Deep Learning-Based Videomics for Automatic Segmentation in Endoscopic Endonasal Surgery

Ann. Ital. Chir., 2026 97, 3: 421–434
https://doi.org/10.62713/aic.4229

Edoardo Agosti[1], Andrea Pagnoni[1], Cesare Zoia[2], Vittorio Rampinelli[3], Alessandro Fiorindi[1], Pier Paolo Panciani[1], Alberto Paderno[4], Marco Maria Fontanella[1]

[1]Neurosurgery Division, Department of Medical and Surgical Specialties, Radiological Sciences and Public Health, University of Brescia, 25123 Brescia, Italy
[2]Neurosurgery Unit, Ospedale Moriggia Pelascini, 22015 Gravedona e Uniti, Italy
[3]Unit of Otolaryngology- Head and Neck Surgery, Department of Medical and Surgical Specialties, Radiological Sciences and Public Health, University of Brescia, 25121 Brescia, Italy
[4]Unit of Otorhinolaryngology, IRCCS Humanitas Research Hospital, 20089 Milan, Italy

AIM: Videomics, the application of deep learning (DL) to endoscopic video, enables real-time tissue segmentation and anatomical recognition. Within endoscopic endonasal approaches, these methods may improve intraoperative visualization, tumor delineation, and surgical precision. Despite growing interest, its translation into routine clinical practice is still limited and not yet fully characterized. This systematic review aimed to synthesize current evidence on DL-based segmentation in endoscopic endonasal surgery, focusing on model architectures, segmentation targets, and reported outcomes.

METHODS: This review was conducted according to Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines. A systematic search of PubMed, Scopus, and Web of Science was performed on 12 January 2025, and updated on 5 June 2025. Studies published between 2018 and 2025 were included, as no eligible studies were available prior to 2018. Studies were included if they involved human endoscopic endonasal procedures and applied DL techniques to endoscopic video for segmentation purposes. Data extraction included sample size, image resolution, annotated datasets, DL architectures, segmentation targets, and model performance metrics. Study quality was assessed using the Newcastle-Ottawa Scale, and descriptive statistics were used to summarize findings.

RESULTS: Out of 223 screened articles, 28 studies met the inclusion criteria, encompassing 154,989 patients and 1,028,440 annotated images. The most common segmentation targets included nasal polyps (25%), nasopharyngeal carcinoma (21.4%), and pituitary adenomas (7.14%). ResNet and YOLO architectures were each used in 5 studies (17.9%), while transformer-based models such as Swin Transformer, NasVLM, and NaMA-Mamba were increasingly utilized in recent years. Performance metrics were high across studies: area under the receiver operating characteristic curve (AUC-ROC) ranged from 87.4% to 99.2%, mean intersection over union [IoU] (mIoU) from 61.2% to 81.7%, and mean average precision (mAP) [0.50] from 53.4% to 94.9%. Inference times varied from 0.14 ms to 100 ms per image. However, only 35.7% of studies reported segmentation tools, and dataset heterogeneity was common.

CONCLUSIONS: DL-based videomics demonstrates high segmentation accuracy across various pathologies and anatomical targets in endoscopic endonasal surgery. Models such as Swin Transformer and YOLO show potential for real-time surgical support. However, translation into clinical practice remains limited by dataset heterogeneity and variability in reporting.

Keywords: videomics; deep learning; endoscopic endonasal surgery; automatic segmentation

Introduction

Videomics, the convergence of video-endoscopy and artificial intelligence (AI), is rapidly emerging as a transformative approach in surgical practice [1–3]. Previous studies on videomics and deep learning (DL) have provided the conceptual and methodological foundation for the present work. In gastrointestinal endoscopy, convolutional neural networks (CNNs) have been successfully applied to polyp detection and segmentation [4,5]. In laryngoscopy, DL has been used to delineate cancer boundaries and enhance diagnostic accuracy [6,7]. Similar applications have been described in sinonasal surgery, where models enabled the recognition and segmentation of nasal polyps and papillomas [8]. These early contributions highlight the feasibility of applying videomics across multiple anatomical regions and serve as a precedent for its potential in the endoscopic endonasal approach (EEA), particularly in skull base surgery [2,3]. By applying computer vision and DL tech-

Submitted: 23 June 2025 Revised: 9 September 2025 Accepted: 18 September 2025 Published: 10 March 2026
Correspondence to: Cesare Zoia, Neurosurgery Unit, Ospedale Moriggia Pelascini, 22015 Gravedona e Uniti, Italy (e-mail: gioiaoffice@gmail.com).
Editor: Abuzer Güngör

niques to endoscopic video data, videomics enables real-time analysis of anatomical structures, lesion recognition, and tissue segmentation [1,2,9,10]. This integration not only enhances diagnostic capabilities but also holds significant promise for improving intraoperative decision-making across various surgical domains [1,3,9,11].

Initially explored in gastrointestinal, dermatological, and upper aerodigestive tract procedures, videomics has demonstrated the capacity to automate lesion detection, guide navigation through complex anatomy, and assist in disease characterization [4–6,9,12,13]. Central to this development is the use of DL models, particularly CNNs, which excel at identifying and segmenting visual features in large-scale image data [14,15]. Through methods such as transfer learning and data augmentation, these models can be effectively trained even with limited clinical datasets, making them viable for specialized surgical environments [16,17].

The EEA has become a cornerstone technique in skull base surgery, providing minimally invasive access to midline structures such as the sellar, suprasellar, and clival regions [18,19]. Despite its advantages, EEA presents inherent challenges, including limited visual fields, complex and variable anatomy, and suboptimal lighting conditions [19,20]. These factors can hinder the surgeon's ability to precisely identify anatomical landmarks or pathological tissue, especially during critical phases of resection [21]. In this context, videomics offers an opportunity to augment the surgeon's perception by delivering automated, high-resolution analysis of intraoperative video in real time [22,23].

Automated segmentation, one of the most promising applications of videomics, can help delineate critical structures, highlight areas of interest, and support safe and complete resection of pathological tissue [13,24]. During EEA, this is particularly relevant for navigating around neurovascular structures and differentiating between normal tissue and residual disease. The potential impact spans not only oncologic outcomes but also the reduction of surgical complications such as cerebrospinal fluid (CSF) leaks or damage to adjacent critical anatomy [18–20].

Among the most studied applications of videomics in EEA is its role in the management of pituitary adenomas (PAs) [2,3]. Although typically benign, PAs present technical challenges during resection due to their proximity to the optic apparatus and normal pituitary tissue. Accurate intraoperative identification of residual tumor remains a key determinant of long-term outcomes [25,26]. Deep learning techniques have shown promise in segmenting adenoma tissue from surrounding structures on endoscopic video, particularly during the final stages of tumor removal, when visual cues are minimal. By enhancing intraoperative awareness, these tools may contribute to more complete resections and improved postoperative results [2,3].

To our knowledge, this is the first systematic review dedicated specifically to DL-based segmentation in endoscopic endonasal surgery. It aims to identify the main anatomical and pathological targets studied, evaluate the performance of different DL architectures, and highlight challenges related to methodological heterogeneity and clinical translation.

Methods

Study Design

This systematic review was carried out following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [27]. A comprehensive literature search was independently performed by two reviewers (EA and AP) using three electronic databases: PubMed, Scopus, and Web of Science. The initial query was conducted on 12 January 2025, and a final search update was completed on 5 June 2025.

