	0.87 ± 0.09	0.49 ± 0.30	0.76 ± 0.14
F1 score	0.80 ± 0.23	0.92 ± 0.05	0.61 ± 0.27	0.81 ± 0.31

Figure 5.2. *Box plots detailing the diagnostic accuracy of the algorithm in different sites according to various metrics. A: Dice Similarity Coefficient (DSC); B: Accuracy; C: Specificity; D: Precision; E: Recall; F: Intersection Over Union (IoU); G: F1 score.*

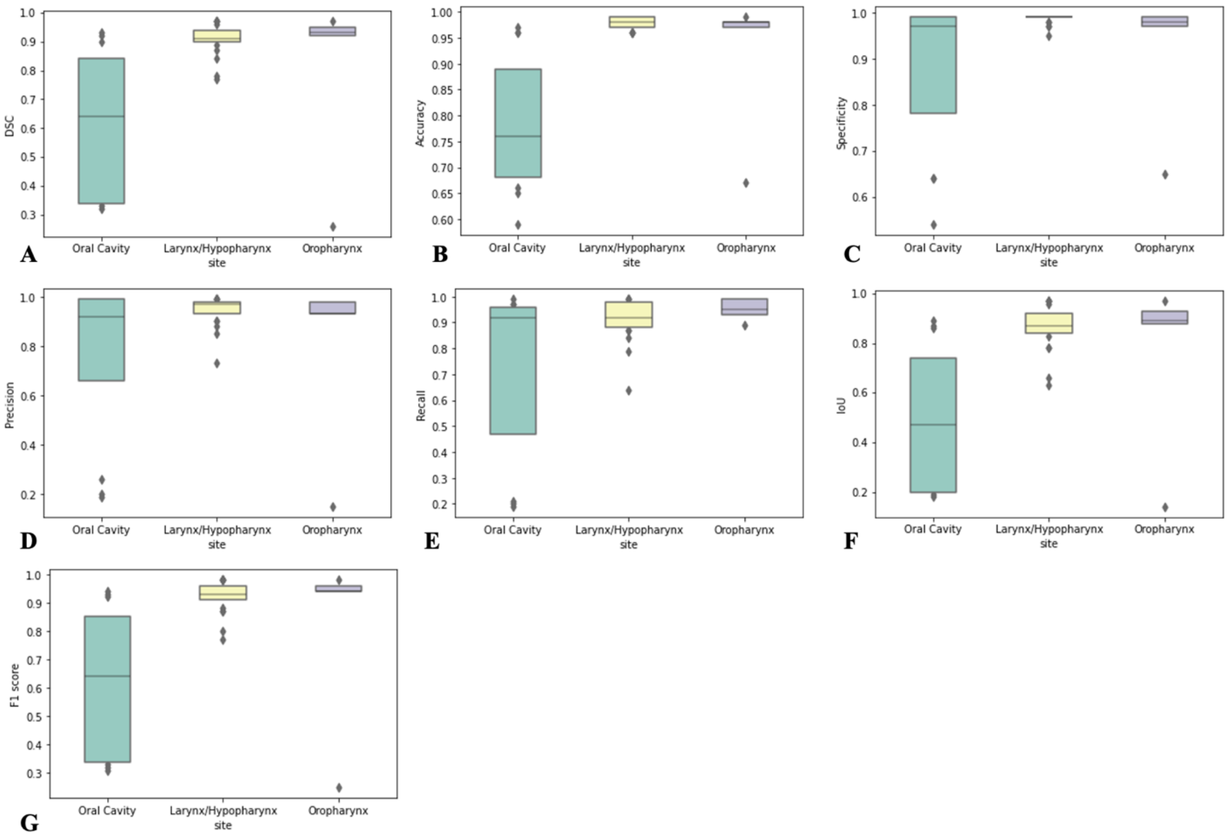
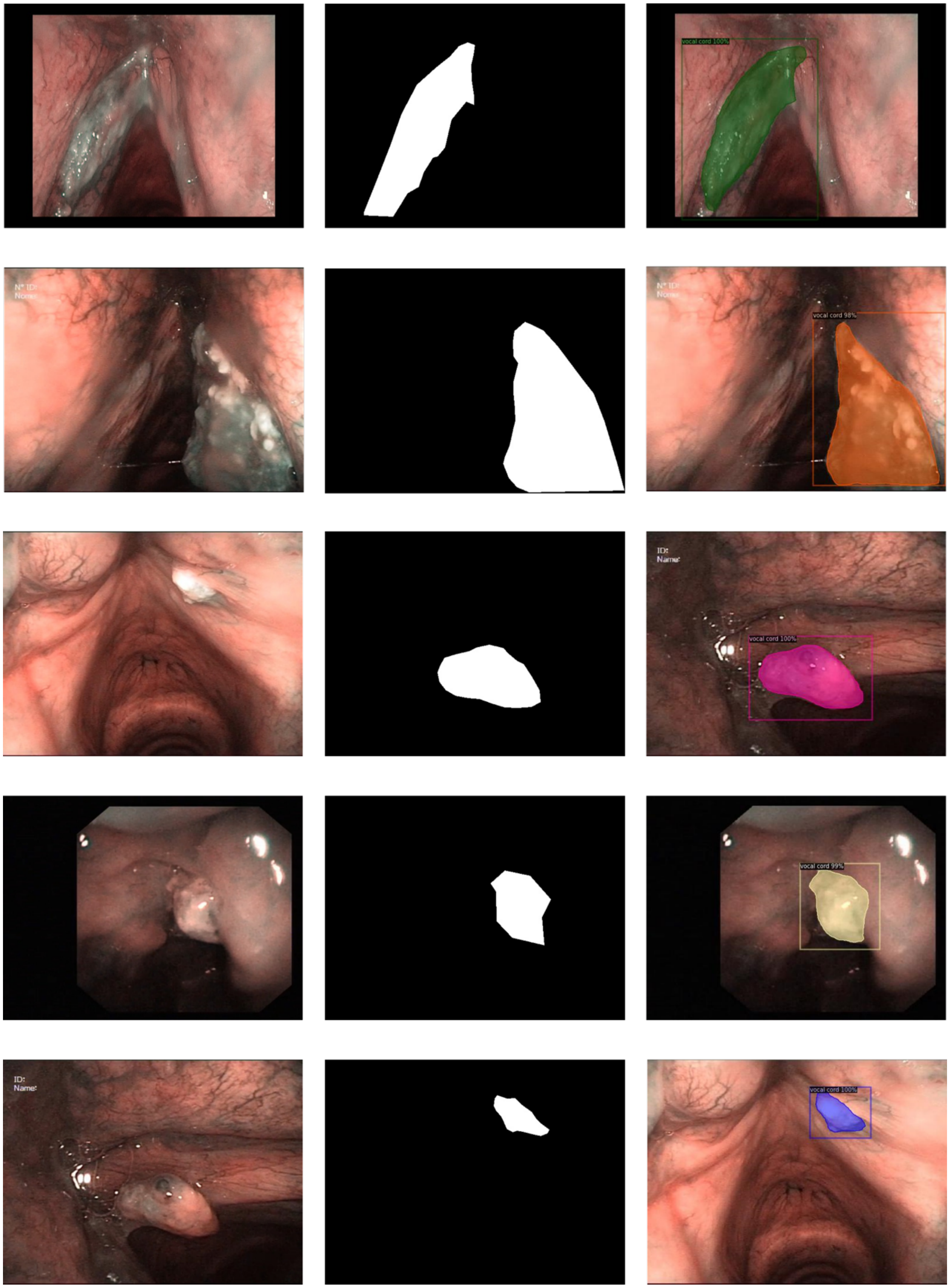


Figure 5.3. *Visual samples of the segmentation results. From left to right: raw endoscopic frames, ground truth annotation, and predictions obtained with the proposed method.*


vocal cord 100%
N° ID:
Nome:
N° ID:
Nome:
vocal cord 98%
ID:
Name:
vocal cord 100%
vocal cord 99%
ID:
Name:
vocal cord 100%

Laryngeal/hypopharyngeal lesions

The total number of images representative for laryngeal and hypopharyngeal lesions in the test set were 27 (52.9% of the test dataset). Out of that number, our algorithm correctly predicted 21 lesions (77.8%). The mean Dsc score was 0.90+/-0.05, the first quartile was 0.90 and the third quartile 0.94 (Table 5.1).

Oral lesions

The oral lesions comprised in the test set were 15 (29.4% of the total). The algorithm performed a correct prediction in 13 cases (86.7%). The mean Dsc score was 0.60+/-0.26, the first quartile was 0.34 and the third quartile 0.84 (Table 5.1).

Oropharyngeal lesions

In the test set, the oropharyngeal lesions were 9 in total out of 51 images (17.6%). The algorithm predicted 5 images (55.5%). The mean value of Dsc score was 0.81+/-0.30, the first quartile was 0.92 and the third quartile 0.95 (Table 5.1).

Comparison between three different UADT sites

Results for each site are summarized in Table 5.1. The overall diagnostic performance, defined by the Dsc score, was significantly different between the different sites (p=0.002). Pairwise analysis showed that the difference was related to significantly inferior results in the oral cavity when compared with larynx/hypopharynx (p<0.001).

Diagnostic results proved to be significantly correlated with the site analyzed also considering other performance metrics: accuracy (p<0.001), specificity (p=0.02), IoU (p=0.002), and F1 score (p=0.002). As above, this difference is related to inferior results in the oral cavity vs. larynx/hypopharynx. However, when considering accuracy, it is also possible to evidence a significant difference between oral cavity and oropharynx (p=0.03).

Discussion

In this study, we evaluated for the first time the specific task of instance segmentation in clinical endoscopy for head and neck SCC. The analysis included three sites of the UADT to allow a comparison of the algorithm's diagnostic performance in different anatomical areas. The algorithm was able to identify and segment the lesion in 76.5% of cases, and showed remarkable diagnostic accuracy, especially in consideration of the complex task to be performed. Interestingly, results were significantly inferior in the oral cavity, where all outcome measures underperformed when compared with larynx/hypopharynx and, in some cases (i.e., accuracy), oropharynx. This is in line with what previously observed by Piazza and coworkers[9] when applying bioendoscopic tools such as NBI. This result is possibly related to the wide array of epithelial subtypes observed in the oral cavity, adjunctive limits specifically correlated with oral examination (e.g., presence of light artifacts), and confounding factors (e.g., tongue blade, teeth, or dentures) that the ML software must learn to take into account.

Instance segmentation represents the ultimate step in video analysis since it allows at the same time detection, classification, and segmentation of multiple elements in each single frame, which is possible thanks to the integration of different analytic components in the same general algorithm. This approach is particularly suited to the context of UADT endoscopy since different alterations (e.g., concomitant inflammatory or benign lesions) can be frequently encountered in the field of view together with the target lesion, and due to the fact that patients with head and neck SCC can develop distinct islands of neoplastic or dysplastic mucosa (i.e., field of cancerization) that might involve various portions of the videoframe, even without continuity.

In general, recent CNN-based methods have demonstrated remarkable results in segmentation of the UADT and proved to be well-suited for such a complex task. Laves et al.[10] first demonstrated that a weighted average ensemble network of UNet and ErfNet were the best suited for laryngeal segmentation of intraoperative images under direct laryngoscopy, with a mean IoU of 84.7%. However, different authors subsequently strived toward development of diagnostic algorithms that could be applied in real time in office-based and intraoperative endoscopy. Paderno et al.[11] explored the use of fully CNNs for real-time segmentation of SCC in the oral cavity and oropharynx. In this work, different architectures were compared detailing their diagnostic performance and inference time, demonstrating the possibility to achieve real-time segmentation. In accordance with previous findings in literature, the present study confirms that the oral cavity may have inferior diagnostic results due to the high variability of subsites when compared with other areas of the UADT (i.e., oropharynx, larynx, and hypopharynx). When dealing with normal laryngeal anatomy, Fehling et al.[12] explored the possibility to achieve a fully automated segmentation of the glottic area using a CNN in high-speed laryngeal videos. The algorithm obtained a Dsc over 0.85 for all subsites analyzed. Finally, Li et al.[13] proposed a method to segment nasopharyngeal malignancies in endoscopic images based on DL, reaching an accuracy of 88.0%. However, progressive advances in automatic segmentation of the UADT can be observed thanks to a recent article by Azam et al.[2], in which SegMENT, a novel CNN-based segmentation model, outperformed previously published results on the external validation cohorts. The model was initially trained on WL and NBI endoscopic frames of laryngeal SCC, but also showed to be effective in the segmentation of independent frames of oral and oropharyngeal cancer. The authors stated that the model demonstrated potential for improved detection of early tumors, more precise biopsies, and better selection of resection margins.

In general, results of automatic segmentation are inferior to those obtained in more straightforward tasks such as frame classification [14] [15] [16] or lesion detection [17] [18] since a more in-depth conceptual model of UADT lesions is required to allow accurate definition of margins. However, semantic segmentation is a key objective when striving towards more complex tasks involving computer vision and human-machine interaction. In fact, other than providing a purely diagnostic tool, a comprehensive understanding of all UADT alterations and suspicious lesions may grant significant aid in intraoperative management. This is even more true when considering instance segmentation, which epitomizes in itself all the needs and requirements of the visual examination of endoscopic images, allowing a full automatic understanding of complex endoscopic scenarios, even those involving more than one lesion and/or more than one pathology.

Potential issues have been addressed to limit biases related to the analysis technique:

- patients (and their related frames) in the training, validation, and test sets have been distinguished into separated groups to avoid overfitting;
- Frames were annotated and reviewed by 3 experts to limit subjective errors;
- Frame selection and data augmentation were performed to reduce the impact of artifacts or technical biases.

However, intrinsic limits should be acknowledged. In particular, the gold standard over which the algorithm has been trained (i.e., the “ground truth”) is represented by an expert opinion of the tumor margins and not by the histopathological definition per se. In fact, as of today, it is not technically possible to provide a direct “in situ”, “in vivo” morphologic correlation between endoscopic images and their histopathological specimen.

References

1. 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

2. 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;12:900451. doi:10.3389/fonc.2022.900451

3. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. IEEE Trans Pattern Anal Mach Intell 2020;42(2):386-397. doi:10.1109/TPAMI.2018.2844175

4. Russell BC.; Torralba A, Murphy KP et al. Label Me: A Database and Web-Based Tool for Image Annotation. International Journal of Computer Vision 2008. 77 (1–3): 157–173. doi:10.1007/s11263-007-0090-8.

5. He K, Gkioxari G, Dollár P, Girshick RB, Mask R-CNN. IEEE International Conference on Computer Vision (ICCV), vol. 2017; 2017. p. 2980–8.

6. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. IEEE Confer Comput Vision Pattern Recogn (CVPR). 2016; 2016:770–8.

7. Lin T, Dollár P, Girshick RB, He K, Hariharan B, Belongie SJ. Feature pyramid networks for object detection. IEEE Conference Computer Vision Pattern Recognition (CVPR). 2017; 2017:936–44.

8. Lin T, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft coco: Common objects in context. In European Conference on Computer Vision. Cham: Springer; 2014. p. 740–55.