Search Strategy

The search strategy was developed to capture studies investigating AI-based video analysis in the context of endoscopic endonasal procedures. A combination of keywords and Boolean operators was used to construct the search string as follows: (videomics OR artificial intelligence OR machine learning OR deep learning OR convolutional neural network OR segmentation) AND (endonasal OR nasal OR pituitary OR sella* OR parasellar) AND (endoscop*). Manual searches of the reference lists from relevant articles were also conducted to identify any additional studies not captured through the initial electronic search.

Eligibility Criteria

Studies were included if they met the following criteria: (1) written in English; (2) involved human subjects undergoing endoscopic endonasal surgery; and (3) explored the use of videomics or DL-based video analysis techniques, particularly segmentation, during endoscopic procedures. Eligible studies had to present original data and describe applications such as anatomical recognition, intraoperative guidance, lesion segmentation, or tissue differentiation. No restrictions were placed on the type of DL architecture: CNNs, transformer-based models, and hybrid approaches were all considered eligible if applied to segmentation tasks in endoscopic endonasal surgery. This ensured the comprehensiveness of the included studies. Studies were included if they applied DL to segmentation in endoscopic endonasal surgery, even when certain methodological details (e.g., image resolution, validation split) were incompletely reported. Missing information was explicitly coded as "Not Available (NA)" in the extraction sheet and in Table 1 (Ref. [2,3,7,8,10–12,28–48]). This approach maximized comprehensiveness while making reporting gaps visible, underscoring the need for greater methodological transparency

Table 1. Summary of videomics data of the studies included in the systematic review.

Author, Year	Patients					Annotation dataset			DL model	Performance (%**)
	Total Patients N	Total images N	Training dataset N (%)	Validation dataset (% or N*)	Testing dataset N (%)	Image resolution (pixel × pixel)	Segmentation software	Evaluated target		
Li *et al.* [39], 2018	7951	27,536	70	20	10	500 × 500	NA	Nasopharingeal malignancy	endoscopic images-based nasopharyngeal malignancy detection model (eNPM-DM)	Accuracy: 88.7
Girdler *et al.* [46], 2021	297	297	75	10	15	224 × 224	NA	IP, NP	ResNet-152	Accuracy: 74.2
Bieck *et al.* [8], 2022	30	180,000	80	20	NA	NA	NA	Five anatomical landmarks during FESS	ResNet-50	F1-Score: 47.0
Ay *et al.* [40], 2022	80	4640	80	300*	20	550 × 550	NA	NP	3CH-CNN	Accuracy: 98.3
Bi *et al.* [48], 2023	NA	3179	80	996*	20	256 × 256	NA	AH grading	MIB-ANet	Accuracy: 76.8
Chen *et al.* [47], 2023	NA	405	80	NA	20	765 × 570	LabelMe	Nasal bleeding	ETU-Net	F1-Score: 76.3 Inference time: 50 ms IoU: 94.6
Das *et al.* [41], 2023	64	635	80	NA	20	1280 × 720	NA	Sella, clival recess, eight parasellar landmarks	PAINet	F1-Score: 97,2 IoU sella: 66.1 IoU clival recess: 54.1 MPCK-20% eight landmarks: 53.2 Inference time: 100 ms
Yui *et al.* [37], 2023	53	143,167	87.5	NA	12.5	224 × 224	NA	IP	MobileNet-V2	AUC-ROC: 87.4 Accuracy: 84.3
Fuse *et al.* [3], 2023	50	605	60	20	20	128 × 128	VoTT	PA	ResNet-50 wide-ResNet-50–2 DenseNet-161	Accuracy: 75.1 F1-Score: 75.1 Accuracy: 76.8 F1-Score: 76.6 Accuracy: 75.3 F1-Score: 75.2
Kwon *et al.* [42], 2024	4340	4340	60	400*	40	224 × 224 or 331 × 331	OpenCV	Normal nasal cavity, NP, benign tumor, malignant tumor	Xception	Accuracy: 82.0 F1-Score: 81.0

Table 1. Continued.

Author, Year	Patients					Annotation dataset			DL model	Performance (%**)
	Total Patients N	Total images N	Training dataset N (%)	Validation dataset (% or N*)	Testing dataset N (%)	Image resolution (pixel × pixel)	Segmentation software	Evaluated target		
Phoommanee *et al.* [29], 2024	73	8972	NA	NA	2257	384 × 384	NA	Nasal obstruction grading	ViT-tin	Accuracy validation set: 94.8 F1-Score validation test: 89.4 Accuracy testing test: 71.1 F1-Score testing set: 33.1 Inference time: 37.04 ms
									SVM	Accuracy validation set: 79.6 F1-Score validation test: 60.3 Accuracy testing test: 69.6 F1-Score testing set: 31.8 Inference time: 0.14 ms
Phoommanee *et al.* [33], 2024	62	779	NA	NA	22	512 × 512	NA	Nasal septum, IT, MT, NP, others, airway	Mask2Former (Swin-T)	mIoU validation set: 81.7 mIoU testing set: 61.2 Inference time: 68.7 ms
Tai *et al.* [43], 2024	1442	1442	70	10	20	256 × 256	NA	NP, IP	InceptionResNetV2	NP accuracy: 82.7 NP F1-Score: 84.1 IP accuracy: 82.7 IP F1-Score: 61.2
Mao *et al.* [11], 2024	64	635	80	NA	20	1280 × 720	NA	Sella, clival recess, four parasellar landmarks	PitSurgRT	IoU sella: 67.0 IoU clival recess: 45.9 MPCK-20% four landmarks: 97.9
									PitSurgRT + TensorRT	Inference time (fps32): 44.1 ms Inference time (fps16): 33.5 ms
Ganeshan *et al.* [10], 2024	NA	2111	80	15	5	1024 × 768	CocoAnnotator	IT, MT	YOLOv8	Accuracy: 91.5 F1-Score: 93.1 Precision: 91.0 Recall: 83.9 mAP [0.50]: 94.9 IoU (mAP [0.50:0.95]): 67.9 Inference time: 9.8 ms
Rampinelli *et al.* [28], 2024	52	342	80	10	10	640 × 640	Roboflow	NP	YOLOv8.0.28	

Table 1. Continued.

Author, Year	Patients					Annotation dataset		DL model	Performance (%**)
	Total Patients N	Total images N	Training dataset N (%)	Validation dataset (% or N*)	Testing dataset N (%)	Image resolution (pixel × pixel)	Segmentation software	Evaluated target	
Xu *et al.* [38], 2024	1050	6300	60	20	20	576 × 720	NA	NP, IP, fungal sinusitis, nasal malignant tumor	Att-Res2-UNet	Accuracy: 97.9
Yue *et al.* [34], 2024	2134	38,073	70	10	20	224 × 224	NA	NPC, AH, allergic rhinitis, chronic rhinosinusitis with nasal polyps, deviated nasal septum, NOR, rhinosinusitis	Nose-Keeper (Swin Transformer)	Accuracy: 92.3
Lei *et al.* [30], 2025	NA	4000	NA	NA	NA	NA	NA	NOR	ResNet-50	Accuracy: 98.6 F1-Score: 98.6 Recall: 98.7
Liu *et al.* [35], 2025	67,900	271,600	90	10	NA	224 × 224	NA	NOR, benign hyperplasia, NPC	NasVLM	NPC accuracy: 94.5 NPC AUC-ROC: 99.2
Wang Z *et al.* [12], 2024	NA	25,278	97	NA	3	224 × 224	NA	AH grading	Xception	Accuracy: 95.5 AUC-ROC grade I: 93.0 AUC-ROC grade II: 94.0 AUC-ROC grade III: 97.0 AUC-ROC grade IV: 91.0 Average:
Bidwell *et al.* [31], 2025	NA	2111	80	5	15	NA	CocoAnnotator	IT, MT	YOLOv8 (all)	F1 score: 92.9 Accuracy: 87.0 Inference time: 90.9 ms
									YOLOv8-nano	F1 score: 92.0 Accuracy: 86.0 Inference time: 25.6 ms
He *et al.* [44], 2025	906	8816	70	10 + 2818*	20	512 × 512	LabelMe	NPC	NPC-SDNet	Diagnostic accuracy: 95.0 (WLI), 95.8 (NBI) Segmentation accuracy: 94.7 (WLI), 92.2 (NBI) IoU: 83.7 (WLI), 83.4 (NBI)
Staartjes *et al.* [7], 2025	146	19,000	NA	NA	20*	NA	VoTT	Anatomical landmarks in the nasal and sellar phase of TSS approach	YOLOv7	mAP [0.50]: 53.4

Table 1. Continued.