9. Piazza C, Del Bon F, Peretti G, Nicolai P. "Biologic endoscopy": optimization of upper aerodigestive tract cancer evaluation. Curr Opin Otolaryngol Head Neck Surg 2011;19(2):67-76. doi:10.1097/MOO.0b013e328344b3ed

10. Laves MH, Bicker J, Kahrs LA, et al. 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

11. 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;11:626602. doi:10.3389/fonc.2021.626602

12. Fehling MK, Grosch F, Schuster ME et al. Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep Convolutional LSTM Network. PloS One 2020;15(2):e0227791. doi:10.1371/journal.pone.0227791

13. Li C, Jing B, Ke L, et al. Development and validation of an endoscopic images-based deep learning model for detection with nasopharyngeal malignancies. Cancer Commun Lond Engl 2018;38(1):59. doi:10.1186/s40880-018-0325-9

14. Song B, Sunny S, Uthoff RD, et al. Automatic classification of dual-modalilty, smartphone-based oral dysplasia and malignancy images using deep learning. Biomed Opt Express 2018;9(11):5318. doi:10.1364/BOE.9.005318

15. Esmaeili N, Sharaf E, Gomes Ataide EJ, et al. Deep Convolution Neural Network for Laryngeal Cancer Classification on Contact Endoscopy-Narrow Band Imaging. Sensors 2021;21(23):8157. doi:10.3390/s21238157

16. Dunham ME, Kong KA, McWhorter AJ, et al. Optical Biopsy: Automated Classification of Airway Endoscopic Findings Using a Convolutional Neural Network. The Laryngoscope 2022;132(S4):S1-S8. doi:10.1002/lary.28708

17. 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

18. 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. The Laryngoscope 2022;132(9):1798-1806. doi:10.1002/lary.29960

6. Classifying vocal folds fixation from endoscopic videos with machine learning

From static vision to motion: The role of vocal fold motility

Introduction

With the advent of artificial intelligence (AI), the last decade has seen a revolution in the field of medical-image analysis, with applications ranging from diagnosis to treatment, guidance, and follow-up.[1] While promising results were obtained for processing anatomical images, such as computerized-tomography,[2] ultrasound,[3] and magnetic-resonance images,[4] the analysis of endoscopic videos still represents a challenge[5] and only few commercially available solutions exist.[6] This may be explained considering the peculiar challenges of endoscopic videos, including poor contrast, low signal-to-noise ratio, presence of motion blurring, and tissue motion. The field of otolaryngology and head and neck surgery makes not an exception.[7] Videoendoscopy is largely used in clinical practice for a number of applications, among which the assessment of vocal cords motility. Paralysis of one or both vocal folds may jeopardize key physiological functions of the larynx, from breathing to airway protection and phonation.[8] The clinical diagnosis relies on the subjective examination and interpretation of vocal folds motion during real-time viewing or playback of videos captured through videoendoscopy. This evaluation is time-consuming, requires a skilled professional to be performed and characterized by high inter- and intra-rater variability.[9] In this context, machine learning (ML) has the potential to tackle the variability of videoendoscopic frames, and to provide a quantitative perspective to the analysis of vocal folds motility. In this paper, we focus on the analysis of endoscopic frames extracted from endoscopic videos, proposing a ML algorithm for the assessment of vocal folds motility.

The literature on ML algorithms for laryngeal videoendoscopic image analysis has been growing since 2017. The work by Moccia et al.[10] is among the first ones to investigate the use of ML algorithms for early stage cancerous laryngeal tissue classification. Since then, several studies have been published.[7] Motility assessment, instead, is mostly addressed with deep learning (DL) methods based on glottal segmentation, from which the motility is evaluated based on the movement of each fold with respect to the midline; or through region of interest detection, and glottal gap delimitation.[11] Hamad et al.[12] developed a DL system for automatic segmentation of the glottal region in laryngoscopy videos using a fully convolutional regression network. More recently, Yousef et al.[13] studied vocal folds kinematics during the running speech, analyzing vocal folds vibrations in adductor spasmodic dysphonia. A U-Net model was deployed for glottal area segmentation in high-speed videoendoscopy to quantitatively analyze vibrations in both healthy and unhealthy patients. Similarly, in the work by Yousef et al.[14] vocal folds dynamics is evaluated in association with voice disorders. They trained a deep neural network with data from laryngeal high-speed videoendoscopy with the aim of segmenting the glottal area, from which the glottal edges are derived during connected speech. Other studies make use of phasegram[15] (a visualization method of system dynamics that can be interpreted as a bifurcation diagram in time) or phonovibrogram,[9,16] (a graphical representation of the vocal folds deflections, automatically extracted from laryngeal high speed recordings) to evaluate vocal folds motility related with voice disorders. However, unlike videoendoscopy, these kinds of tests are not usually performed in clinical practice.

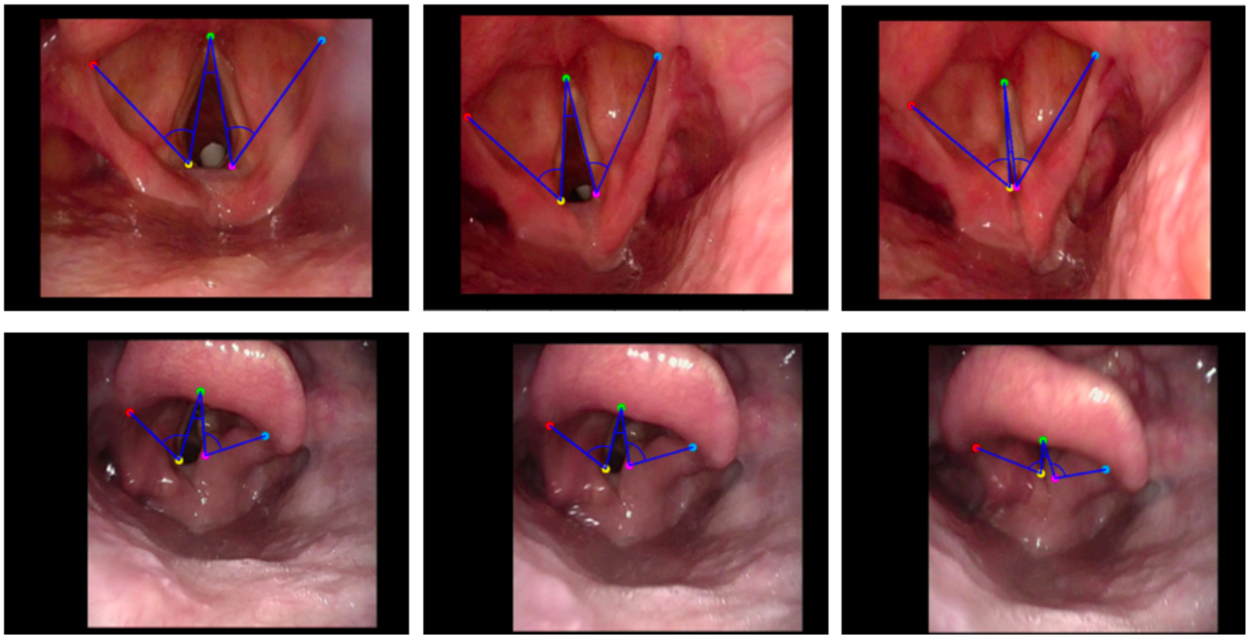
Differently from the work in the literature, we rely on ML for vocal folds motility estimation. We propose, in fact, a method to classify motility into two classes (namely: preserved motility and fixation) based on keypoints. This method is advantageous as it allows to directly obtain a classification, without the need of post-processing, as in the case of glottal segmentation. Each of the selected keypoints represents an important clinical landmark for the analysis, providing a close approximation of both glottic and arytenoid movements. Starting from the coordinates of the five key-points, clinically relevant features were handcrafted to train the classification models.

Methods

Vocal folds model and keypoints annotation

The dataset used for this analysis is made of videoendoscopic frames of patients treated at the Unit of Otorhinolaryngology - Head and Neck Surgery, University of Brescia, Italy. Data were acquired following the principles of the Helsinki Declaration, and approval was obtained by the local ethical committee of Spedali Civili of Brescia. A total of 558 endoscopic images from 186 patients was collected from a dedicated archive and anonymized, and for each video three representative frames were selected. The motility was estimated among these three endoscopic frames from five keypoints located at specific sites of the larynx: the epiglottic insertion point of the left and right aryepiglottic folds (LE and RE, respectively), the posterior angle of the left and right vocal folds (LV and RV, respectively), and the anterior commissure (A), as shown in Figure 6.1. RE and LE represent the insertion of the aryepiglottic folds in the epiglottis. In particular, they are one of the pivot points that remain fixed when the arytenoid moves (together with the aryepiglottic fold). Hence, they are suitable reference points when trying to assess the movements of the supraglottic larynx, using the angle formed by them and the vocal folds.

***Figure 6.1**. Representation of three consecutive frames (from left: abducted, normal, and adducted vocal cords, respectively) with ground truth keypoints annotation. The images in the first row refer to a subject with preserved motility, while the ones in the second row to a subject with fixation. The colored points represent the keypoints: left epiglottic in red, left vocal fold in yellow, anterior commissure in green, right vocal fold in magenta, right epiglottic in cyan.*



Frames annotation was performed by an expert (more than 10 years of experience) laryngologist using LabelMe1. Only subjects for which three frames representing a specific vocal folds position (abducted, neutral, adducted) were available, were included in the study. After this process of data selection, the collected dataset counted 101 subjects with preserved motility and 51 subjects with fixation. The dataset includes both oncologic and non-oncologic patients.

Feature extraction and classification

To assess vocal folds motility, we extracted the following features from the labeled frames:

- The central and the two external angles for each frame (as shown in Figure 6.1).

- The static index: the difference between the two external angles for each of the three frames.
- The dynamic index: the ratio between the difference of the right angle in the first and third frames and the difference of the left angle in the first and third frames.

We investigated common ML classification algorithms, including support vector machines with linear (SVC) and non linear (SVM) kernels, XGBoost (XGB), and random forest (RF). The optimal hyperparameters for each classifier were retrieved via grid-search and cross validation on the training set, using stratified three-fold cross validation. This ensures that every patient in our dataset appears at least once in the testing set. In particular, the three-fold cross validation cyclically splits the dataset into three equally sized folds, of which two are used to train and one to validate and tune the parameters. Before classification, features were normalized by removing the mean (centering) and scaling to unit variance. Given the unbalance between the two classes, the minority class was over-sampled using the synthetic minority oversampling technique (SMOTE). Also, class weights were balanced according to the number of samples of each class.

Experimental Analysis

The performance of the classifiers was evaluated using classification precision (Prec), recall (Rec), and F1-score (F1) on the test set. Considering the unbalance of our dataset, the area under the precision-recall curve (AUC) and the average precision (AP) were also computed.

Results

The performance of all the classifiers is shown in Table 6.1, results are reported in terms of the metrics computed on the test set. Figure 6.2 shows the precision-recall curves of all the classifiers.

All the tested models showed comparable results, however, the best-performing classification algorithm resulted to be the XGB, with an AP of 0.76 and 0.94, and an AUC of 0.76 and 0.93 for the fixation and preserved motility class, respectively. Features importance of the XGB classifier is reported in Figure 6.3. Specifically, on 152 test subjects (among the three cross validation folds), XGB achieved the lowest number of incorrect predictions (27 subjects). Samples of misclassified frames are shown in Figure 6.4.

Table 6.1. *Performance evaluation metrics. Precision (Prec), recall (Rec), F1-score (F1), accuracy (Acc), average preci- sion (AP), and area under the precision-recall curve (AUC) are reported. For each classifier, the first row refers to the class fixation, while the second to the class preserved motility.*

Classifier	Prec	Rec	F1	Acc	AP	AUC
	0.73	0.73	0.73		0.75	0.73
SVC	0.86	0.86	0.86	0.82	0.90	0.90
	0.71	0.76	0.74		0.72	0.71
SVM	0.88	0.84	0.86	0.82	0.93	0.92
	0.67	0.59	0.62		0.64	0.63
RF	0.80	0.85	0.83	0.76	0.89	0.89
	0.76	0.69	0.72		0.76	0.76
XGB	0.85	0.89	0.87	0.82	0.94	0.93

Figure 6.2*. Precision-Recall curves calculated on the test set, for all the classifiers. The best-performing classifier resulted to be XGB, showing the highest average precision and area under the precision-recall curve for both classes.*

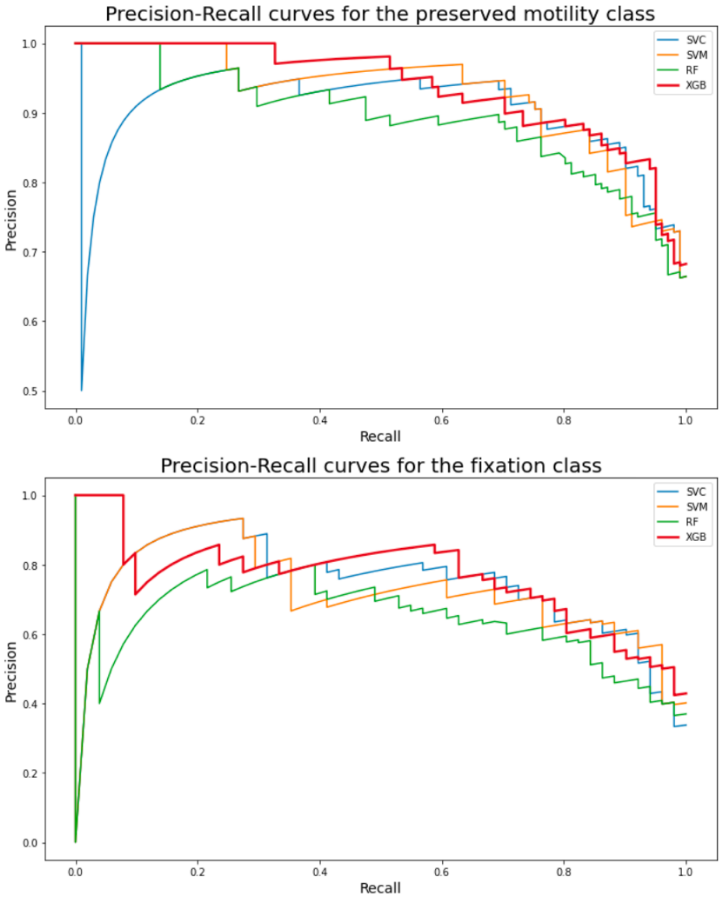


Figure 6.3. *Features importance of the XGB classifier. Features from 0 to 8 refer to the three angles (central and externals) of the three successive frames, features from 9 to 11 refer to the static indexes of the three frames, and feature 12 refers to the dynamic index.*

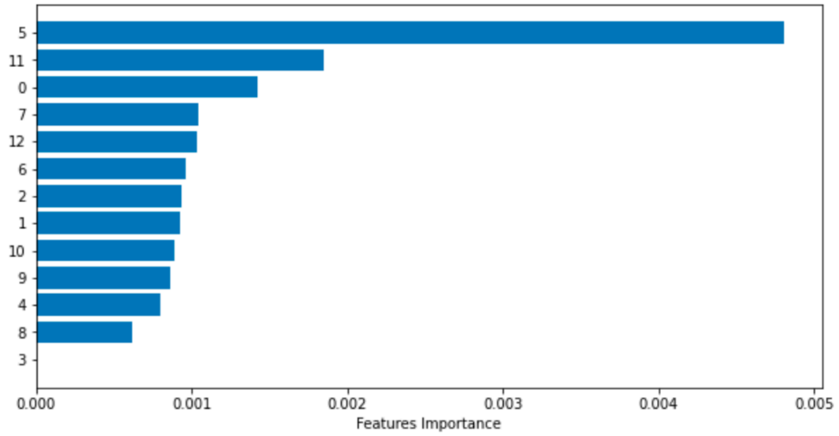


Figure 6.4. *Visual samples of misclassified frames. The images in the first row were erroneously predicted as belonging to the fixation class, while the images in the second row were erroneously predicted as belonging to the preserved motility class. In the latter*

case, the vocal folds area occupies a small portion of the frame, which makes the prediction more challenging.

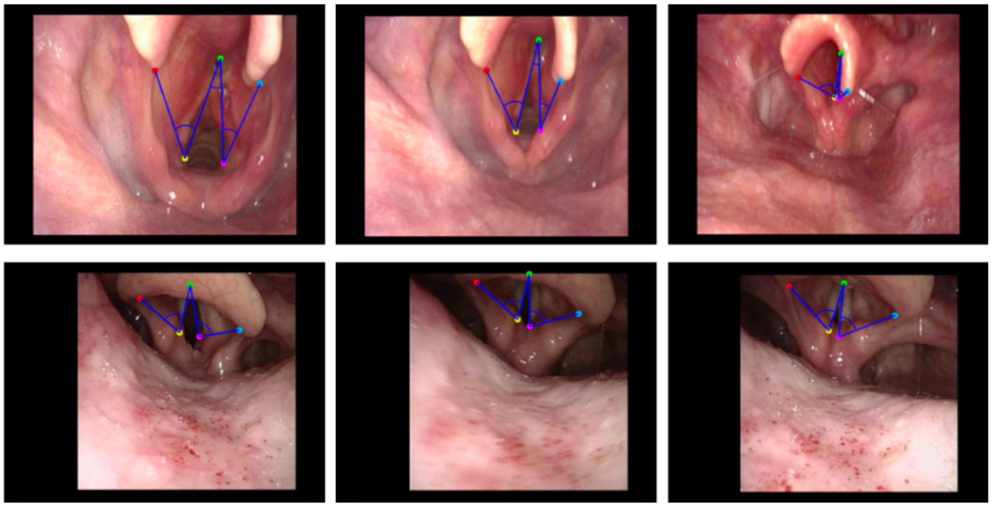
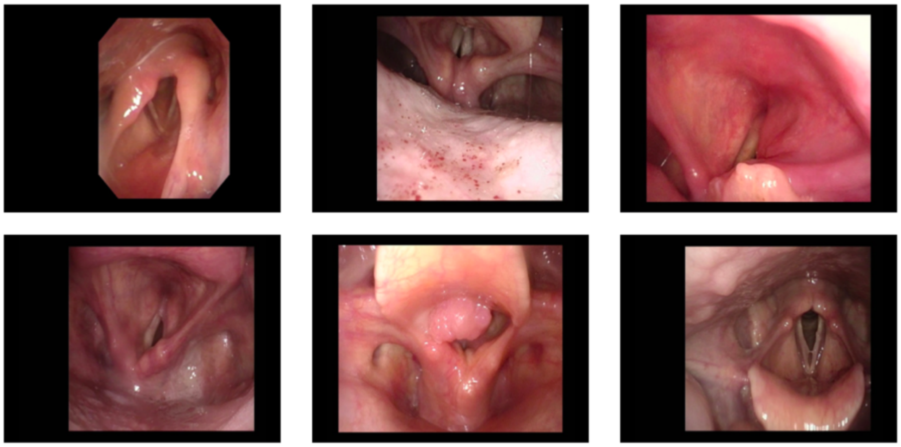


***Figure 6.5**. Visual samples of frames from the used dataset. It is characterized by high variability among the frames, which reflects also on the variability of the features used to train the models.*



Discussion

The main objective of this study was to evaluate the ability of ML algorithms to discriminate between vocal cords preserved motility and fixation. To do so, we extracted a number of relevant features from triplets of videoendoscopic frames, representing specific vocal folds positions. The extracted features were used to train and test four different classifiers, which showed good results, and the best-performing resulted to be the XGB. Even though the results of this model do not depart from the others, the use of this specific ML classifier could be useful in case of some not labeled keypoints, as it is able to handle missing values.[17] From the results, it is also possible to appreciate the ability of all the tested models in assessing vocal cords motility. This is an expected behavior[18,19] and confirms that the application on ML may have a positive impact to assist clinicians in their practice.

To the best of our knowledge, this is the first study to rely on keypoints to evaluate vocal cords motility. Previous work in literature, in fact, focused on the segmentation of the glottis to evaluate the motility. The advantage of relying on keypoints, as already demonstrated in precedent work from other fields,[20,21] is the possibility to obtain a direct classification. Methods relying on segmentation, in fact, need a post-processing step to obtain a diagnosis. A limitation of the proposed work could be seen in the relatively limited size of the dataset, which is due to the time needed to label each frame, and to the lack of available annotated dataset online. The time consuming annotation procedure also makes it difficult, at the moment, to evaluate intra-observer variability. Moreover, the dataset used in this work includes frames with very high variability among each other, as shown in Figure 6.5, which is typical of videoendoscopic frames. This characteristic of the dataset reflects also on the extracted features and on the achieved results. For this reason, adding the classification algorithm downstream of a frame selection process might improve the results. As future work, to support clinicians in the actual clinical practice, the classification model could be included within other computer assisted algorithms for diagnostic support, e.g., frames selection and automatic keypoints regression.

Conclusion

Vocal folds fixation is typically assessed by visually evaluating videoendoscopic frames. This process is time consuming and requires an expert eye. To make the evaluation more objective, in this paper we compared four ML models to classify vocal cords motility into two classes: preserved motility and fixation. The best-performing model, XGB, proved to be a useful tool to investigate vocal cords motility in a more objective and reliable way. It is, in fact, able to distinguish between the two classes, which makes it a potential tool to support clinicians in the clinical practice.

References

1. Ligtens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Medical Image Analysis 2017;42:60-88.
2. Domingues I, Pereira G, Martins P, et al. Using deep learning techniques in medical imaging: a systematic review of applications on ct and pet. Artificial Intelligence Review 2020;53(6):4093-4160.
3. Fiorentino MC, Villani FP, Di Cosmo M, et al. A review on deep-learning algorithms for fetal ultrasound-image analysis. Medical Image Analysis 2022.
4. Mazurowski MA, Buda M, Saha A, Bashir MR. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on mri. Journal of Magnetic Resonance Imaging 2019;49(4):939-954.
5. Maier-Hein L, Eisenmann M, Sarikaya D, März T, Collins T, et al. Surgical data science–from concepts toward clinical translation. Medical Image Analysis 2022;76:102306.

6. Repici A, Badalamenti M, Maselli R, Correale L, Radaelli F, et al. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology 2020;159(2):512-520.

7. Paderno A, Gennaini A, Sordi C, Montenegro D, Lancini D, et al. Artificial intelligence in clinical endoscopy: insights in the field of videomics. Frontiers in Surgery 2022;9:361.

8. Menon JK, Nair RM, Priyanka S. Unilateral vocal fold paralysis: can laryngoscopy predict recovery? a prospective study. The Journal of Laryngology Otology 2014;128(12):1095-1104.

9. Voigt D, Döllinger M, Yang A, Eysholdt U, Lohscheller J. Automatic diagnosis of vocal fold paresis by employing phonovibrogram features and machine learning methods. Computer Methods and Programs in Biomedicine 2010;99(3):275-288.

10. Moccia S, De Momi E, Guarnaschelli M, Savazzi A, Laborai L, et al. Confident texture-based laminae tissue classification for early stage diagnosis support. Journal of Medical Imaging 2017;4(3):034502.

11. Andrade-Miranda G, Stylianou Y, Deliyski DD, Godino-Llorente N, Henrich N. Laryngeal image processing of vocal folds motion. Applied Sciences 2020;10(5).

12. Hamad AS, Haney MH, Lever TE, Bunyak F. Automated segmentation of the vocal folds in laryngeal endoscopy videos using deep convolutional regression networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2019;140-148.

13. Yousef AM, Deliyski DD, Zacharias SR, Naghibolhosseini M. Deep-learning-based representation of vocal fold dynamics in adductor spasmodic dysphonia during connected speech in high-speed videonoscopy. Journal of Voice 2022.

14. Yousef AM, Deliyski DD, Zacharias SR, de Alarcon A, Orlikoff RF, Naghibolhosseini M. A deep learning approach for quantifying vocal fold dynamics during connected speech using laryngeal high-speed videonoscopy. Journal of Speech, Language, and Hearing Research 2022;65(6):2098-2113.

15. Herbst CT, Unger H, Herzel H, Švec JG, Lohscheller J. "Phagein" analysis of vocal fold vibration documented with laryngeal high-speed video endoscopy. Journal of Voice 2016;30(6).

16. Lohscheller J. Towards evidence based diagnosis of voice disorders using phonovibrograms. 2009 2nd International Symposium on Applied Sciences in Biomedical and Communication Technologies 2009;1-4.

17. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. 2016;785-794.

18. Singh V, Gourisaria MK, Das H. Performance analysis of machine learning algorithms for prediction of liver disease. 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON) 2021;1-7.

19. Refat MA, Amin M, Kaushal M, Yasmin M, Islam K. A comparative analysis of early stage diabetes prediction using machine learning and deep learning approach. 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC) 2021;654-659.

20. Colleoni E, Moccia S, Du X, De Momi E, Stoyanov D. Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers. IEEE Robotics and Automation Letters 2019;4(3):2714-2721.

21. Moccia S, Migliorelli L, Carnielli V, Frontoni E. Preterm infants' pose estimation with spatio-temporal features. IEEE Transactions on Biomedical Engineering 2020;67(8):2370-2380.

7. Artificial intelligence for automatic detection and segmentation of nasal polyposis: a pilot study

Artificial intelligence models as a commodity: Standardized workflows for out-of-the-box training of pivotal vision models – The nasal polyps case-study.

Introduction

Chronic rhinosinusitis with nasal polyps (CRSwNP) is a specific subtype of chronic rhinosinusitis (CRS) usually related to an overactive Type-2 immune response. This association is of significant interest because it provides therapeutic targets for monoclonal antibodies, new generation molecules that have substantially improved the quality of life of patients with CRSwNP. The recent introduction of these new drugs has rekindled the interest of the scientific community towards CRSwNP, resulting in a remarkable increase in the related scientific production.[1]

However, understanding the full complexity of this pathological entity is still an ongoing area of research. Moreover, given the direct and indirect costs related to CRwNP, amongst the highest in healthcare,[1] many aspects of its practical management should be improved.

To date, the diagnosis and quantification of CRwNP is based on the integration by the physician of endoscopic, radiological, pathological, laboratory, anamnestic and subjective data, with the former and latter with a preponderant weight in the decision-making process.[1]

Since the costs of monoclonal therapy remain relatively high,[1] a precise and standardized selection process is mandatory, to propose it only to patients who will potentially benefit from the drugs. However, both endoscopy and quality of life assessment are characterized by a high grade of subjectivity. If on the symptoms reported by the patients the clinician can optimize only recording them as precisely as possible on validated questionaries, an effort to a sound endoscopic examination should be done in every case.

Nasal polyps usually show a typical greyish, translucent round aspect. However, the superficial aspect may vary significantly based on the presence of secretions, nasal congestion, highly active inflammation, and scar tissue due to previous surgeries. In these conditions, distinguishing normal mucosa from pathological tissue may be not obvious, thus jeopardizing diagnosis and quantification of the polyps. Moreover, an otolaryngology with the necessary expertise in rhinology may not be available in all facilities.

In this scenario, a computer aided decision support system might be beneficial for early diagnosis of nasal polyps. In recent years, among the most transformative applications of artificial intelligence (AI) is its integration with medical computer vision, which holds immense promise for revolutionizing healthcare diagnostics. By leveraging the power of AI algorithms and image analysis techniques, medical professionals can extract valuable insights from medical data with great accuracy, speed, and objectivity.[2,3] Medical computer vision refers to the intersection of computer vision techniques and medical imaging, encompassing the analysis, interpretation, and understanding of visual data obtained from various medical imaging modalities.[4]

In recent years several applications of image recognition guided by AI algorithms have been studied to face different challenges of otorhinolaryngological pathology,[5-8] especially regarding oncological and otological diseases. Regarding the management of CRSwNP, mainly radiological and histological image analysis tools have been exploited to extract information on prognosis, response to treatment and differential diagnosis,[7,8] whereas little has been done with the huge number of clinical videos depicting the objective situation of patient's nasal fossae collected in everyday practice.[8] For this reason, the aim of our pilot study is to verify the development feasibility of an AI based image analysis system, capable of identifying and delineate nasal polyps from nasal endoscopy videos, a task commonly referred as “segmentation”.[9,10]

Materials and Methods

Recorded nasal videoendoscopies of patients treated for CRwNP between 2019 and 2022 were retrospectively revised. Selection criteria comprised a pathology report positive for inflammatory nasal polyposis. All patients underwent high definition (HD) endoscopy (HD Video Rhino-laryngoscope Olympus ENF-VH, Olympus Medical System Corporation, Tokyo, Japan).

Nasal endoscopic images were retrospectively manually collected by deriving screenshots from videos in two different settings: endoscopic examination videos recorded during recruitment or follow-up of patients receiving biologic therapy for CRSwNP; intraoperative pre-intervention assessment in video of nasal endoscopic surgery performed for CRSwNP.

Images included in the dataset were required to have a sufficiently clear view of NPs of various degrees, without blood or crusting, at different levels of depth of the nasal fossa.

Three ENTs (CC, GT, CM) classified independently 1/3 of the screenshots each by manually segmenting each nasal polyp present in the image using the web application Roboflow (© 2023 Roboflow, Inc.).

After segmentation the following pre-processing was applied to each image:

- Auto-orientation of pixel data (with EXIF-orientation stripping)
- Resize to 640x640 (Stretch)

Moreover, the following augmentation was applied to create 3 versions of each source image:

- Randomly crop between 0 and 45 percent of the image

- Random rotation of between -23 and +23 degrees
- Random brightness adjustment of between -33 and +33 percent

Training of the AI Model

For the automated segmentation of nasal polyps, we employed the Ultralytics YOLOv8.0.28, a popular and efficient deep learning model designed for object detection and segmentation tasks. The YOLO model, specifically the YOLOv8s-seg variant, was trained using a dataset that was previously prepared as mentioned.

The training was conducted for 100 epochs with an image size of 640x640 pixels. This process was executed on a GPU with CUDA support, specifically the Tesla T4.

After the training was completed, the optimizer data was stripped from the saved model weights.

Training, validation, and test sets were split with a ratio of 80%:10%:10%. After the split, data augmentation was performed on the training set.

Results

A total of 342 images were screenshotted from the video nasoendoscopy of 52 patients affected by CRSwNP, 199 (58%) of images were of the right nasal fossa while the remaining 143 (42%) images were of the left nasal fossa. A mean of 3.8 and 2.5 screenshots were taken from the right and left nasal fossa respectively.

Training, validation, and test set were split with a ratio of 80%:10%:10%. After the split, data augmentation was performed on the training set.

A total of 816 images (92%) were used in the training set, 34 images (4%) in the validation set and 36 images (4%) in the testing set.

The resulting YOLOv8s-seg model consisted of 195 layers, had 11,779,987 parameters, and required 42.4 GFLOPs for operation.

When tested against the validation set, the model achieved satisfying results. For the box detection (detecting the location of nasal polyps), the model achieved a precision (P) of 0.91, recall (R) of 0.839, and mAP50 (mean average precision at 50% IoU) of 0.949. The mAP from 50% to 95% IoU was 0.675.

For the segmentation task (identifying and delineating the exact shape of nasal polyps), the model produced similar metrics with a precision (P) of 0.91, recall (R) of 0.839, mAP50 of 0.949, and mAP from 50% to 95% IoU of 0.679 (**Figure 1-6**).

In terms of processing speed, the model took an average of 0.2ms for pre-processing, 9.8ms for inference, 0.0ms for loss calculation, and 1.2ms for post-processing per image.

Discussion

Over the past few years, the medical imaging sector has seen a surge in the application of deep learning techniques, particularly convolutional neural networks (CNNs). Most research papers focusing on the application of AI in rhinology delve into diagnostics,[8,11] primarily aiming to automate the identification of anatomical and pathological features. Out of these, a significant number concentrate on radiology, including CT, MRI, and X-ray imaging.[8]

Within papers addressing AI applications for non- radiological diagnostic tools, only few focused on endoscopy.[8,11-17] The main spot of interest was the nasopharynx, and CNNs have been mainly used to develop tools for detecting nasopharyngeal carcinomas and grading adenoid hypertrophy.[11-13,15,16]

Girdler et al. developed and validated a CNN-based model to detect nasal polyposis and inverted papilloma from nasal endoscopic image. Despite the low overall accuracy rates, the authors showed the potential of deep learning for the diagnosis of nasal cavity masses,[14] paving the way for future studies.