Author, Year	Patients					Annotation dataset			DL model	Performance (%**)
	Total Patients N	Total images N	Training dataset N (%)	Validation dataset (% or N*)	Testing dataset N (%)	Image resolution (pixel × pixel)	Segmentation software	Evaluated target		
Mao *et al.* [45], 2025	64	635	80	NA	20	1280 × 720	NA	Sella, clival recess, four parasellar landmarks	ConsisTNet	IoU sella: 66.8 IoU clival recess: 45.1 F1-Score sella: 80.6 F1-Score clival recess: 60.2 MPCK-20% four landmarks: 96.3
									ConsisTNet + Tensor RT	Inference time (fps32): 8.8 ms Inference time (fps16): 5.0 ms
Levi *et al.* [32], 2025	311	1242	80	10	10	224 × 224	Encord	Sinonasal masses	EfficientNet-B2	Accuracy: 90.5
									nnUNet	F1-Score: 89.7 Segmentation accuracy: 72.3 (polyp) and 72.8 (tumor)
Wang W *et al.* [36], 2025	67,900	271,600	90	10	NA	224 × 224	NA	NOR, NPC	NaMA-Mamba	NPC accuracy: 92.7 NPC F1-Score: 90.7 NPC AUC-ROC: 96.3
Agosti *et al.* [2], 2025	20	700	80	10	10	640 × 640	Label Studio	PA	YOLOv8x-seg	mAP [0.50]: 83.7 mAP [0.50:0.95]: 50.9 DSC: 83.0
									Swin Transformer	mAP [0.50]: 91.3 mAP [0.50:0.95]: 60.1 DSC: 89.0 Recall: 91.0 Precision: 88.0

Abbreviations: AH, adenoidal hypertrophy; AUC-ROC, area under the receiver operating characteristic curve; DL, deep learning; DSC, Dice Similarity Coefficient; FESS, Functional Endoscopic Sinus Surgery; IoU, intersection over union; IP, inverted papilloma; IT, Inferior Turbinate; mIoU, mean IoU; MPCK-20, mean percentage of correct keypoints within 20%; MT, Middle Turbinate; ms, milliseconds; N, Number; NA, Not Available; NBI, Narrow Band Imaging; NOR, NORmal nasal cavity and nasopharynx; NP, Nasal Polyps; NPC, Nasopharhyngeal Carcinoma; PA, Pituitary Adenoma; TSS, Trans-Sphenoidal Surgery; VoTT, Visual object Tagging Tool; WLI, White Light Imaging; eNPM-DM, endoscopic images-based nasopharyngeal malignancy detection model; 3CH-CNN, 3-Chamber view Convolutional Neural Network; MIB-ANet, Modified Inception Block-Adenoid Network; ETU-Net, Efficient Transformer and convolutional U-style connected attention network; PAINet, Pituitary Anatomy Identification Network; PitSurgRT, Pituitary Surgery Real-Time; ViT-tin, Vision Transformers Tiny; mAP [0.50], mean average precision at IoU threshold 0.50; mAP [0.50:0.95]: mean average precision across IoU thresholds from 0.50 to 0.95 in steps of 0.05.

N*: Additional images out of the total images; **: The inference time is always specified to be measured in ms. When the articles reported it in frames per second (fps) (i.e., Mao *et al.* [11], 2024 and Mao *et al.* [45], 2025), it was converted to ms.

in future studies. Exclusion criteria comprised: (1) review articles, editorials, or opinion pieces; (2) conference abstracts lacking peer review; (3) technical or methodological studies without clinical context or human data; and (4) animal or phantom-only research. Additionally, studies were excluded if they did not clearly describe the AI methodology or failed to report clinically relevant outcomes.

Study Selection

All references were managed using EndNote X9 software (Build 15659, Clarivate Plc, Philadelphia, PA, USA). Duplicate entries were removed prior to the screening process. Two reviewers (EA and AP) independently screened all titles and abstracts for relevance. Full texts of potentially eligible studies were then assessed. Inter-rater reliability was assessed using Cohen's kappa at the title/abstract screening stage, resulting in $\kappa = 0.82$, which indicates strong agreement. Discrepancies were resolved through consensus or, when necessary, consultation with a third reviewer (PPP).

Data Extraction

A structured data collection sheet was employed to systematically extract information from each included study. The following variables were recorded: first author, year of publication, total number of patients, number of patients included in the annotated dataset, total number of images, and the distribution of data into training, validation, and testing sets (reported as either number or percentage). Technical specifications such as image resolution (in pixels), the segmentation software used, the specific evaluated anatomical or pathological target, and DL models were also collected. When methodological details such as segmentation software or image resolution were not reported, these were recorded as "Not Available (NA)" in the extraction sheet. Authors were not contacted for additional information, as the aim of this review was to synthesize published evidence. This limitation was explicitly considered when interpreting the reproducibility of included studies.

In terms of performance evaluation, key metrics were extracted to assess the effectiveness of each videomics approach. These included: accuracy, F1-score, sensitivity, specificity, precision, recall, area under the receiver operating characteristic curve (AUC-ROC), intersection over union (IoU), mean IoU (mIoU), mean average precision at IoU threshold 0.50 (mAP [0.50]), IoU across specific anatomical regions (e.g., sella, clival recess), mean percentage of correct keypoints within 20% (MPCK-20), and inference time.

Outcomes

The primary objective of this systematic review was to catalog and evaluate the current clinical applications of videomics and DL-based segmentation in endoscopic endonasal surgery. This included assessment of model functionality in real-time surgical guidance, anatomical structure identification, tumor or tissue boundary detection, and impact on intraoperative workflow or surgical outcomes. Secondary outcomes included classification of DL techniques used, dataset characteristics, and performance metrics reported.

Risk of Bias Assessment

Given that most studies included in this review were retrospective observational studies based on clinical endoscopy datasets, the Newcastle–Ottawa Scale (NOS) was employed to assess study quality and potential bias [49]. The scale assesses research methodology across three domains: cohort selection, definition of inclusion and exclusion criteria, and outcome assessment (**Supplementary material 1**).

Methodological quality assessment was conducted for each study using the NOS. Scores guided the narrative synthesis, explored the robustness of study findings, and identified potential sources of heterogeneity. Quality scores for each included study are summarized in **Supplementary material 2**.

Statistical Analysis

Descriptive statistics were applied to summarize the characteristics of the included studies, videomics methodologies, and model performance metrics. Results were presented using medians, ranges, and proportions as appropriate. Due to the methodological heterogeneity across studies, including differences in DL architectures, segmentation targets, validation techniques, and clinical endpoints, a meta-analysis was not performed. All statistical analyses and visualizations were performed using R (version 4.2.0; https://www.r-project.org).