In this context, we wanted to develop and test a highly accurate, AI-based, image analysis system, able to identify nasal polyps from endoscopies of patients affected by CRSwNP.

Leveraging an expansive dataset of 342 high quality images, we crafted an algorithm for segmentation of nasal polyps. The precision, recall rate, and mean average precision of 0.91, 0.839 and 0.949, respectively, emphasize its reliability in detecting nasal polyps with substantial accuracy. The integration of this tool into everyday clinical practice could lead to a significant improvement in patient management by potentially facilitating the development of personalized therapeutic strategies, including the targeted application of biological therapies. Furthermore, the integration of this AI-driven tool could mitigate the existing subjectivity in diagnosis, catalyzing a shift towards more data-driven and objective clinical management of CRSwNP.

However, it is pertinent to acknowledge the limitations of our current pilot project. The first is the lack of comparison with non-inflammatory polypoid lesions, given that our sample was a priori specifically aligned with CRSwNP. Despite having demonstrated the prowess of AI in segmenting nasal polyps with high precision, in line with Girdler et al.,[14] we should expand the spectrum of our study to encompass different nasal lesions, both benign and malignant. This would provide a tool with broader clinical applicability, able to orient diagnosis and consequently management of lesions with potentially opposite biological behaviors.

The rarity of nasal lesions different from inflammatory polyposis, in parallel with their superficial heterogeneity, are the main elements hampering this type of analysis. However, multicentric efforts in collecting large dataset of endoscopic images could overcome these limits.

Furthermore, to enhance the practical utility of this technological advancement, it is imperative to develop a reliable polyp quantification algorithm. This step will be pivotal to assimilate this tool into clinical practice, particularly concerning the indication of biological therapy, which nowadays is based on endoscopic (i.e., Nasal Polyp Score) and/or subjective (i.e., SNOT-22) quantification methods.[1] Moreover, it would help in monitoring in a reproducible way the response of patients to steroid and biologic therapy.

In this view, the research effort should be in the direction of segmenting not pathological nasal structures (i.e., turbinates, nasal septum, nasal floor), to allow for a quantification of polyps in respect to these structures. Moreover, the dataset should include images obtained from patients who underwent previous surgery. The large proportion of patients with recurrent polyposis have some grade of altered anatomy (i.e., cut or reshape of middle turbinate) and cannot be neglected in future analyses, even if this will make them more complex.

Conclusions

The pilot study represents a significant advancement in the application of AI to the management of CRSwNP. Leveraging a robust dataset, the developed algorithm yielded good precision and recall metrics, suggesting it could serve as a useful complement to existing diagnostic and treatment paradigms. The algorithm has the potential to assist clinicians in making more accurate and cost-effective treatment decisions, including the use of monoclonal therapies. However, limitations such as the lack of polyps grading and comparison with non-inflammatory polypoid lesions indicate room for further research. Overall, the study contributes to a growing body of evidence supporting the integration of data-driven methods in medical diagnostics and treatment planning.

References

1. Fokkens WJ, Viskens AS, Backer V, et al. EPOS/EUFOREA update on indication and evaluation of Biologics in Chronic Rhinosinusitis with Nasal Polyps 2023. Rhinology. 2023;61(3):194-202. doi: 10.4193/Rhin22.489.
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56. doi: 10.1038/s41591-018-0300-7. Epub 2019 Jan 7.
3. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal. 201742:60-88. doi: 10.1016/j.media.2017.07.005.
4. 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.
5. Wu Q, Wang X, Liang G, et al. Advances in Image-Based Artificial Intelligence in Otorhinolaryngology-Head and Neck Surgery: A Systematic Review. Otolaryngol Head Neck Surg. 2023. doi: 10.1002/ohn.391. Epub ahead of print.
6. Mäkitie AA, Alabi RO, Ng SP, et al. Artificial Intelligence in Head and Neck Cancer: A Systematic Review of Systematic Reviews. Adv Ther. 2023;40(8):3360-3380. doi: 10.1007/s12325-023-02527-9.

7. Bulfamante AM, Ferella F, Miller AM, et al. Artificial intelligence, machine learning, and deep learning in rhinology: a systematic review. Eur Arch Otorhinolaryngol. 2023;280(2):529-542. doi: 10.1007/s00405-022-07701-3.

8. Osie G, Darbari Kaul R, Alvarado R, et al. A Scoping Review of Artificial Intelligence Research in Rhinology. Am J Rhinol Allergy. 2023;37(4):438-448.

9. Paderno A, Gennarini F, Sordi A, et al. Artificial intelligence in clinical endoscopy: Insights in the field of videomics. Front Surg. 2022;9:933297. doi: 10.3389/fsurg.2022.933297.

10. Paderno A, Villani FP, Fior M, et al. Instance segmentation of upper aerodigestive tract cancer: site-specific outcomes. Acta Otorhinolaryngol Ital. 2023;43(4):283-290. doi: 10.14639/0392-100X-N2336.

11. Bi M, Zheng S, Li X, et al. MIB-ANet: A novel multi-scale deep network for nasal endoscopy-based adenoid hypertrophy grading. Front Med (Lausanne). 2023 Apr 14;10:1142261. doi: 10.3389/fmed.2023.1142261.

12. Li C, Jing B, Ke L, et al. Development and validation of an endoscopic images-based deep learning model for detection with nasopharyngeal malignancies. Cancer Commun (Lond). 2018;38(1):59. doi: 10.1186/s40880-018-0325-9.

13. Liu X, Sinha A, Ishii M, et al. Dense Depth Estimation in Monocular Endoscopy With Self-Supervised Learning Methods. IEEE Trans Med Imaging. 2020;39(5):1438-1447. doi: 10.1109/TMI.2019.2950936.

14. Girdler B, Moon H, Bae MR, et al. 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.

15. Shu C, Yan H, Zheng W, et al. Deep Learning-Guided Fiberoptic Raman Spectroscopy Enables Real-Time In Vivo Diagnosis and Assessment of Nasopharyngeal Carcinoma and Post-treatment Efficacy during Endoscopy. Anal Chem. 2021;93(31):10898-10906. doi: 10.1021/acs.analchem.1c01559.

16. 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.

17. Staartjes VE, Volokitin A, Regli L, et al. Machine Vision for Real-Time Intraoperative Anatomic Guidance: A Proof-of-Concept Study in Endoscopic Pituitary Surgery. Oper Neurosurg (Hagerstown). 2021;21(4):242-247. doi: 10.1093/ons/opab187.

8. Computer vision foundation models in medical images: Proof of concept in oropharyngeal cancer classification

Foundation models in clinical endoscopy

Introduction

Otolaryngology – Head and Neck Surgery has long been in pursuit of tools that can augment the diagnostic precision in the endoscopic evaluation of mucosal lesions. The complex anatomy of the upper aerodigestive tract (UADT) and the heterogeneity of lesions arising from this area require diagnostic methodologies that are both accurate and objective in the identification of early malignancies. While the mainstay of our diagnostic approaches has been rooted in subjective endoscopic evaluation and conventional radiologic techniques, the dynamic landscape of technological advancements offers a constantly growing range of innovative approaches.

Artificial intelligence (AI) has been heralded as a transformative force across various medical disciplines.[1,2] Within AI, computer vision stands out for its potential in medical imaging, endoscopy, and surgery. Here, the ability to 'teach' algorithms to interpret and infer from complex visual data offers unprecedented opportunities. One of the notable advancements in this space is the emergence of computer vision foundation models based on the vision transformer architecture [3], complex algorithms that can be trained by self-supervised approaches on extensive unlabeled datasets, leading to a broad understanding of visual patterns.[4] These algorithms are able to process input images and to create embeddings, multi-dimensional vectors that provide a summary of the visual pattern in each image. A notable example is represented by DINOv2,[5] a self-supervised vision transformer model by Meta AI that thanks to its robustness and adaptability is particularly enticing for applications where fine pattern recognition is required. However, current medical applications of image-based foundation models are still limited.

In our study, the overall endeavor is twofold: first, to rigorously assess the efficacy of DINOv2-derived embeddings (i.e., the quantitative features extracted from each image) in distinguishing between normal mucosa and neoplastic tissue, using the oropharynx as a representative model; and second, to position this exploration as a proof of concept within the broader narrative of foundation models in the evaluation of complex medical images.

Materials and Methods

This study was performed following the principles of the Declaration of Helsinki and dataset collection was approved by the Institutional Review Board, Ethics Committee of ASST Spedali Civili of Brescia, University of Brescia, Italy.

Objective

This study was designed to determine how effectively features extracted using DINOv2, when processed with a standard Support Vector Machine (SVM), can differentiate between neoplastic and normal tissue in the oropharynx.

Dataset Source and Image Processing

Images utilized in this research originated from a previously published dataset by the Department of Otolaryngology, Head and Neck Surgery, University of Brescia, Italy.

All patients were examined using Narrow Band Imaging (NBI). The examinations were carried out via transnasal videolaryngoscopy utilizing HD video rhino-laryngoscope models Olympus

ENF-VH, ENF-VQ, or ENF-V2 (Olympus Medical System Corporation, Tokyo, Japan). Alternatively, some examinations were conducted through transoral endoscopy, leveraging a 0° rigid telescope that was paired with an Evis Exera II HD camera. This camera was connected to an Evis Exera II CLV-180B/III CV-190 light source (Olympus Medical Systems Corporation, Tokyo, Japan).

From full endoscopic frames, multiple 300x300 pixel image patches were extracted by an otolaryngologist – head and neck surgeon with experience in UADT endoscopy, selecting the most representative areas of neoplastic tissue and regions with normal mucosa that were judged as at least 1 cm from the tumor margins. The image patches were clustered according to their tissue of origin (i.e., normal mucosa vs. neoplastic tissue).

To ensure uniformity and mitigate biases stemming from lighting conditions or equipment variations, images underwent a normalization process. This involved adjustments in luminosity, color, and contrast. OpenCV (Open Source Computer Vision Library), an open-source computer vision and machine learning software library was employed for histogram equalization across all color channels, followed by Z-score normalization. The final step involved rescaling the images to a [0, 255] range.

Embeddings Extraction

The DINOv2 model[5] was then employed with no fine-tuning to extract embeddings from each image patch. For this study, we utilized the "ViT-S/14 distilled" variant of DINOv2. This particular variant stands out as the smallest and less resource-intensive version of the model (21 million parameters), making it an ideal choice for applications where computational efficiency is paramount. This model had been pre-trained on an expansive dataset comprising 142 million images, which endowed it with a broad understanding of visual features across diverse domains, and had been distilled from the original model containing 1 billion parameters.

Classification Using SVM

The extracted embeddings, representing the processed images in the form of a tensor, were then used to train a Support Vector Machine (SVM) algorithm. Emphasis was placed on achieving high accuracy, making SVM an apt choice given its efficacy in high-dimensional spaces, as is typical with image embeddings.

To assess the model's discriminative capacity, the dataset was divided into an 80% training and 20% test split. A validation set was not required due to the absence of hyperparameters tuning. The Receiver Operating Characteristic (ROC) curve and the area under the curve (AUC) were calculated for the test set.

All image processing, feature extraction, training, and models testing was conducted using Google Colaboratory in a Python 3.10 environment.

Results

Dataset Overview

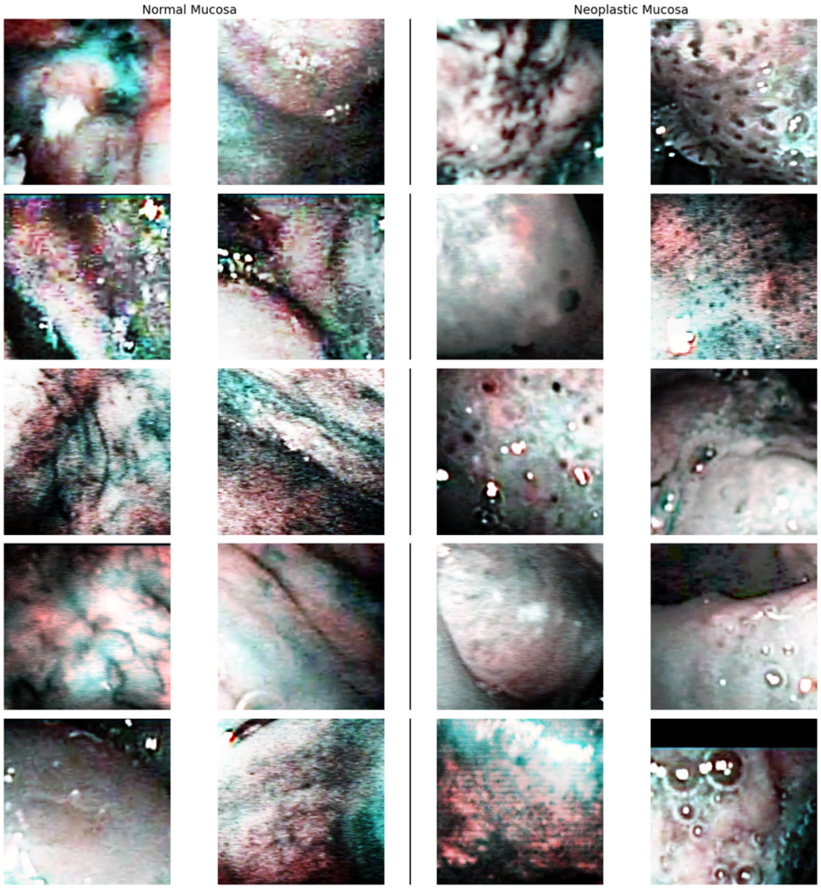
A quality selection process from the original dataset (45 videos and 116 frames) was performed to ensure the absence of biases, 7 endoscopic videos were excluded since they did not meet the quality criteria for the current task: sufficient quality, illumination, white balance, and magnification in order to allow the extraction of at least 1 patch from the image. The dataset used in this study comprised 38 endoscopic NBI videos of biopsy-proven oropharyngeal cancer, yielding a total of 88 frames.

From this dataset, a total of 327 image patches were extracted, ranging from 1 to 6 patches for each image, according to one of the following criteria:

- Presence of neoplastic tissue of a sufficient size to completely occupy the square.
- Presence of non-neoplastic mucosa of a sufficient size to completely occupy the square and with an adequate distance from the neoplastic tissue.

Of these, 205 patches were representative of neoplastic tissue, while the remaining 122 were of normal mucosa. A sample of their appearance after normalization is provided in Figure 8.1.

Figure 8.1. *Comparative visualization of normalized image patches from normal and neoplastic mucosa in oropharyngeal endoscopy. Displayed are representative images from both normal (left) and neoplastic (right) oropharyngeal mucosa samples (first ten images of each dataset). Each couple of rows corresponds to samples juxtaposing the morphological differences evident between the two conditions. The highly complex features and patterns discernible in these images underline the challenges inherent in the classification task.*



Diagnostic Performance

The process of embedding all image patches required a total computation time of 2 minutes and 55 seconds using an Intel(R) Xeon(R) CPU 2.20GHz or 6 seconds using an Nvidia Tesla T4 GPU. The resulting performance metrics for tumor classification were:

- Accuracy: 92%
- Precision: 89%
- Recall: 100%
- F1-Score: 94%

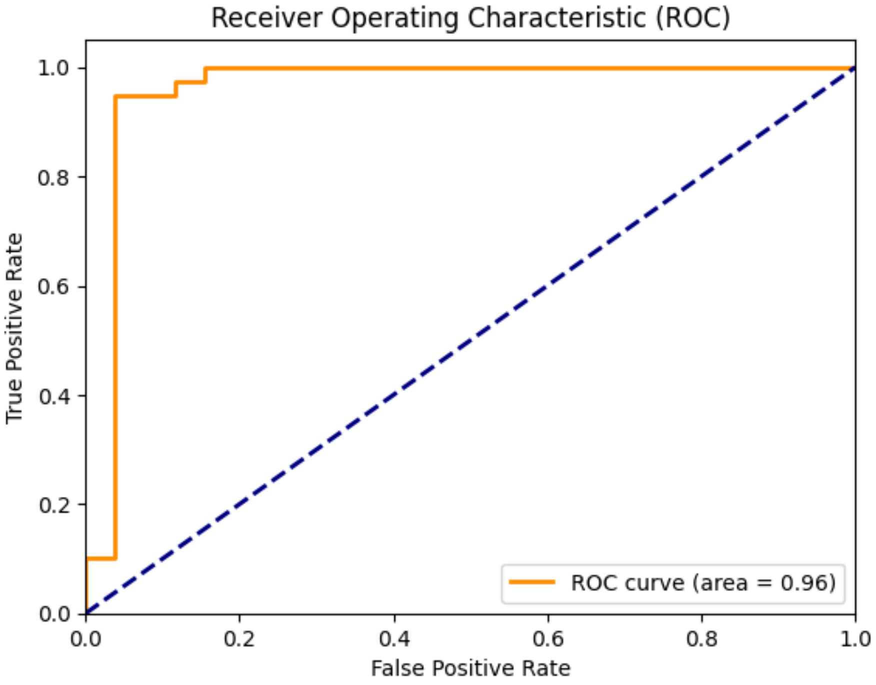
Detailed metrics are presented in Table 8.1.

Table 8.1. *Classification performance with an 80 to 20 split: detailed metrics. Legend: Avg, average.*

TRAINING SET METRICS				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	1.00	0.95	0.97	95
Tumor	0.97	1.00	0.99	165
Accuracy	0.98			260
Macro Avg	0.99	0.97	0.98	260
Weighted Avg	0.98	0.98	0.98	260
TEST SET METRICS				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	1.00	0.81	0.89	26
Tumor	0.89	1.00	0.94	39
Accuracy	0.92			65
Macro Avg	0.94	0.90	0.92	65
Weighted Avg	0.93	0.92	0.92	65

The Receiver Operating Characteristic (ROC) curve for the test set, displayed in Figure 8.2, yielded an area under the curve (AUC) of 0.96, highlighting the model's robust capability in tumor detection.

Figure 8.2. *Receiver Operating Characteristic (ROC) curve for the classification model.*



Discussion

Our study highlights the potential impact that computer vision foundation models can have on the field of medical image evaluation (i.e., Videomics),[6] particularly when considering endoscopy and surgery, visual inputs that, when compared with radiologic imaging, are closer to the natural images employed in the training of currently available solutions. These algorithms, based on the transformers architecture and optimized using vast datasets, have showcased their prowess in discerning intricate patterns and subtle variations, often challenging to detect through conventional means.[7]

In this specific application, the DINOv2 model was employed to extract and distill features present in the oropharyngeal mucosa evaluated by NBI endoscopy, considering both normal and neoplastic areas. Of note, DINOv2 did not receive any form of fine-tuning for this type of task. The extracted features, in the form of tensors, were then used to train and test a conventional machine learning algorithm (i.e., SVM). NBI was used as a form of image pre-processing in view of the limited number of image patches available, allowing to enhance the contrast of the submucosal microvascular pattern without the need for dedicated computational approaches.

The high average accuracy of 92% demonstrates the model's ability in differentiating between normal and neoplastic tissues. Furthermore, the high recall (100%) underscores its strength in correctly identifying tumors, a crucial metric in this specific clinical context. Finally, the model exhibited an AUC of 0.96 in the test set, underlining its outstanding performance in balancing sensitivity and specificity for tumor classification.

While this represents a selected proof of concept, the discourse around AI 'foundation models' in healthcare is particularly pertinent.[4] These models, often pre-trained on expansive datasets, serve as a robust base that can be directly used or fine-tuned for specific tasks. Their generalizability and adaptability hint at a future where AI tools might be tailored to diverse clinical scenarios with minimal additional training. This is particularly relevant in fields like Otolaryngology – Head and Neck Surgery where the prevalence of each specific condition might never be sufficient to provide extensive enough datasets for more complex tasks. The recent research by Zhou et al.[8] offers a compelling illustration of this paradigm. Zhou and colleagues introduced RETFound, an innovative foundation model for retinal images, designed to learn universally applicable representations from unlabeled retinal images, setting the stage for efficient model adaptation across diverse applications. RETFound, was trained on 1.6 million unlabeled retinal images through self-supervised learning, and it was subsequently fine-tuned for disease detection tasks using explicit labels. Remarkably, the adapted RETFound consistently surpassed several benchmark models in evaluating ocular conditions and signs of systemic pathologies. This approach represents a significant paradigm shift, offering a scalable solution to augment model performance while reducing the annotation responsibilities of experts.

The trajectory of medical computer vision in endoscopy, particularly when considering specialized fields like Otolaryngology, can be distilled into three distinct subsequent phases of development, each bringing forth its unique methodologies and innovations:

1. Handcrafted feature extraction and subsequent feature analysis by machine learning (ML) algorithms.
2. Direct application of convolutional neural networks (CNN) with automatic feature extraction.
3. Application of vision transformer-based foundation models with subsequent task-oriented optimization.

The first attempts were made using handcrafted feature extraction and traditional ML classifiers by focusing on the tissue texture (especially from NBI images)[9–11] or the microvascular structure.[12,13] Nevertheless, this approach was limited by the number and quality of the features selected by the investigators, and was less efficient compared to current deep learning (DL) models.

In the subsequent phase, CNN-classifiers became the architecture of choice to analyze endoscopic images and they have been widely investigated to classify different sites and diseases of the UADT, while the oropharynx remained an under-researched area due to the relative rarity of lesions and challenges in obtaining high-quality endoscopic images.[14–18,18–27] The advancement of these approaches has been significantly influenced by the size and quality of training datasets. Larger and more diverse datasets tended to lead to improved model accuracy, especially as the complexity of classification tasks increased. For example, Xiong et al.[16] utilized a dataset of 14,897 laryngeal images and achieved an 89.7% accuracy in binary classification, but the model's accuracy dipped to 77.3% when faced with a more complex 4-class categorization task. This decline in performance with increased classification complexity, even with such a vast dataset, underscores the challenges of differentiating between minimal variations in medical images, even with relatively large datasets. Contrastingly, Ren et al.'s[18] study stands out for its use of 24,667 laryngoscopy training images, and thanks to this impressive effort in data collection, their DL model was able to maintain an accuracy above 95% even across multiple classes.

However, not every task can depend on amassing vast amounts of data. Many analytical steps hinge on a comprehensive grasp of overarching visual concepts, anatomical structures, and common pathological traits. Foundation models based on vision transformers[3] provide this basic requirements and allow supervised training to be focused only on the domain-specific component of the classification problem, as proved by Zhou et al.[8] Our study was structured to serve as a benchmark for employing such an approach, with specific design choices to fully test this concept:

- We deliberately chose an anatomical site that remains under-researched (in terms of computer vision approaches) and comes with a limited pool of training images. Adjunctively, we refrained from any data augmentation.
- Our methodology was tailored to heavily rely on the intricate discernment of patterns and general structures. By selecting 300x300 pixel segments that exclusively displayed neoplastic or normal mucosa and normalizing for color and illumination, we constrained

the model to solely assess the mucosal and microvascular patterns, coupled with their spatial distribution.

- Leveraging the inherent capabilities of the DINOv2 model, we opted against any fine-tuning. Also, instead of appending a classification head to the model's backbone, we employed a straightforward SVM classifier. This decision was strategic, aiming to transparently and quantitatively gauge the efficiency and diagnostic potential of the embeddings derived from DINOv2 in feature extraction.

Even with these premises, we achieved an impressive average accuracy of 91% with solid validation and only using 327 image patches. This performance, when juxtaposed against the studies mentioned, underscores the efficacy of foundation models, even when operating with a limited dataset.

Conclusions

Our exploration into foundation models, particularly DINOv2, for endoscopic image analysis has yielded promising insights. While traditional methods have been predominantly localized, relying on extensive data and specific training, foundation models like DINOv2 offer a paradigm shift, leveraging vast pre-existing knowledge to discern intricate patterns in lesser-studied anatomical sites. Our study underscored the model's proficiency in detecting subtle variations, even with limited training data. The generalizability and adaptability of such models pave the way for more efficient and flexible medical image analyses, suggesting a transformative trajectory for AI-driven diagnostic tools in otolaryngology and endoscopy at large.

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7

2. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018;2(10):719-731. doi:10.1038/s41551-018-0305-z

3. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Published online June 3, 2021. Accessed September 15, 2023. http://arxiv.org/abs/2010.11929

4. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259-265. doi:10.1038/s41586-023-05881-4

5. Oquab M, Darcet T, Moutakanni T, et al. DINOv2: Learning Robust Visual Features without Supervision. Published online April 14, 2023. doi:10.48550/arXiv.2304.07193

6. 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

7. Patel P, Thakkar A. The upsurge of deep learning for computer vision applications. IJECE. 2020;10(1):538. doi:10.11591/ijece.v10i1.pp538-548

8. Zhou Y, Chia MA, Wagner SK, et al. A foundation model for generalizable disease detection from retinal images. Nature. Published online September 13, 2023. doi:10.1038/s41586-023-06555-x

9. 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

10. 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

11. Moccia S, De Momi E, Guarnaschelli M, et al. Confident texture-based laryngeal tissue classification for early stage diagnosis support. J Med Imaging (Bellingham). 2017;4(3):034502. doi:10.1117/1.JMI.4.3.034502

12. Irem Turkmen H, Elif Karsligil M, Kocak I. Classification of laryngeal disorders based on shape and vascular defects of vocal folds. Computers in Biology and Medicine. 2015;62:76-85. doi:10.1016/j.compbiomed.2015.02.001

13. Barbalata C, Mattos LS. Laryngeal Tumor Detection and Classification in Endoscopic Video. IEEE Journal of Biomedical and Health Informatics. 2016;20(1):322-332. doi:10.1109/JBHI.2014.2374975

14. Hanif U, Kezirian E, Kiær EK, Mignot E, Sorensen HBD, Jennum P. Upper Airway Classification in Sleep Endoscopy Examinations using Convolutional Recurrent Neural Networks. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). ; 2021:3957-3960. doi:10.1109/EMBC46164.2021.9630098

15. Cho WK, Lee YJ, Joo HA, et al. Diagnostic Accuracies of Laryngeal Diseases Using a Convolutional Neural Network-Based Image Classification System. The Laryngoscope. 2021;131(11):2558-2566. doi:10.1002/lary.29595

16. 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

17. Ay B, Turker C, Emre E, Ay K, Aydin G. Automated classification of nasal polyps in endoscopy video-frames using handcrafted and CNN features. Computers in Biology and Medicine. 2022;147:105725. doi:10.1016/j.compbiomed.2022.105725

18. Ren J, Jing X, Wang J, et al. Automatic Recognition of Laryngoscopic Images Using a Deep-Learning Technique. The Laryngoscope. 2020;130(11):E686-E693. doi:10.1002/lary.28539

19. 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. International Forum of Allergy & Rhinology. 2021;11(12):1637-1646. doi:10.1002/alr.22854

20. Cho WK, Choi SH. Comparison of Convolutional Neural Network Models for Determination of Vocal Fold Normality in Laryngoscopic Images. Journal of Voice. 2022;36(5):590-598. doi:10.1016/j.jvoice.2020.08.003

21. Dunham ME, Kong KA, McWhorter AJ, Adkins LK. Optical Biopsy: Automated Classification of Airway Endoscopic Findings Using a Convolutional Neural Network. The Laryngoscope. 2022;132(S4):S1-S8. doi:10.1002/lary.28708

22. Heo J, Lim JH, Lee HR, et al. Deep learning model for tongue cancer diagnosis using endoscopic images. Sci Rep. 2022;12(1):6281. doi:10.1038/s41598-022-10287-9

23. He Y, Cheng Y, Huang Z, et al. A deep convolutional neural network-based method for laryngeal squamous cell carcinoma diagnosis. Annals of Translational Medicine. 2021;9(24):1797-1797. doi:10.21037/atm-21-6458

24. Xu J, Wang J, Bian X, et al. Deep Learning for nasopharyngeal Carcinoma Identification Using Both White Light and Narrow-Band Imaging Endoscopy. The Laryngoscope. 2022;132(5):999-1007. doi:10.1002/lary.29894

25. Ye G, Du C, Lin T, Yan Y, Jiang J. Deep Learning for Laryngopharyngeal Reflux Diagnosis. Applied Sciences. 2021;11(11):4753. doi:10.3390/app11114753

26. 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 Recognition Letters. 2021;150:207-213. doi:10.1016/j.patrec.2021.06.034

27. 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. Computers in Biology and Medicine. 2022;144:105339. doi:10.1016/j.compbiomed.2022.105339

9. A comparison between manual feature extraction and transformer-based embeddings in oropharyngeal cancer classification: Evaluating the progress of computer vision

Foundation vision models for automatic feature extraction: New approaches to reduce subjectivity.

Introduction

As presented in the previous section, pre-trained foundation models based on vision transformers have the potential to extract reliable feature representations from Narrow Band Imaging (NBI) frames depicting normal or neoplastic oropharyngeal mucosa. This is a new and interesting concept that can lead to the development of large scale foundation models finetuned on endoscopic images. These models can potentially improve feature extraction for a wide series of downstream tasks that will require only a limited amount of training frames to reach peak performance, up to the concept of "few-shot learning".[1]

NBI conceptually represents a type of pre-processing image enhancement that focuses the visualization on certain wavelengths of the visible light to highlight the mucosal and submucosal microvasculature.[2] This approach has been developed to help the "human expert" in the endoscopic diagnosis of undetermined lesions, a concept termed "optical biopsy".[3] However, NBI might also limit the amount of available information in the frame by excluding some wavelengths that are present in white light (WL) endoscopy and this should be taken into consideration when assessing the optimal input data for each algorithm.

As mentioned in the previous sections, Mascharak et al.[4] described one of the first approaches to automatic classification of normal or neoplastic mucosa in the upper aerodigestive tract, specifically the oropharynx. The authors employed manual features extraction and subsequent analysis using a Bayes naïve classifier and reported the different diagnostic outcomes for WL and NBI frames. Manual feature extraction and selection represents the first technical approach to computer vision tasks, but it is technically challenging, time consuming, and it is burdened by some degree of subjectivity. On the other hand, large vision models are capable of extracting image features in the form of embeddings, that can be subsequently used for classification tasks.

Here we perform a secondary analysis on Mascharak et al.'s[4] dataset using automatic feature extraction with a foundation vision model, DINOv2.[5]

The aim of this study is to answer to the following questions:

- What is the evolution in terms of performance from the previous approach of manual feature extraction to automatic feature extraction using a vision foundation model?
- Are image enhancing techniques still relevant when performing automatic feature extraction (i.e., what is the performance gap between WL and NBI evaluation)?

Materials and Methods

The study objective is to compare the diagnostic performance of different feature extraction approaches for image classification: manual feature extraction and automatic feature extraction.

Data Acquisition and Preprocessing

The dataset comprised of 144 WL and 168 NBI image patches, sourced from the dataset made available by Mascharak et al.[4] Each image was labeled according to the tissue type it represented, falling into one of two categories: Normal or Tumor. The images were stored in PNG format, ensuring high-quality and lossless data retention.

The images underwent a series of preprocessing steps to make them compatible with the DINOv2 model. This included conversion to tensor format, resizing to a 244x244 resolution, center cropping to a final size of 224x224, and normalization of pixel values. The preprocessing pipeline ensured that the images were in the appropriate format and dimensionality for feature extraction.

Feature Extraction

The DINOv2 model, specifically the "dinov2_vits14" variant, was employed for feature extraction. The model was loaded using PyTorch's hub functionality and set to evaluation mode to disable any training-specific operations. The extracted features were then saved to a JSON file for subsequent analysis.

Classifier Training and Evaluation

A Support Vector Machine (SVM) classifier with a radial basis function kernel was utilized for the classification task. The gamma parameter was set to 'scale', and probability estimates were enabled. The feature vectors extracted by DINOv2, along with their corresponding labels, formed the input data for the classifier.