Results

Literature Review

Following the removal of duplicate records, a total of 223 records were screened. After screening titles and abstracts, 44 studies were deemed potentially relevant and were retrieved for full-text evaluation. Upon detailed assessment, 28 studies satisfied the eligibility criteria and were included in the final review. The remaining 16 articles were excluded for the following reasons: 14 were not aligned with the core objectives of this review and 2 lacked relevant outcome data. A detailed summary of the study selection process is illustrated in the PRISMA flow diagram (Fig. 1). The PRISMA checklist is available in **Supplementary material 3**. Study quality, as assessed by the NOS, ranged between 7 and 9 points for all the included studies, classifying them as high quality (**Supplementary material 2**).

Data Analysis

A total of 28 studies published between 2018 and 2025 were included, collectively analyzing 154,989 patients and 1,028,440 annotated endoscopic images. All studies re-

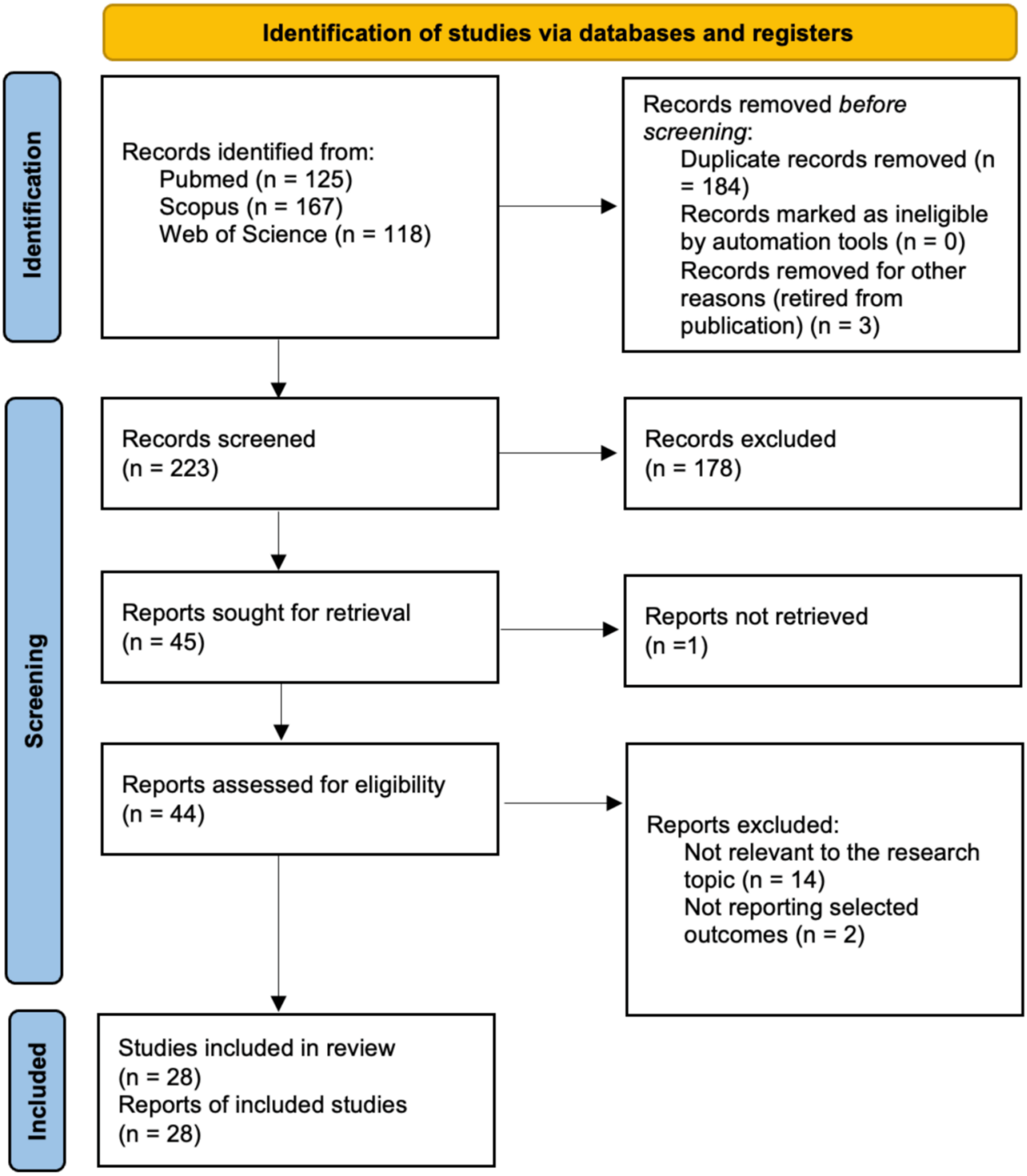

Fig. 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow chart.

ported the use of annotated datasets, but the number of frames varied substantially, from fewer than 700 to more than 270,000 images. Data were typically split into training, validation, and testing sets in ratios ranging between 60–97%, 10–20%, and 3–40%, respectively, although several studies also reported absolute numbers (e.g., 400 images in validation sets). Studies allocating less than 10% of their data to test sets frequently reported disproportionately high-performance metrics, raising the possibility of overfitting and inflated accuracy estimates. This methodological weakness highlights the importance of balanced data splitting to ensure generalizable performance. Image resolution ranged from 128 × 128 to 1280 × 720 pixels, with 224 × 224 (28.6% of studies) being the most frequently used format. Despite the importance of annotation methodology, only 10 studies (35.7%) disclosed the software used,

including LabelMe, Visual object Tagging Tool (VoTT), CocoAnnotator, Roboflow, OpenCV, Encord, and Label-Studio, whereas 64.3% omitted this information. These methodological differences illustrate a significant source of heterogeneity that affects reproducibility.

Segmentation targets were diverse, reflecting both oncologic and anatomical applications. Nasal polyps were the most frequent focus (7 studies, 25%), followed by nasopharyngeal carcinoma (6 studies, 21.4%). Other targets included sinonasal masses such as inverted papilloma (5 studies, 17.9%), inferior and/or middle turbinates (5 studies, 17.9%), clival/sellar landmarks (4 studies, 14.3%), adenoidal hypertrophy (3 studies, 10.7%), and PAs (2 studies, 7.1%). This distribution demonstrates the dual role of videomics in disease characterization and intraoperative navigation.

The included studies covered a wide spectrum of DL architectures, ranging from traditional CNNs (ResNet, Xception, InceptionResNetV2, MobileNet-V2) [3,8,30] to object detection models (YOLOv7, YOLOv8) [2,10,28,31], and emerging transformer-based frameworks (Swin Transformer, NasVLM, NaMA-Mamba) [2,33–36]. This distribution reflects the methodological diversity of the field. CNNs such as ResNet-50 and ResNet-152 were reported in 5 studies (17.9%), while YOLO-based architectures (YOLOv7, YOLOv8) were also used in 5 studies (17.9%). Other CNN-based models included Xception, InceptionResNetV2, MobileNet-V2, EfficientNet-B2, and nnUNet. More recently, transformer-based approaches were increasingly adopted, including Swin Transformer (3 studies), NasVLM (1 study), and NaMA-Mamba (1 study). This trend reflects a methodological shift toward architectures with greater capacity for multi-scale contextual analysis. Models such as YOLOv8 and Swin Transformer achieved top-performing metrics and serve as benchmarks of both technical advancement and translational innovation in videomics.