To assess the model's performance and generalizability, a train-validation split was performed, allocating 80% of the data for training and 20% for test. The classification report revealed precision, recall, and F1-score values for both classes on the training and validation sets, providing a comprehensive view of the model's performance. A 5-fold cross-validation was performed to evaluate the model’s stability and performance across different subsets of the data.

The Receiver Operating Characteristic (ROC) curve was plotted for the test set to visualize the trade-off between the True Positive Rate and False Positive Rate at various threshold settings. The area under the ROC curve (AUC) was calculated, providing a single scalar value to summarize the classifier's performance.

Results

1. Performance on WL images:

The SVM classifier, when applied to white light images, demonstrated a satisfying ability to differentiate between Normal and Tumor tissues.

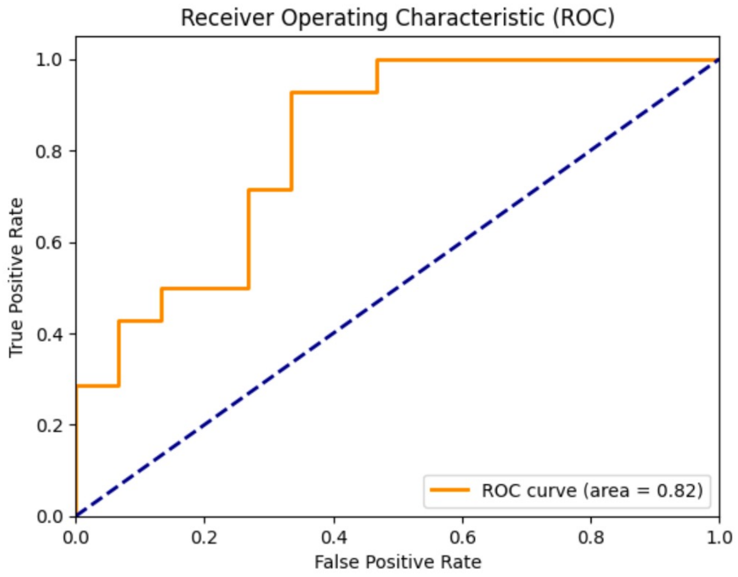
- **Test set accuracy**: The accuracy on the test set was recorded at 72.4%, highlighting a competent model performance.

<u>In-depth test set metrics:</u>

- **Precision**: The model exhibited a high precision of 89% for Normal tissues, suggesting that when it predicted a tissue as Normal, it was correct 89% of the time. However, the precision for Tumor predictions was lower, at 65%.
- **Recall**: There was a noticeable imbalance in recall, with the model demonstrating a recall of 93% for Tumor tissues, minimizing the false negatives, which is crucial in medical diagnoses. However, the recall for Normal tissues was significantly lower at 53%.
- **F1-score**: The F1-score, a balance between precision and recall, was 67% for Normal and 76% for Tumor tissues.
- **Support**: The model was tested on 15 Normal and 14 Tumor samples.

- **ROC AUC**: 0.73 (Figure 9.1)

Figure 9.1. *ROC curve for white light images.*



<u>Cross-validation test metrics</u>:

- **Test Accuracy**: The average test accuracy across the 5 folds was 72%, consistent with the single split test accuracy.
- **Precision**: 72%
- **Recall (Sensitivity)**: 82%
- **F1-score**: 76%
- **Specificity**: 63%

2. Performance on NBI images:

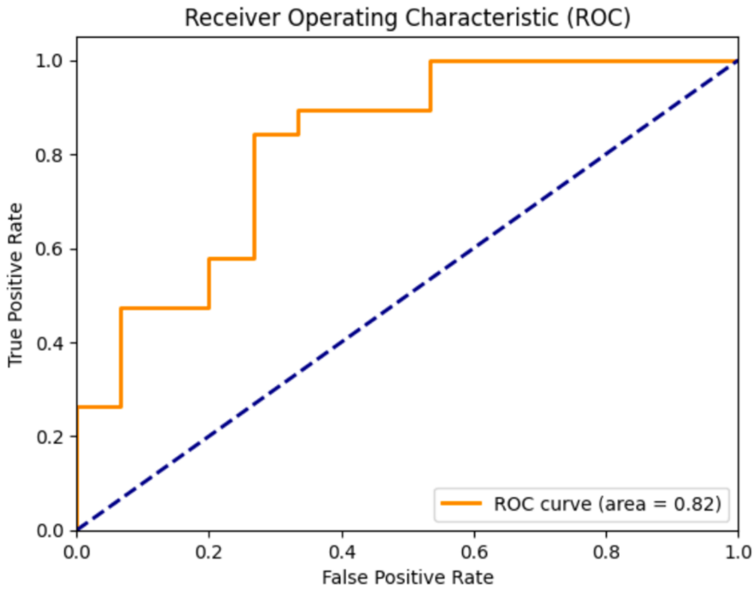
The SVM classifier's application to NBI images painted a slightly different picture, displaying a balanced and strong performance across metrics.

- **Test set accuracy**: The model achieved an accuracy of 79.4% on the test set.

In-depth test set metrics:

- **Precision**: The model achieved a precision of 79% for Normal and 80% for Tumor.
- **Recall**: The recall was also balanced, with 73% for Normal and 84% for Tumor, showcasing the model's ability to correctly identify both tissue types, though with slightly better performance in minimizing false negatives for Tumor tissues.
- **F1-score**: F1-scores of 76% and 82% for Normal and Tumor respectively, reflect the balanced precision and recall.
- **ROC AUC**: 0.82

Figure 9.2. *ROC curve for NBI images.*



Cross-validation test metrics:

- **Test Accuracy**: A consistent average test accuracy of 79% was noted across the folds.
- **Precision**: 78%
- **Recall (Sensitivity)**: 75%
- **F1-score**: 75%
- **Specificity**: 82%

Discussion

With WL images, the model showcased its strength in identifying Tumor tissues, but it also revealed a conservative approach towards predicting Normal tissues, leading to a lower recall for Normal and lower precision for Tumor predictions. The good ROC AUC suggests promising discriminative power, still leaving some room for improvement. With NBI images, the model displayed a more balanced and slightly superior performance compared to the white light model. The precision and recall were notably more balanced across the two classes, and the high sensitivity and specificity indicate the model's potential in medical diagnostics.

Both models, trained on WL and NBI images respectively, demonstrated promising capabilities in distinguishing between Normal and Tumor tissues. The high recall for Tumor tissues across both models is a highlight, emphasizing their potential in clinical settings where the cost of missing a Tumor tissue (false negative) is high. Nonetheless, the observed discrepancies in precision and recall, alongside the evident drop in performance from training to testing, signal the need for further optimization and validation. Enhancing model generalization and precision is paramount to paving the way for more reliable and effective applications in medical image analysis.

As delineated in Table 9.1, the comparative analysis between two distinctive methodologies— one leveraging the nuances of automatic feature extraction with Support Vector Machines (SVM) and the other rooted in the traditional paradigms of manual feature selection with naïve Bayes— reveals multifaceted insights into endoscopic computer vision.

Table 9.1*. Diagnostic performance using the two different methodologies: comparison.*

Metrics	Automatic (WL)	Automatic (NBI)	Manual (NBI, Color)	Manual (NBI, Color + Texture)	Manual (WL, Color)	Manual (WL, Color + Texture)
Accuracy	72.4%	79.4%	65.9%	70.0%	52.3%	52.4%
Recall	93%	84%	66.8%	70.9%	44.8%	47.0%
Specificity	63%	82%	64.9%	68.9%	59.9%	57.9%
ROC AUC	73%	82%	72.3%	80.1%	54.6%	52.3%

First, a commonality emerges: both methodologies manifest superior diagnostic efficacy when utilizing NBI over traditional WL imaging. However, delving deeper, it is possible to observe a divergence in the performance metrics of the two approaches. For the naïve Bayes approach, which necessitates meticulous manual feature selection, the reliance on the robustness of NBI is paramount. This is evidenced by its consistent diagnostic performance metrics across the board. The NBI, in this context, acts as a compensatory mechanism, bridging the potential gaps and limitations intrinsic to manual feature discernment.

Conversely, evaluating the automatic feature extraction approach coupled with SVM classification, an interesting peculiarity unfolds. While NBI's superiority still holds, the disparity in performance between NBI and WL imaging diminishes. This suggests that automatic feature extraction using a large vision model is capable of gleaning critical diagnostic insights even from conventional WL images. The prowess of foundation model-based embeddings, in essence, manages to narrow the gap between the two imaging modalities.

This observation opens a broader discourse on the evolving synergy between imaging technologies and machine learning in contemporary medical diagnostics. As machine learning models continue their ascent towards higher sophistication, their dependency on external imaging enhancements like NBI could potentially attenuate. Instead, the future might see a paradigm where these advanced algorithms can autonomously parse and interpret even raw, unenhanced visual data, translating them into diagnostic insights with precision.

Such a trajectory also raises pertinent questions about the future of medical diagnostics. Could we be on the cusp of a revolution where the diagnostic process is overwhelmingly automated, powered by machine learning algorithms that can seamlessly interpret vast spectrums of visual data? And as these models evolve, will the need for specialized imaging modalities, like NBI, become obsolete, or will they coexist, complementing the ever-evolving landscape of automated diagnostics? This discourse merely scratches the surface, serving as an invitation to delve deeper and chart the unexplored terrains of this emerging synergy.

References

1. Chen R, Liu Y, Kong L, et al. Towards Label-free Scene Understanding by Vision Foundation Models. 2023; ArXiv, abs/2306.03899.
2. de Kleijn BJ, Heldens GTN, Herruer JM, et al. Intraoperative Imaging Techniques to Improve Surgical Resection Margins of Oropharyngeal Squamous Cell Cancer: A Comprehensive Review of Current Literature. Cancers (Basel). 2023;15(3):896. doi: 10.3390/cancers15030896

3. Halicek M, Little JV, Wang X, Chen AY, Fei B. Optical biopsy of head and neck cancer using hyperspectral imaging and convolutional neural networks. J Biomed Opt. 2019;24(3):1-9. doi: 10.1117/1.JBO.24.3.036007

4. 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.

5. Oquab M, Darcet T, Moutakanni T, et al. Dinov2: Learning robust visual features without supervision. 2023; arXiv preprint arXiv:2304.07193

10. ENDO-CLOUD: Enhanced Network for Deep learning-Oriented Classification and Leveraging of Optimized Upper aerodigestive tract Data: A cloud-based annotation system with informative frame classification in videolaryngoscopy

Working at scale: MLOps platforms to allow large-scale training and deployment of computer vision AI models.

Introduction

In the rapidly developing field of contemporary laryngeal endoscopy there are multiple challenges, both from the scientific (medical), as well as technological (informatics) perspectives.[1] Machine learning (ML) approaches, or more specifically – deep learning (DL), constitute a core methodic for applying artificial intelligence (AI) concepts to biomedical problems, such as videoendoscopic image segmentation and classification.[2-5] Utilizing AI methods for bioimaging is of course not only restricted to the 2D image pattern analysis;[6-8] however, the emphasis in the current section is focused predominately on the videolaryngoscopic image processing for laryngeal pathology detection.[9,10]

In order to further improve diagnostic accuracy of the aerodigestive tract mucosal lesions, as well as to provide additional information for specialists during training, AI models can be developed and utilized. For instance, there are several support systems for (semi-)automatic lesions segmentation of videoendoscopic frames being developed.[2,11,12] Importantly, such systems can also provide statistical confidence intervals of the model decision, in terms of assigning the segment to a given pathology class. Finally, such system can be also utilized for radiological images (i.e., source DICOM images can be converted to PNG raster graphics on the per-slice basis, prior to being introduced to the system).[13]

The initial, critical stage in the development of AI models for laryngeal videoendoscopy involves gathering video data captured during the examination of a patient's aerodigestive tract. Despite the global prevalence of videoendoscopy as a standard hospital procedure, a single site can only yield a restricted quantity of individual patient data. This inherent limitation underscores the necessity for diversified and substantial data collection, ensuring the robustness and generalizability of the developed AI models. For specialized issues, there are a few hundred patients per year at a given hospital or clinic. From the technical standpoint, it means that the number of samples available for training and testing ML model is relatively small at a single site, hence the necessity of a multi-centric cooperation to provide thousands of annotated images of different pathologies.

In parallel, the role of advanced endoscopic techniques, such as Narrow Band Imaging (NBI), has become increasingly prominent in improving diagnostic accuracy.[14-20]

Building upon earlier studies,[21] our initial results affirm the notion that the right tools and methods can meet the hefty data demands necessary to train and test ML models. Moreover, these strategies can make the tasks of collecting and labelling data substantially more manageable. However, in order to achieve this goal, an advanced system allowing an effective multi-centric cooperation is required.

Past efforts to develop segmentation systems and DL training workflows for aerodigestive endoscopy imaging have largely been targeted at desktop platforms.[12,22-24] However, often several different desktop applications had to be used, in order to: (1) extract the frames from the video, (2) segment the image (e.g., LabelMe – http://labelme.csail.mit.edu, VGG Image Annotator (VIA) –

https://www.robots.ox.ac.uk/~vgg/software/via/; labelimg – https://www.v7labs.com/blog/best-image- annotation-tools#labelimg; or RIL-Contour[25]), and, finally, (3) train the model. A significant drawback of this approach lies in the necessary involvement of each participating specialist in the image segmentation process. They must independently install and configure all requisite software on their personal laptops. Only upon completion of this setup can they commence the extraction and segmentation of frames. These frames are subsequently shared with the AI specialists through an additional application or a web-based file transfer platform (such as Google Drive or WeTransfer). This process introduces a layer of complexity and potential inefficiency.

Although, in general, multi-purpose segmentation systems are pretty common,[13,26,27] there is a lack of such solutions dedicated to 2D videoendoscopic frames. Moreover, from the medical point of view, the existing workflows, as well as the available datasets, are often restricted to a single anatomical structure of the larynx, e.g., the glottis,[28] while with the use of the system herein described, it is possible, if needed, to perform the segmentation of all laryngeal structures and/or all tissue types.

Materials and Methods

IBSU-1432 dataset

The dataset of 1432 frames used in this study comprised three subsets: (1) 720 frames shared as part of Moccia et al.;[24] (2) 237 frames from the videos recorded in the Department of Otorhinolaryngology – Head and Neck Surgery, University of Brescia, Italy; and (3) 475 frames extracted from the videos acquired in the Clinic of Otolaryngology, Head and Neck Surgery of the Poznan University of Medical Sciences, Poland. The rationale behind the extension of the initial dataset of 720 frames was to provide a more robust model for the selection of informative frames for the system described in the following sections. Moreover, in the extended 1432-frame dataset, there were not only NBI frames (720 from Moccia at al.,[24] 123 from Brescia, and 475 from Poznan), but also 99 WL frames (from the Brescia dataset). In terms of frame quality, four classes were distinguished – after Moccia et al.[24] in: informative frames (436), blurred frames (383), frames with saliva/specular reflections (321), and underexposed frames (292).

The model created with 1432 frames was quantitatively validated, according to the practices routinely utilized in ML model workflows,[29-39] as well as qualitatively, as part of the annotation practice in the ENDO-CLOUD system by clinical experts. In details, two aspects of the quantitative classifier accuracy were analyzed.

We analyzed our system's performance using two distinct metrics. The first assessed the algorithm's efficiency in differentiating among the four categories of frames under review: informative, blurred, frames with saliva/light reflections, and underexposed frames. This measure, termed 'general accuracy,' gauges the algorithm's overall classification ability. However, in this context, not all classification errors carry equal weight. Specifically, the system's primary task is to distinguish informative frames from non-informative ones, rendering errors among various types of non-informative frames less impactful. Recognizing this practical nuance, we incorporated a second metric into our analysis: the accuracy of a binary classifier that focuses solely on separating informative from non-informative frames. This measure provides a more targeted evaluation of the system's performance in its most critical function.