Performance outcomes were generally high, but with considerable variability. Reported accuracy ranged from 71.1% to 98.6%, F1-score from 33.1% to 98.6%, and recall from 83.9% to 98.7%. Area under the receiver operating characteristic curve (AUC-ROC) values spanned 87.4% to 99.2%. Segmentation-specific metrics demonstrated similar heterogeneity: IoU ranged from 45.1% to 94.6%, mIoU from 61.2% to 81.7%, mAP [0.50] from 53.4% to 94.9%, and MPCK-20 from 53.2% to 97.9%. Notably, studies with larger and higher-resolution datasets, or with standardized annotation protocols, tended to achieve superior boundary delineation and generalization ability.

Inference time was inconsistently reported, with only 8 studies (28.6%) providing data. Reported times ranged from 0.14 ms per image for lightweight SVM-based models to approximately 100 ms for more complex multi-stage networks. YOLOv8 achieved inference times as low as 9.8 ms per image, while transformer-based frameworks (e.g., Swin Transformer, NaMA-Mamba) reported longer but still clinically feasible times when optimized with engines such as TensorRT. These findings suggest a trade-off between segmentation accuracy and processing speed, with CNNs favoring efficiency and transformers excelling in precision. Reported inference times were not normalized to hardware specifications, as only a minority of studies provided full details of GPU type, memory, or acceleration engines. As a result, cross-study comparisons should be interpreted cautiously. Subgroup comparison of CNN-based versus transformer-based models showed that transformer architectures achieved superior segmentation accuracy (mAP [0.50], mIoU), whereas CNN-based models, particularly YOLO variants, consistently achieved faster inference times. This highlights a trade-off between precision and computational efficiency, which is relevant for clinical translation.

Discussion

This systematic review highlights the rapid advancement of deep learning-based videoomics approaches (including segmentation-focused research protocols) in transnasal endoscopic surgery and their potential clinical applications, particularly in PA resection [2,3]. Among the 28 studies reviewed, a diverse array of convolutional and transformer-based architectures was employed for intraoperative video analysis, with consistently high performance across metrics. Notably, models such as ResNet-50 [3,8,30], YOLOv8 [2,10,28,31], and nnUNet [32], along with emerging transformer-based architectures like Swin Transformer [2,33,34], NasVLM [35], and NaMA-Mamba [36], demonstrated superior multi-scale spatial representation, positioning them as leading candidates for clinical translation.

Emerging Role of Transformer-Based Architectures

Among the DL models identified, transformer-based architectures such as Swin Transformer, NasVLM, and NaMA-Mamba have demonstrated superior capabilities in anatomical segmentation by capturing long-range dependencies and multi-scale contextual features [2,33–36]. While traditional CNNs like ResNet-50 and MobileNet-V2 were widely utilized, they often rely on local receptive fields and may fall short when dealing with the complex, variable anatomy of the skull base [3,8,30,37]. In contrast, Swin Transformer achieved state-of-the-art performance, surpassing CNN-based models such as YOLOv8 and Mask R-CNN [2,33,34]. These results demonstrate the innovative role of DL in surgical video analysis. While earlier CNNs established feasibility [3,8,30,37], transformer-based models represent a step change by enabling fine-grained anatomical delineation and robust generalization [2,33–36]. These features are critical for clinical adoption and underscore the importance of DL beyond feasibility, highlighting tangible improvements in accuracy, boundary detection, and resilience to intraoperative variability. These findings are

consistent with prior reports (e.g., PitSurgRT and U-Net-based systems), where CNNs improved localization but were constrained by segmentation accuracy and generalization ability, particularly in challenging intraoperative environments [11,32,38]. Comparative benchmarking indicates that transformers achieve their best performance in high-resolution, well-annotated datasets, while CNN-based models, particularly YOLO variants, remain more robust in smaller or heterogeneous datasets. This underscores that dataset quality and annotation practices are decisive modifiers of model performance. CNNs also maintain a consistent advantage in inference speed, reinforcing their suitability for real-time integration, whereas transformers deliver superior segmentation accuracy at a higher computational cost.

Overall, CNNs and transformers should be considered complementary. CNNs provide lightweight, rapid solutions for intraoperative support, while transformers expand the frontier of precision and generalization. Together, they represent a major advance toward clinically viable videomics in endoscopic endonasal surgery.

Videomics in PA Surgery

The clinical utility of videomics in PA surgery is increasingly evident, particularly in aiding the differentiation between tumor and normal pituitary tissue, an essential requirement for maximizing resection while minimizing complications. Fuse *et al*. [3] conducted one of the earliest pilot studies applying DL to intraoperative endoscopic images for PA identification, using a Wide-ResNet architecture. Although their model achieved a modest accuracy of 76.8%, it demonstrated the feasibility of using deep neural networks for intraoperative tissue classification, highlighting limitations related to small datasets and variability in intraoperative imaging conditions.

Building on this foundational work, Agosti *et al*. [2] developed a more comprehensive videomics pipeline incorporating both YOLO-based and transformer-based segmentation models trained on a curated dataset of 700 annotated endoscopic frames. Among the proposed models, the Swin Transformer outperformed all others across nearly every evaluation metric. Its hierarchical self-attention mechanism enabled robust capture of multi-scale spatial features, crucial for delineating complex anatomical boundaries such as residual adenoma margins in the sellar region (Fig. 2). Compared to the YOLO variants—which, while fast and efficient, tended to trade off fine-grained segmentation accuracy for inference speed—the Swin Transformer maintained high boundary precision even under challenging intraoperative conditions, achieving a test segmentation mAP [0.50] of 0.913 and mAP [0.50:0.95] of 0.601 [2,7,10,28,31,33,38]. Mask R-CNN models, traditionally strong performers in image segmentation tasks, showed good overall accuracy but were more prone to overfitting and underperformed in generalization compared to Swin, particularly after prolonged training [2,32]. YOLOv8x-seg, the best-performing YOLO model, delivered a respectable mAP [0.50] of 0.837 but demonstrated reduced generalization on more nuanced segmentation tasks, as reflected in its lower mAP [0.50:0.95] [2]. These results suggest that while YOLO-based models may be suitable for rapid detection tasks, transformer-based architectures are better suited for tasks demanding high-resolution spatial awareness and precision. The combination of superior generalization, accurate edge delineation, and resilience to intraoperative visual noise marks the Swin Transformer as a promising candidate for integration into real-time surgical guidance systems [2,33,34].

Segmentation of Anatomical and Pathological Targets

The application of videomics to the segmentation of anatomical and pathological targets in endoscopic endonasal surgery extends beyond PAs, encompassing a broad spectrum of mucosal pathologies (e.g., nasal polyps and nasopharyngeal carcinoma) and endoscopic and skull base landmarks, including turbinates, and clival and sellar regions [2,3,7,8,11,28,32–36,38–45]. This breadth reflects the growing recognition of intraoperative video as a rich source of spatial-temporal data for surgical navigation and decision support. Several studies demonstrated the capacity of DL models to generalize across different target types with consistently high performance, suggesting robust feature extraction and contextual learning capabilities [7,8,10,32,33,35,38,41,42,45,46].