EfficientNet's open-source code with examples and pre-trained models helped implement the network quickly and transparently in our tool. This allowed us to focus more on the proper adaptation of the dataset and the appropriate configuration of the learning process.

In the current study the EfficientNetB7 DL model was utilized, which is a convolutional neural network architecture and scaling method that uniformly scales all dimensions of depth, width, and resolution using a compound coefficient.[31] The structure of this architecture includes five different modules and includes seven blocks for feature extraction, merging, scaling, multiplication, etc., as well as input and output modules. Because it its relatively high performance in image feature detection, as compared to other commonly utilized convolutional neural networks,[31] this network can be successfully applied to the frame quality classification. Additionally, EfficientNetB7 is an open-source solution, available with a pre-trained model, which allows for faster adaptation of this architecture to specific computer vision problems, such as the ones being the subject of the current article. The architecture was implemented using Python programming language (v. 3.8), with a TensorFlow framework, and several Python modules, commonly utilized in ML applications, such as: pandas, NumPy, and scikit-learn.

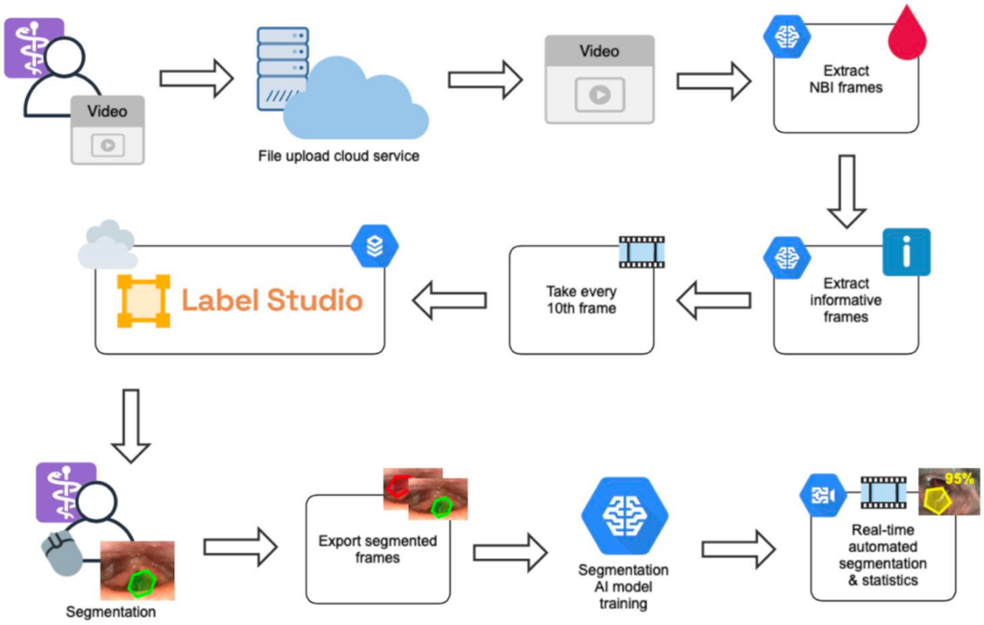
The accuracy metrics for the model were obtained with a 10-fold cross-validation procedure, i.e., when using the batch size of 4, the first 9 folds contained 36 batches, which means 144 images per fold, which gives 1296 frames for the training (9*144=1296), while the last fold consisted of 34 batches, i.e., 136 example frames for validation (34*4=136). In total, for each fold, there were 1296 frames for training, and 136 frames for testing, which gives 1432 frames in the whole dataset. The following metrics will be reported: accuracy score (for the unbalanced classes), precision, recall, F1 score, and G-mean.

ENDO-CLOUD – Segmentation powered by the cloud

The ENDO-CLOUD system described in the current article is fully web-based, meaning that the segmentation process can be easily shared between multiple users and user roles. For example: (1)

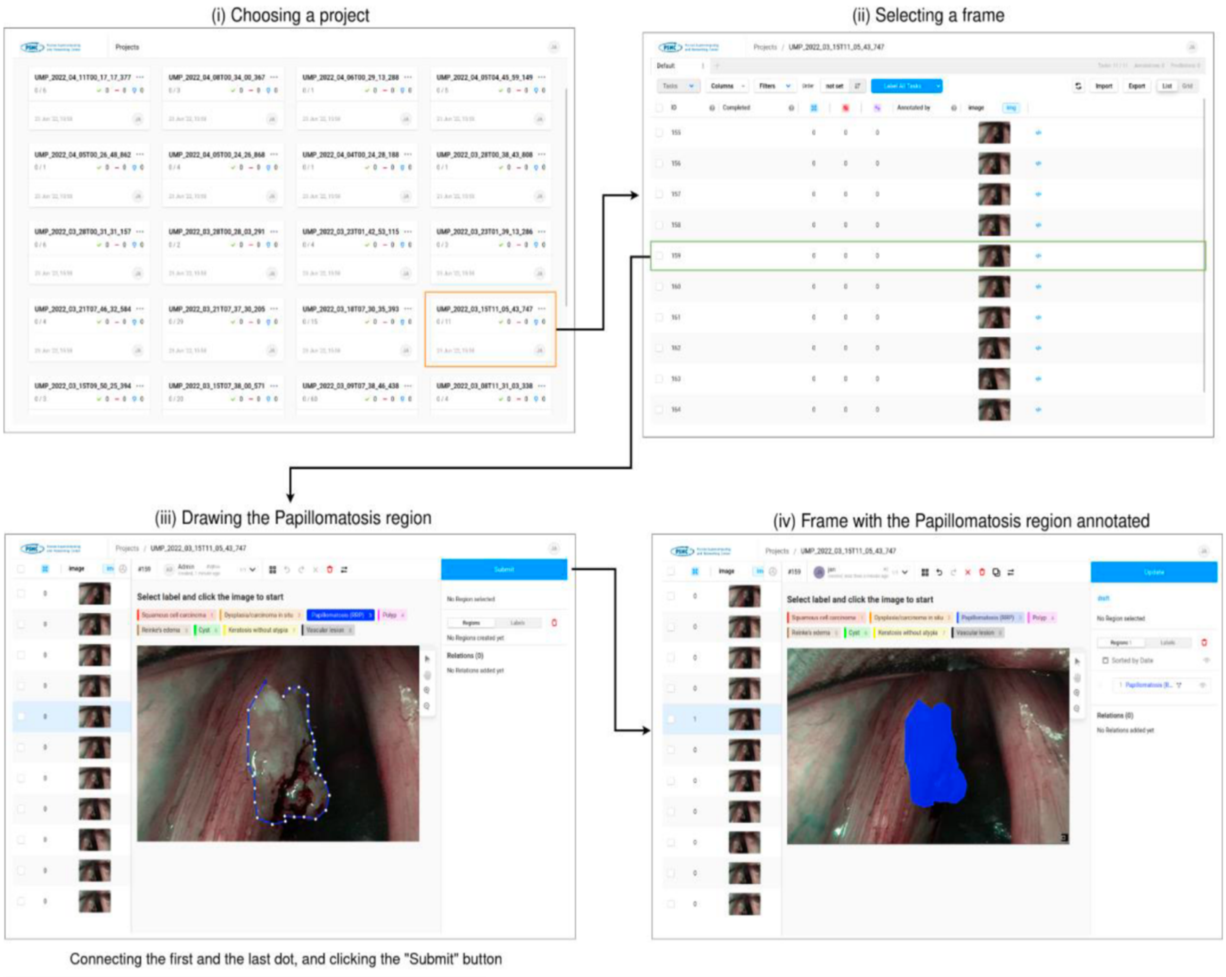
medical experts can access data resources and perform segmentation, (2) DL specialists can use the annotated data to develop AI models, and (3) system administrators can be in charge of the security and privacy issues, and maintain the remote resources. From the perspective of an end-user who is performing the segmentation (medical expert), the system is available via browser – as a web application, which means that no desktop applications has to be installed on the user computer, and no further files have to be shared with the partners. All data management and data access functionalities are provided by the system. The segmentation process is based on the open source Label Studio web application, originally developed as a general-purpose natural scene segmentation toolkit (see https://labelstud.io – Apache 2.0 License)[29,30], that was adapted for this specific use case. The contribution of the authors of the current article was to specifically adapt the base Label Studio application source code to medical image segmentation. The complete ENDO-CLOUD system is more than the segmentation tool alone (Figure 10.1): it includes the whole process of semi-automatic video frame extraction, combined with segmentation functionalities provided by Label Studio, which allows working with medical materials such as videolaryngoscopies.

Figure 10.1. *Workflow diagram for uploading and performing segmentation of the NBI images, followed by training and usage of a DL model.*


Video
File upload cloud service
Video
Extract NBI frames
Extract informative frames
Take every 10th frame
Label Studio
Segmentation
Export segmented frames
Segmentation AI model training
95%
Real-time automated segmentation & statistics

In the ENDO-CLOUD implementation, the Label Studio has been stripped of any unnecessary graphics interface elements, thus making it more straightforward and user friendly. Importantly, the predefined labels for particular laryngeal pathologies have been created using Label Studio templates, so that the medical experts can start labeling right away – as soon as they start using an instance of the system. In the Label Studio implementation terminology, each patient's videoendoscopic data is treated as a separate "project". For the particular patient's project, all the informative NBI frames are stored on the server, and are accessible through a database, with a convenient user interface (Figure 10.2). The use of separate projects for each patient allows an efficient data organization, and the workflow is characterized with better usability. After the original videos are anonymized and uploaded to the server, projects for particular patients are generated by a custom script that transforms whole-length medical videos into images (separate frames), which are stored on the server-side in the standardized directory structure.

***Figure 10.2**. Segmentation process with web-based application – the system is based on the Label Studio open source software (https://labelstud.io). Phases (i)-(iv) go from patient (case) selection (i), through deciding which frame to annotate (ii), up to performing (iii), and finishing (iv) the segmentation on a chosen frame.*


(i) Choosing a project
Projects
(ii) Selecting a frame
Projects / UMP_2022_03_15T11_05_43_747
(iii) Drawing the Papillomatosis region
(iv) Frame with the Papillomatosis region annotated
Select label and click the image to start
Squamous cell carcinoma
Dysplasia/carcinoma in situ
Papillomatosis (RRP)
Polyp
Reinke's edema
Cyst
Keratosis without atypia
Vascular lesion
Submit
Update

Connecting the first and the last dot, and clicking the "Submit" button

The selection of the final frames (i.e., the ones to be segmented), as well as the segmentation process itself, is facilitated in the ENDO-CLOUD system by a convenient user interface running in the web browser. The user of the system (medical expert) can choose a patient from the list of projects – each representing data from a single patient (Figure 10.2, phase “(i)”). For each of the project's fields, the annotation progress status is presented, to provide an overview of how many frames have already been segmented for a given subject. After selecting a patient’s project, a list of videoendoscopic frames for the selected patient is displayed (Figure 10.2, “(ii)”). Subsequently, the user can select a frame of interest and begin the annotation procedure. The initial step of the segmentation is choosing a type of pathology to be segmented – by selecting a label (button) with a given type of disease. Then, by clicking on the image, the user creates an outline of the mucosal abnormality (Figure 10.2, “(iii)”). When the user clicks the first “dot” (i.e., user “closes” the segment delineation), a mask/region representing a given laryngeal lesion is created (Figure 10.2, “(iv)”). In order to save the created region, the "Submit" button in the top right corner of the screen has to be clicked.

The Label Studio software also provides a feature for recording the time spent on annotating each image (frame). This allows the system to keep track of how much time it took to perform a given annotation, thus providing an estimate of how long all annotations for all project tasks (videoframes) may take. It can help to organize the work of a team segmenting the images across multiple sites. It allows the team to focus on high-priority tasks and finish them according to the schedule. Time spent on annotating particular frames is stored in the database, and can be exported with the annotation dataset in the following formats: JSON, JSON-MIN, CSV, and TSV (as a “lead_time” field, in seconds). Additionally, all of the Label Studio annotations can be exported to formats commonly utilized in training deep neural network models (such as deep convolutional neural networks). The available export formats for annotations and images are: COCO, CONLL2003, Pascal VOC XML, and YOLO.

The particular types of laryngeal diseases that are the default labels for the segmented delineations on the frames from the WL/NBI videos are: (1) squamous cell carcinoma, (2) dysplasia/carcinoma in situ, (3) recurrent respiratory papillomatosis, (4) polyp, (5) Reinke’s edema, (6) cyst, (7) keratosis without atypia, and (8) vascular lesion (for the visualization of the labels, see Figure 10.2 (iii) and (iv)). Importantly, once defined, the same, constant set of labels can be used by all the different research groups or clinics who are using a specific version of the web-based ENDO-CLOUD system. The modification of this set of pathology labels is possible (including renaming labels, adding new labels, removing them, etc.), and the changes are automatically applied to all the images that have already been segmented, as well as to the new annotation views. Interestingly, it is also possible to adjust the system, so that it can be used to segment not only larynx pathologies, but also the healthy anatomical sites.

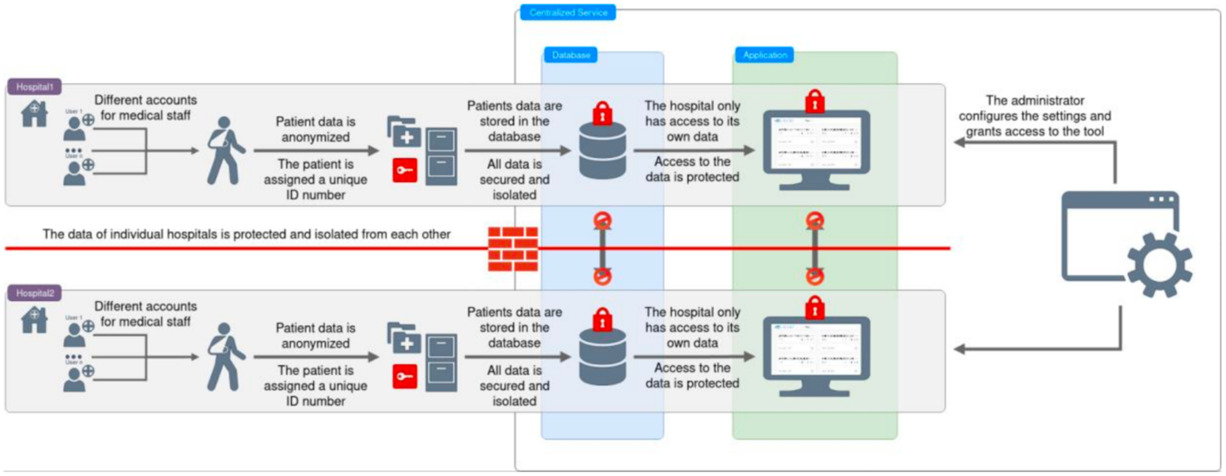
Data sharing and process automation

Data privacy and security have the highest priority when developing systems that are being used for storing and processing medical data. In the ENDO-CLOUD system, the anonymity of the patient is provided at the stage of data acquisition, where the patient is assigned a unique ID number (identifying him/her in the hospital electronic database, or hospital information system). Hence, no name, surname, or any identity-related information is required as the input to the cloud-based segmentation system. Furthermore, in the ENDO-CLOUD system, stable, robust and secure software components are used (such as Ubuntu operating system and Django web framework) in order to maximize the security of the system, thus reducing the risk of data being intercepted by an untrusted third party (thus minimizing the risk of data leaks).

The key software component of the ENDO-CLOUD system, which is Label Studio, provides on its own high security of the data that are being processed. It is also possible that each institution can have its own data storage, and the only common aspect would be providing key functionalities by the application. Additionally, each medical or research facility may have independent accounts for each member of the medical staff, providing better control over data management and segmentation process, also allowing collaboration between experts and cross-validation of the segmentation performed. An example of such a collaboration can be that different users from the same hospital (or clinic/research center) are allowed to access data from the same patient (Figure 10.3).