The segmentation of oncologic lesions, such as Nasopharhyngeal Carcinoma (NPC), is particularly promising for enhancing intraoperative margin assessment and minimizing residual disease [34–36,39,42,44]. Likewise, accurate identification of anatomical landmarks can aid in avoiding critical structures, potentially reducing complications and operative time. These applications align with broader trends in surgical data science, where real-time tissue recognition and context-aware assistance are pivotal for precision surgery [7,8,11,41,45]. However, variability in annotation tools (e.g., LabelMe, COCO Annotator, Roboflow, Label Studio) [2,10,28,31,44,47] and the frequent omission of labeling protocols in over 60% of studies point to a pressing need for standardized dataset curation. Furthermore, the wide range of input resolutions (i.e., from 128 × 128 to 1280 × 720 pixels) introduces heterogeneity that may affect model reproducibility and generalizability. Establishing consensus on image acquisition standards and benchmarking metrics (e.g., IoU, mIoU, MPCK-20) will be essential to validate model performance across institutions and pathologies, thereby facilitating clinical translation.

Integration and Real-Time Feasibility

Real-time integration remains a critical factor for clinical adoption. While several studies reported inference times under 10 milliseconds, the variation in computational se-

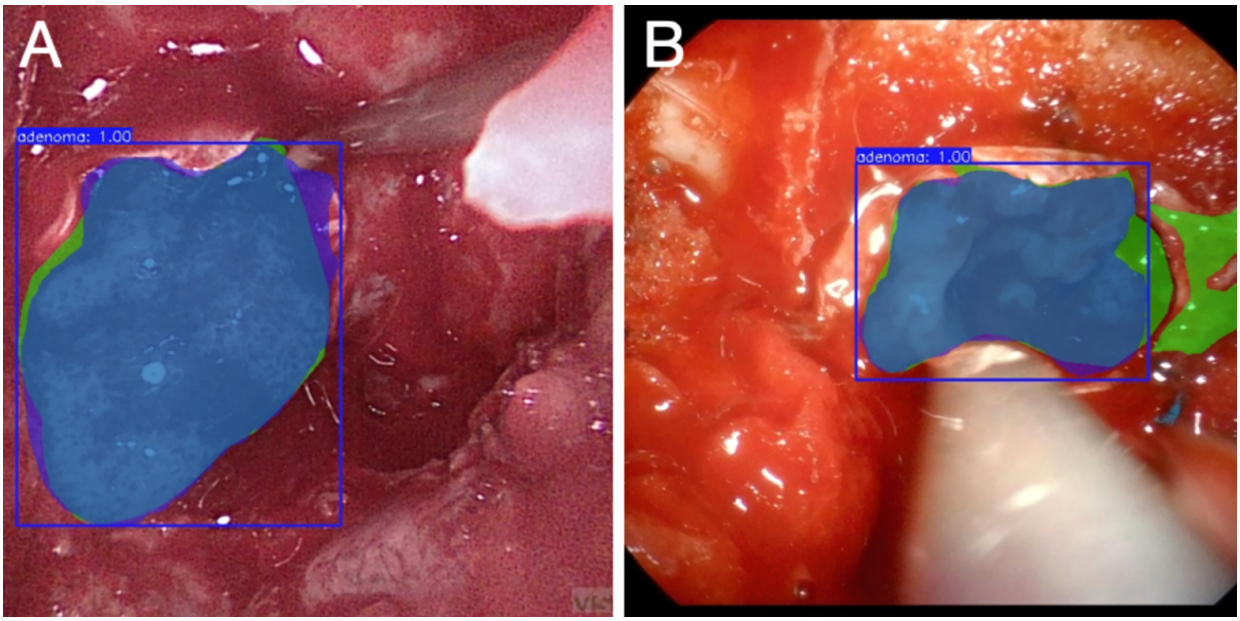

Fig. 2. Sequential endoscopic frames from a PA procedure during Swin Transformer testing. Green: manual segmentation; blue: model prediction; purple: overlap. AUC-ROC: 1.00. (A) Early stage: PA identified after dural flap retraction. (B) Mid-stage: PA tissue removed with a curette. Source: Original image material from the University of Brescia. Patient consent was obtained.

tups and model complexities limits direct comparison [11, 28,29,31,33,41,45,48]. Nonetheless, models like YOLOv8 and NaMA-Mamba showed high potential for intraoperative deployment due to their balance of speed and segmentation accuracy [2,28,31]. Real-time surgical guidance, including detection of neurovascular landmarks and residual tumor margins, was evaluated in multiple studies, indicating not only feasibility but a tangible impact on surgical decision-making and outcomes such as extent of resection and complication rates (e.g., CSF leaks) [3,7,10,41,44,45]. Beyond technical performance, the clinical implications of DL-based videomics are substantial. Real-time segmentation can enhance intraoperative decision-making, reduce the risk of complications such as CSF leaks, and support surgical education by providing automated visual feedback. These aspects highlight the potential of videomics not only for research but also for tangible improvements in patient safety and surgical outcomes.

However, despite these promising results, the translation of DL-based videomics into clinical practice remains limited. Key barriers include regulatory and ethical approval processes, which require demonstration of safety, robustness, and explainability before clinical deployment; difficulties in cross-institutional data sharing due to privacy and standardization concerns; hardware and computing costs, especially in resource-limited settings; and the challenge of ensuring real-time performance within the constraints of operating room environments. Furthermore, clinicians may be hesitant to rely on algorithmic outputs without transparent validation and interpretability. These obstacles help explain why, despite excellent experimental performance, widespread clinical implementation has not yet occurred.

Limitations of the Study

Many of the included studies were single-center and retrospective, potentially limiting generalizability. Sample sizes varied widely, and some used fewer than 400 annotated images for validation. Manual segmentation remains the primary method of annotation, which introduces inter-observer variability. Furthermore, less than half of the studies disclosed full methodological details, including segmentation software and DL pipeline parameters. Moreover, performance validation across multiple institutions and endoscopic platforms remains scarce. Another important limitation relates to the lack of systematic discussion of non-technical barriers. Even though the algorithms demonstrate excellent accuracy, issues of regulatory approval, ethical governance, data privacy, and infrastructure costs remain largely unaddressed in the literature, further delaying clinical adoption. To address these issues, future studies should focus on multi-institutional collaborations, standardized data sharing frameworks, and the development of semi-automated annotation tools to scale dataset creation while maintaining accuracy. Equally, prospective clinical trials and partnerships with regulatory bodies will be essential to assess real-world feasibility and ensure compliance with ethical and legal frameworks. Another limitation is that most of the included studies were published within the last three years, reflecting the very recent development of this field. This short research period means that evidence is still preliminary, follow-up is scarce, and long-term clinical outcomes remain underreported. These factors limit the strength of conclusions and highlight the need for longitudinal and multicenter validation before widespread clinical implementation.

Conclusions

Deep learning-based segmentation has shown potential to transform intraoperative guidance into endoscopic endonasal surgery. With architectures like ResNet, YOLO, and Swin Transformer achieving high performance on intraoperative endoscopic video data, these tools could offer a pathway to safer and more precise surgical orientation and tumor resection. Nevertheless, translation into routine clinical practice remains constrained by regulatory, ethical, technical, and economic challenges, in addition to the need for robust multicenter validation. Addressing these barriers will be crucial for moving from promising research results toward widespread clinical adoption.

Availability of Data and Materials

The data used and analyzed during the current study are available from the corresponding author on reasonable request.

Author Contributions

EA, AP, CZ, VR, AF, PPP, AP, and MMF designed the research study. EA and AP performed the research. EA drafted the manuscript. CZ, VR, AF, PPP, AP, and MMF participated in the data analysis. All authors contributed to critical revision of the manuscript for important intellectual content. All authors read and approved the final manuscript. All authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

Ethics Approval and Consent to Participate

Not applicable.

Acknowledgment

Not applicable.

Funding

This research received no external funding.

Conflict of Interest

Cesare Zoia is serving as one of the Editorial Board Members of this journal. We declare that Cesare Zoia had no involvement in the peer review of this article and has no access to information regarding its peer review. Other authors declare no conflict of interest.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.62713/aic.4229.

References

[1] Paderno A, Gennarini F, Sordi A, Montenegro C, Lancini D, Villani FP, *et al*. Artificial intelligence in clinical endoscopy: Insights in the field of videomics. Frontiers in Surgery. 2022; 9: 933297. https://doi.org/10.3389/fsurg.2022.933297.

[2] Agosti E, Paderno A, Pagnoni A, Rampinelli V, Panciani PP, Fiorindi A, *et al*. Deep learning for automatic segmentation of pituitary adenomas using narrow band imaging: preliminary experience in a clinical perspective. Journal of Neurological Surgery Part B: Skull Base. 2025; 86: 576. https://doi.org/10.1055/s-0045-1803861.

[3] Fuse Y, Takeuchi K, Hashimoto N, Nagata Y, Takagi Y, Nagatani T, *et al*. Deep learning based identification of pituitary adenoma on surgical endoscopic images: a pilot study. Neurosurgical Review. 2023; 46: 291. https://doi.org/10.1007/s10143-023-02196-w.

[4] Jha D, Ali S, Tomar NK, Johansen HD, Johansen D, Rittscher J, *et al*. Real-Time Polyp Detection, Localization and Segmentation in Colonoscopy Using Deep Learning. IEEE Access: Practical Innovations, Open Solutions. 2021; 9: 40496–40510. https://doi.org/10.1109/ACCESS.2021.3063716.

[5] Oukdach Y, Garbaz A, Kerkaou Z, El Ansari M, Koutti L, El Ouafdi AF, *et al*. UViT-Seg: An Efficient ViT and U-Net-Based Framework for Accurate Colorectal Polyp Segmentation in Colonoscopy and WCE Images. Journal of Imaging Informatics in Medicine. 2024; 37: 2354–2374. https://doi.org/10.1007/s10278-024-01124-8.

[6] Sampieri C, Azam MA, Ioppi A, Baldini C, Moccia S, Kim D, *et al*. Real-Time Laryngeal Cancer Boundaries Delineation on White Light and Narrow-Band Imaging Laryngoscopy with Deep Learning. The Laryngoscope. 2024; 134: 2826–2834. https://doi.org/10.1002/lary.31255.

[7] Staartjes VE, Sarwin G, Carretta A, Zoli M, Mazzatenta D, Regli L, *et al*. AENEAS Project: Live Image-Based Navigation and Roadmap Generation in Endoscopic Neurosurgery Using Machine Vision. Operative Neurosurgery (Hagerstown, Md.). 2025. https://doi.org/10.1227/ons.0000000000001583.

[8] Bieck R, Heuermann K, Sorge M, Neumuth T, Pirlich M. Saliency-assisted multi-label classification for explainable deep learning applications in endoscopic ENT navigation. Current Directions in Biomedical Engineering. 2022; 8: 596–599. https://doi.org/10.1515/cdbme-2022-1152.

[9] Azam MA, Sampieri C, Ioppi A, Benzi P, Giordano GG, De Vecchi M, *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. Frontiers in Oncology. 2022; 12: 900451. https://doi.org/10.3389/fonc.2022.900451.

[10] Ganeshan V, Bidwell J, Gyawali D, Nguyen TS, Morse J, Smith MP, *et al*. Enhancing nasal endoscopy: Classification, detection, and segmentation of anatomic landmarks using a convolutional neural network. International Forum of Allergy & Rhinology. 2024; 14: 1521–1524. https://doi.org/10.1002/alr.23384.

[11] Mao Z, Das A, Islam M, Khan DZ, Williams SC, Hanrahan JG, *et al*. PitSurgRT: real-time localization of critical anatomical structures in endoscopic pituitary surgery. International Journal of Computer Assisted Radiology and Surgery. 2024; 19: 1053–1060. https://doi.org/10.1007/s11548-024-03094-2.

[12] Wang Z, Liu Z, Yu J, Gao Y, Liu M. Multi-scale nested UNet with transformer for colorectal polyp segmentation. Journal of Applied Clinical Medical Physics. 2024; 25: e14351. https://doi.org/10.1002/acm2.14351.

[13] Madani A, Namazi B, Altieri MS, Hashimoto DA, Rivera AM, Pucher PH, *et al*. Artificial Intelligence for Intraoperative Guidance: Using Semantic Segmentation to Identify Surgical Anatomy During Laparoscopic Cholecystectomy. Annals of Surgery. 2022; 276: 363–369. https://doi.org/10.1097/SLA.0000000000004594.

[14] Guanbin Li, Yizhou Yu. Visual Saliency Detection Based on Multiscale Deep CNN Features. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society. 2016; 25: 5012–5024. https://doi.org/10.1109/TIP.2016.2602079.

[15] Yan Z, Jagadeesh V, DeCoste D, di W, Piramuthu R. HD CNN: hierarchical deep convolutional neural network for image classification.

2014. Available at: https://arxiv.org/abs/1410.0736 (Accessed: 10 June 2025).

[16] Dubey AK, Chabert GL, Carriero A, Pasche A, Danna PSC, Agarwal S, *et al*. Ensemble Deep Learning Derived from Transfer Learning for Classification of COVID-19 Patients on Hybrid Deep-Learning-Based Lung Segmentation: A Data Augmentation and Balancing Framework. Diagnostics (Basel, Switzerland). 2023; 13: 1954. https://doi.org/10.3390/diagnostics13111954.

[17] Sanford TH, Zhang L, Harmon SA, Sackett J, Yang D, Roth H, *et al*. Data Augmentation and Transfer Learning to Improve Generalizability of an Automated Prostate Segmentation Model. AJR. American Journal of Roentgenology. 2020; 215: 1403–1410. https://doi.org/10.2214/AJR.19.22347.

[18] Alfieri A, Jho HD. Endoscopic endonasal approaches to the cavernous sinus: surgical approaches. Neurosurgery. 2001; 49: 354–354–60; discussion 360–2. https://doi.org/10.1097/00006123-200108000-00017.

[19] Silveira-Bertazzo G, Manjila S, Carrau RL, Prevedello DM. Expanded endoscopic endonasal approach for extending suprasellar and third ventricular lesions. Acta Neurochirurgica. 2020; 162: 2403–2408. https://doi.org/10.1007/s00701-020-04368-9.

[20] Kasemsiri P, Carrau RL, Ditzel Filho LFS, Prevedello DM, Otto BA, Old M, *et al*. Advantages and limitations of endoscopic endonasal approaches to the skull base. World Neurosurgery. 2014; 82: S12–21. https://doi.org/10.1016/j.wneu.2014.07.022.

[21] Ishii Y, Tahara S, Teramoto A, Morita A. Endoscopic endonasal skull base surgery: advantages, limitations, and our techniques to overcome cerebrospinal fluid leakage: technical note. Neurologia Medico-chirurgica. 2014; 54: 983–990. https://doi.org/10.2176/nmc.st.2014-0081.

[22] Gribaudo M, Piazzolla P, Porpiglia F, Vezzetti E, Violante MG. 3D augmentation of the surgical video stream: Toward a modular approach. Computer Methods and Programs in Biomedicine. 2020; 191: 105505. https://doi.org/10.1016/j.cmpb.2020.105505.

[23] Sampieri C, Baldini C, Azam MA, Moccia S, Mattos LS, Vilaseca I, *et al*. Artificial Intelligence for Upper Aerodigestive Tract Endoscopy and Laryngoscopy: A Guide for Physicians and State-of-the-Art Review. Otolaryngology–head and Neck Surgery: Official Journal of American Academy of Otolaryngology-Head and Neck Surgery. 2023; 169: 811–829. https://doi.org/10.1002/ohn.343.