The workflow with the ENDO-CLOUD segmentation system is as follows: medical specialists provide patient data through a secure, dedicated channel. All the data provided to the system have to be anonymized, and just then they can be uploaded (once the new batch of data will be detected), using a custom script, into the segmentation system. Although the Label Studio application by default uses the same database and instances for all users, in the implementation described in this article it is ensured that the data are fully secure and isolated between groups of users. This means, for example, that one center cannot view data from other medical facilities. The only person who can view and add new groups of users (e.g., for a new medical facility that wants to join the multi-centric effort of data annotation) is the system administrator. Importantly, patient data have also been anonymized prior to uploading the data to the system. The procedure for allowing a new facility to use the system is that, first, a new instance/accounts are being created for the institution, and afterwards the medical staff can work on its data, as presented in Figure 10.2. The roles of users involved in the data processing workflow are presented in Figure 10.3.

***Figure 10.3**. Specialists from different domains can work together in the ENDO-CLOUD videoendoscopy data processing workflow.*


Centralized Service
Database
Application
Hospital1
User 1
User n
Different accounts for medical staff
Patient data is anonymized
The patient is assigned a unique ID number
Patients data are stored in the database
All data is secured and isolated
The hospital only has access to its own data
Access to the data is protected
The administrator configures the settings and grants access to the tool
The data of individual hospitals is protected and isolated from each other
Hospital2
User 1
User n
Different accounts for medical staff
Patient data is anonymized
The patient is assigned a unique ID number
Patients data are stored in the database
All data is secured and isolated
The hospital only has access to its own data
Access to the data is protected

Results

Classification results for the IBSU-1432 dataset

The reliability of the informative frames selection model was validated using the following metrics: accuracy, precision, recall, F1 score, and G-mean. The overall accuracy for the 10-fold cross- validation of the 4-class classifier (I vs B vs S vs U) was 84.64%. In the case of the binary (2-class) classifier differentiating between informative and non-informative frames (I vs B+S+U) the following metrics were obtained: accuracy (non-balanced classes), 92.36%; precision, 90.01%; recall, 83.80%; F1 score, 86.26%; and G-mean, 89.39%.

Semi-automatic frames selection

In the first step of this algorithm, one of the several ways of classifying a frame as NBI vs WL is being utilized. For certain version of the Olympus system, the devised workflow uses the optical character recognition (OCR) system to recognize if a frame has been recorded in NBI or in WL mode. Another approach utilizes histogram value analysis – this method yields slightly worse results than the OCR algorithm, however, the histograms are a more universal way of determining the type of the frame, and it can be used for different versions from various endoscopic system manufacturers. For the purpose of the current implementation of the ENDO-CLOUD system, if the frame is not classified as an NBI picture, it is not included in further processing. In the next stage, for all the frames selected as NBI frames, the EfficientNetB7 neural network[31] is used in order to create a feature map of each frame and recognize if it is informative or not (based on the pre-trained model). In the final stage, every 10th frame is selected to reduce the number of those recognized as "NBI informative" ones. Finally, the selected frames are imported into the Label Studio segmentation web application program. One of the key advantages of the semi-automatic selection of video frames is that the expert does not have to watch the entire videoendoscopy to manually select the informative frames (for a similar approach, in terms of ML technologies in general, see Breck et al.[32]). Thanks to the semi-automatic script, a lot of the manual work done by the medical experts can be avoided, thus allowing more frames to be segmented in the same amount of time.

Importantly, our work is based on the MLOps pipeline[33] which allows better management of the model development, and helps to provide more reliable results in a shorter period of time. As a first step of the MLOps pipeline, the workflow has been designed, and all the key terms and functionalities have been defined and described as follows: (1) dataset definition (including all three data sources: Moccia et al.,[24] Poznan, and Brescia); (2) DL architecture preparation; (3) splitting the dataset into training and testing subsets; (4) obtaining the model performance metrics; (5) reporting the results. In detail, the dataset has been obtained from the following sources: (i) the repository shared as part of the study by Moccia et al.;[24] (ii) the new data provided by the co-authors of this article from the Department of Otolaryngology, Head and Neck Surgery, Poznan University of Medical Sciences (Poznan, Poland); and (iii) data from the Department of

Otorhinolaryngology – Head and Neck Surgery, ASST Spedali Civili of Brescia, University of Brescia, School of Medicine (Brescia, Italy). After the dataset was prepared, as part of the transfer learning method, a pre-trained EfficientNetB7 neural network has been utilized.[34] The final, combined dataset has been then split into training and testing subsets, and the higher layers of the neural network have been trained. The resulting ML model has been used in a semi-automatic frame selection script (described in detail in the "Semi-automatic frame selection" section above). The project itself, as well as the models described in the current article, are in continuous development and integration (continuous delivery),[35,36] in order to ensure that the best possible and up-to-date models are being utilized in the system.

Discussion

Future developments

In this article, a web-based system has been described that enables a multi-centric approach to laryngeal pathologies segmentation within videoendoscopies. From the ML perspective, the AI algorithms that are currently utilized in the system can still be improved, including: NBI frames detection method, informative frames selection model, and a separate algorithm for the automatic segmentation and statistics calculation. These improvements can also be achieved by choosing different neural network implementations, extending (or changing) the training datasets, or by including more pathology types in the models.

Conclusions

A web-based system has been herein presented that allows semi-automatic frames selection and their manual segmentation once extracted from NBI videoendoscopies of the larynx. We believe that the implementation and wide usage of such a system can immensely contribute to the effort of creating reliable ML models supporting the work of medical experts in the field of otolaryngology.

References

1. Nogal P, Buchwald M, Staśkiewicz M, et al. Endoluminal larynx anatomy model - towards facilitating deep learning and defining standards for medical images evaluation with artificial intelligence algorithms. Otolaryngol Pol 2022; 76:1-9. https://doi.org/10.5604/01.3001.0015.9501
2. 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
3. Esmaeili N, Illanes A, Boese A, et al. Laryngeal Lesion Classification Based on Vascular Patterns in Contact Endoscopy and Narrow Band Imaging: Manual Versus Automatic Approach. Sensors 2022; 20:4018. https://doi.org/10.3390/s20144018
4. Esmaeili N, Sharaf E, Gomes Ataide EJ, et al. Deep Convolution Neural Network for Laryngeal Cancer Classification on Contact Endoscopy-Narrow Band Imaging. Sensors 2021; 21:8157. https://doi.org/10.3390/s21238157

5. Tran BA, Dao TT, Dung HD, et al. Support of deep learning to classify vocal fold images in flexible laryngoscopy. Am J of Otolaryng 2023; 44:103800. https://doi.org/10.1016/j.amjoto.2023.103800
6. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism 2017; 69S:S36-S40.
1. https://doi.org/10.1016/j.metabol.2017.01.011
7. Buchwald M, Przybylski Ł, Króliczak G. Decoding Brain States for Planning Functional Grasps of Tools: A Functional Magnetic Resonance Imaging Multivoxel Pattern Analysis Study. J Int Neuropsychol Soc 2018; 24:1013-1025. https://doi.org/10.1017/S1355617718000590
8. Wasilewicz R, Mazurek C, Pukacki J. Influence of cardiovascular system on 24 hour ocular volume changes, measured with contact lens sensor in healthy and POAG subjects. 8 World Glaucoma Congress, Melbourn 27-30.03.2019.
9. Esmaeili N, Illanes A, Boese A, et al. A Preliminary Study on Automatic Characterization and Classification of Vascular Patterns of Contact Endoscopy Images. Annu Int Conf IEEE Eng Med Biol Soc 2019:2703-2706. https://doi.org/10.1109/EMBC.2019.8857145
10. Esmaeili N, Illanes A, Boese A, et al. Novel automated vessel pattern characterization of larynx contact endoscopic video images. Int J Comput Assist Radiol Surg 2019; 14:1751-1761. https://doi.org/10.1007/s11548-019-02034-9
11. Mahmood H, Shaban M, Rajpoot N, et al. Artificial Intelligence-based methods in head and neck cancer diagnosis: an overview. Br J Cancer 2021; 124:1934-1940. https://doi.org/10.1038/s41416-021-01386-x

12. 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; 11:626602. https://doi.org/10.3389/fonc.2021.626602

13. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging 2012; 30:1323-41. https://doi/org/10.1016/j.mri.2012.05.001

14. Kuznetsov K, Lambert R, Rey JF. Narrow-band imaging: potential and limitations. Endosc 2006; 38:76-81. https://doi.org/10.1055/s-2005-921114

15. Ni XG, He S, Xu ZG, et al. Application of narrow band imaging endoscopy in the diagnosis of laryngeal cancer. Chin J Otorhinolaryngol Head and Neck Surg 2010; 45:143-7.

16. Ni XG, Wang GQ. The role of narrow band imaging in head and neck cancers. Curr Oncol Rep 2016; 18:1-7. https://doi.org/10.1007/s11912-015-0498-1

17. Witkiewicz J, Klimza H, Piersiala K, et al. The usefulness of the narrow band imaging (NBI) in decision-making process regarding second look procedure (SL) in laryngeal cancer follow-up after transoral laser microsurgery. PLoS One 2020; 15:e0236623. https://doi.org/10.1371/journal.pone.0236623

18. Pietruszewska W, Morawska J, Rosiak O, et al. Vocal Fold Leukoplakia: Which of the Classifications of White Light and Narrow Band Imaging Most Accurately Predicts Laryngeal Cancer Transformation? Proposition for a Diagnostic Algorithm. Cancers 2021; 13:3273. https://doi.org/10.3390/cancers13133273

19. Klimza H, Jackowska J, Tokarski M, et al. Narrow-band imaging (NBI) for improving the assessment of vocal fold leukoplakia and overcoming the umbrella effect. PLoS One 2017; 12:e0180590. https://doi.org/10.1371/journal.pone.0180590

20. Mehlum CS, Døssing H, Davaris N, et al. Interrater variation of vascular classifications used in enhanced laryngeal contact endoscopy Eur Arch Otorhinolaryngol 2020; 277:2485-2492. https://doi.org/10.1007/s00405-020-06000-z

21. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS One 2019; 14:e0224365. https://doi.org/10.1371/journal.pone.0224365

22. Gong J, Holsinger FC, Noel JE, et al. Using deep learning to identify the recurrent laryngeal nerve during thyroidectomy. Sci Rep 2021; 11:14306. https://doi.org/10.1038/s41598-021-93202-y

23. Yao P, Usman M, Chen YH, et al. Applications of Artificial Intelligence to Office Laryngoscopy: A Scoping Review. Laryngoscope 2022;132(10):1993-2016. https://doi.org/10.1002/lary.29886

24. Moccia S, Vanone GO, Momi E, et al. Learning- based classification of informative laryngoscopic frames. Comput Methods Programs Biomed 2018; 158:21-30. https://doi.org/10.1016/j.cmpb.2018.01.030

25. Philbrick KA, Weston AD, Akkus Z, et al. RIL-Contour: a Medical Imaging Dataset Annotation Tool for and with Deep Learning. J Digit Imaging 2019; 32:571-581.

2. https://doi.org/10.1007/s10278-019-00232-0

26. Rao D, K P, Singh R, J V. Automated segmentation of the larynx on computed tomography images: a review. Biomed Eng Lett 2022; 12:175-183. https://doi.org/10.1007/s13534-022-00221-3

27. El Naqa I, Ruan D, Valdes G, et al. Machine learning and modeling: Data, validation, communication challenges. Med Phys 2018; 45:e834-e840. https://doi.org/10.1002/mp.12811

28. Gómez P, Kist AM, Schlegel P, et al. BAGLS, a multihospital Benchmark for Automatic Glottis Segmentation. Sci Data 2022; 7:186. https://doi.org/10.1038/s41597-020-0526-3
29. Tkachenko M, Malyuk M, Shevchenko N, et al. Label Studio: Data labeling software (2020) Open source software available from https://github. com/heartexlabs/label-studio
30. Huang Y, Zhang H, Wen Y, et al. Modelci-e: Enabling continual learning in deep learning serving systems. arXiv preprint arXiv:2106.03122. 2021. https://doi.org/10.48550/arXiv.2106.03122
31. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks (2019) In International conference on machine learning 2019 May 24 (pp. 6105-6114). PMLR. Available from: https://proceedings.mlr.press/v97/tan19a.html
32. Breck E, Polyzotis N, Roy S, et al. Data Validation for Machine Learning. InMLSys 2019 Apr 2. Available from: https://mlsys.org/Conferences/2019/doc/2019/167.pdf
33. Symeonidis G, Nerantzis E, Kazakis A, et al. MLOps-Definitions, Tools and Challenges. In 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC) 2022 Jan 26 (pp. 0453-0460). IEEE. https://doi.org/10.1109/CCWC54503.2022.9720902
34. You K, Liu Y, Wang J, et al. LogME: Practical Assessment of Pre-trained Models for Transfer Learning. Proceedings of the 38th International Conference on Machine Learning 2021, in Proceedings of Machine Learning Research 139:12133-12143 Available from: https://proceedings.mlr.press/v139/you21b.html
35. Fowler M, Foemmel M. Continuous integration. https://tinyurl.com/ycbl2uhj [Online; accessed 28-June- 2023]
36. Meyer M. Continuous integration and its tools. IEEE Software 2014; 31:14-6. https://doi.org/10.1109/MS.2014.58

37. Patrini I, Ruperti M, Moccia S, et al. Transfer learning for informative-frame selection in laryngoscopic videos through learned features. Med Biol Eng Comput 2020; 58:1225-1238. https://doi.org/10.1007/s11517-020-02127-7

38. Behnke M, Buchwald M, Bykowski A, et al. Psychophysiology of positive and negative emotions, dataset of 1157 cases and 8 biosignals. Sci Data 2022; 9:10. https://doi.org/10.1038/s41597-021-01117-0

39. Stevens LM, Mortazavi BJ, Deo RC, et al. Recommendations for reporting machine learning analyses in clinical research. Circ-Cardiovasc Qual 2020; 13:e006556. https://doi.org/10.1161/CIRCOUTCOMES.120.006556

Conclusion

This dissertation underscores the transformative potential of computer vision in medicine, particularly in Otolaryngology – Head and Neck Surgery. The primary applications identified are in endoscopy and surgical procedures, but there is a clear connection emerging with radiologic imaging, pathology, and molecular biology.

The field of computer vision is evolving at an unprecedented pace. This rapid evolution necessitated a strategic approach in structuring the deliverables of this research. Consequently, the projects were divided into small, diverse work-packages. This methodology was adopted to circumvent the potential pitfall of investing excessively in a singular approach or technology, which bore the risk of becoming outdated or obsolete before the conclusion of the investigation. This flexible and adaptive strategy ensured that the research remained relevant, cutting-edge, and reflective of the current state of the field.

However, it is crucial to underline that the technical challenges associated with integrating computer vision and artificial intelligence into medicine are only one facet of the issue. The implementation of artificial intelligence-based applications in clinical practice is a complex endeavor that necessitates a thorough and deliberate ethical and regulatory evaluation. While this manuscript does not delve into these aspects in detail, they have been identified and acknowledged as a critical variable in each experimental setup, underlining their importance in the responsible and ethical development of artificial intelligence applications in medicine.

In conclusion, it is pertinent to highlight the burgeoning impact of large language models on society, a development that, while not directly linked to computer vision, is undeniably relevant. In particular, the demarcation lines between different artificial intelligence applications are increasingly becoming less distinct, paving the way for the advent of multimodal models. These models, capable of processing and interpreting various types of inputs (text, images, and more) and generating diverse outputs without necessitating distinct models for each data type, can potentially revolutionize the medical field. Furthermore, the emergence of independent agents powered by artificial intelligence is a development that is showing immense promise, and it is anticipated that they will contribute significantly to advancements in many fields, including medicine. This research, therefore, not only underscores the potential of computer vision in Otolaryngology but also sets the stage for future interdisciplinary innovations and integrations.