[24] Ding H, Cen Q, Si X, Pan Z, Chen X. Automatic glottis segmentation for laryngeal endoscopic images based on U-Net. Biomedical Signal Processing and Control. 2022; 71: 103116. https://doi.org/10.1016/j.bspc.2021.103116.

[25] Namvar M, Iranmehr A, Fathi MR, Sadrhosseini SM, Tabari A, Shirzad N, *et al*. Complications in Endoscopic Endonasal Pituitary Adenoma Surgery: An Institution Experience in 310 Patients. Journal of Neurological Surgery. Part B, Skull Base. 2022; 84: 255–265. https://doi.org/10.1055/a-1838-5897.

[26] Staartjes VE, Togni-Pogliorini A, Stumpo V, Serra C, Regli L. Impact of intraoperative magnetic resonance imaging on gross total resection, extent of resection, and residual tumor volume in pituitary surgery: systematic review and meta-analysis. Pituitary. 2021; 24: 644–656. https://doi.org/10.1007/s11102-021-01147-2.

[27] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ (Clinical Research Ed.). 2021; 372: n71. https://doi.org/10.1136/bmj.n71.

[28] Rampinelli V, Paderno A, Conti C, Testa G, Modesti CL, Agosti E, *et al*. Artificial intelligence for automatic detection and segmentation of nasal polyposis: a pilot study. European Archives of Oto-rhino-laryngology: Official Journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS): Affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery. 2024; 281: 5815–5821. https://doi.org/10.1007/s00405-024-08809-4.

[29] Phoommanee N, Andrews PJ, Leung TS. Grade classification of nasal obstruction from endoscopy videos using machine learning. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference. 2024; 2024: 1–4. https://doi.org/10.1109/EMBC53108.2024.10781696.

[30] Lei J, Yang W, Yang R. A Deep Learning Method for Automated Site Recognition of Nasopharyngeal Endoscopic Images. Journal of Medical and Biological Engineering. 2025; 45: 240–251. https://doi.org/10.1007/s40846-025-00936-5.

[31] Bidwell J, Gyawali D, Morse J, Ganeshan V, Nguyen T, McCoul ED. Real-time augmentation of diagnostic nasal endoscopy video using AI-enabled edge computing. International Forum of Allergy & Rhinology. 2025; 15: 191–194. https://doi.org/10.1002/alr.23458.

[32] Levi L, Ye K, Fieux M, Renteria A, Lin S, Xing L, *et al*. Machine Learning of Endoscopy Images to Identify, Classify, and Segment Sinonasal Masses. International Forum of Allergy & Rhinology. 2025; 15: 524–535. https://doi.org/10.1002/alr.23525.

[33] Phoommanee N, Andrews PJ, Leung TS. Segmentation of endoscopy images of anterior nasal cavity using deep learning. In Chen W, Astley SM (eds.) Progress in Biomedical Optics and Imaging - Proceedings of SPIE (pp.1292706). 2024. https://doi.org/10.1117/12.2691427.

[34] Yue Y, Zeng X, Lin H, Xu J, Zhang F, Zhou K, *et al*. A deep learning based smartphone application for early detection of nasopharyngeal carcinoma using endoscopic images. NPJ Digital Medicine. 2024; 7: 384. https://doi.org/10.1038/s41746-024-01403-2.

[35] Liu X, Gong W, Chen X, Li Z, Liu Y, Wang L, *et al*. Vision-language foundation model for generalizable nasal disease diagnosis using unlabeled endoscopic records. Pattern Recognition. 2025; 165: 111646. https://doi.org/10.1016/j.patcog.2025.111646.

[36] Wang W, Jin Z, Liu X, Chen X. NaMA-Mamba: Foundation model for generalizable nasal disease detection using masked autoencoder with Mamba on endoscopic images. Computerized Medical Imaging and Graphics: the Official Journal of the Computerized Medical Imaging Society. 2025; 122: 102524. https://doi.org/10.1016/j.compmedimag.2025.102524.

[37] Yui R, Takahashi M, Noda K, Yoshida K, Sakurai R, Ohira S, *et al*. Preoperative prediction of sinonasal papilloma by artificial intelligence using nasal video endoscopy: a retrospective study. Scientific Reports. 2023; 13: 12439. https://doi.org/10.1038/s41598-023-38913-0.

[38] Xu X, Yun B, Zhao Y, Jin L, Zong Y, Yu G, *et al*. Neoplasms in the Nasal Cavity Identified and Tracked with an Artificial Intelligence-Assisted Nasal Endoscopic Diagnostic System. Bioengineering (Basel, Switzerland). 2024; 12: 10. https://doi.org/10.3390/bioengineering12010010.

[39] 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 Communications (London, England). 2018; 38: 59. https://doi.org/10.1186/s40880-018-0325-9.

[40] 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. https://doi.org/10.1016/j.compbiomed.2022.105725.

[41] Das A, Khan DZ, Williams SC, Hanrahan JG, Borg A, Dorward NL, *et al*. A Multi-task Network for Anatomy Identification in Endoscopic Pituitary Surgery. Lecture Notes in Computer Science. 2023; 14228: 472–482. https://doi.org/10.1007/978-3-031-43996-4_45.

[42] Kwon KW, Park SH, Lee DH, Kim DY, Park IH, Cho HJ, *et al*. Deep learning algorithm for the automated detection and classification of nasal cavity mass in nasal endoscopic images. PloS One. 2024; 19: e0297536. https://doi.org/10.1371/journal.pone.0297536.

[43] Tai J, Han M, Choi BY, Kang SH, Kim H, Kwak J, *et al*. Deep learning model for differentiating nasal cavity masses based on nasal endoscopy images. BMC Medical Informatics and Decision Making. 2024; 24: 145. https://doi.org/10.1186/s12911-024-02517-z.

[44] He R, Jie P, Hou W, Long Y, Zhou G, Wu S, *et al*. Real-time artificial intelligence-assisted detection and segmentation of nasopharyngeal carcinoma using multimodal endoscopic data: a multi-center, prospective study. EClinicalMedicine. 2025; 81: 103120. https://doi.org/10.1016/j.eclinm.2025.103120.

[45] Mao Z, Das A, Khan DZ, Williams SC, Hanrahan JG, Stoyanov D, *et al*. ConsisTNet: a spatio-temporal approach for consistent anatomical localization in endoscopic pituitary surgery. International Journal of Computer Assisted Radiology and Surgery. 2025; 20: 1239–1248. https://doi.org/10.1007/s11548-025-03369-2.

[46] 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: 1637–1646. https://doi.org/10.1002/alr.22854.

[47] Chen J, Liu Q, Wei Z, Luo X, Lai M, Chen H, *et al*. ETU-Net: efficient Transformer and convolutional U-style connected attention segmentation network applied to endoscopic image of epistaxis. Frontiers in Medicine. 2023; 10: 1198054. https://doi.org/10.3389/fmed.2023.1198054.

[48] Bi M, Zheng S, Li X, Liu H, Feng X, Fan Y, *et al*. MIB-ANet: A novel multi-scale deep network for nasal endoscopy-based adenoid hypertrophy grading. Frontiers in Medicine. 2023; 10: 1142261. https://doi.org/10.3389/fmed.2023.1142261.

[49] Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. European Journal of Epidemiology. 2010; 25: 603–605. https://doi.org/10.1007/s10654-010-9491-z.