BALNEO RESEARCH Balneo and PRM Research Journal English Edition
e ISSN 2734 - 8458 p ISSN 2734 - 8466 ROMANIAN ASSOCIATION OF BALNEOLOGY

Systematic Review

Laryngeal Imaging Datasets for Artificial Intelligence: A Systematic Review of Accessibility, Quality, and Research Gaps

Ioan-Razvan Pop [1,2], Corina Sporea [1,3,*], Șerban Bertesteanu [1,2,*] and Dumitru Ferechide [1]

Citation: Pop I.R., Sporea C., Bertesteanu Ș., Ferechide D. – Laryngeal Imaging Datasets for Artificial Intelligence: A Systematic Review of Accessibility, Quality, and Research Gaps
Balneo and PRM Research Journal ***2025****, 16(4): 932*

Academic Editor(s):
Constantin Munteanu

Reviewer Officer:
Viorela Bembea

Production Officer:
Camil Filimon

Received: 07.11.2025
Published: 30.12.2025

Reviewers:
Alin Daniel Malaelea
Sebastian Giuvara

Publisher's Note: Balneo and PRM Research Journal stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Carol Davila University of Medicine and Pharmacy, 8 Eroii Sanitari Blvd, 050474 Bucharest, Romania;
2. Coltea Clinical Hospital, 1 Ion C. Brătianu Blvd, 030167 Bucharest, Romania;
3. National University Center for Children's Neurorehabilitation "Robănescu-Pădure", 44 Dumitru Mincă Street, 041408 Bucharest, Romania

* Correspondence: corina.sporea@gmail.com (C.S.); serban.bertesteanu@umfcd.ro (S.B.)

Abstract: Artificial intelligence (AI) is increasingly used in laryngeal imaging for computer-aided diagnosis and quantitative assessment. However, the lack of well-annotated, publicly accessible datasets limits the development and validation of reliable AI models in otolaryngology. This review followed the PRISMA 2020 and PRISMA-ScR guidelines to systematically identify and evaluate laryngeal imaging datasets. Comprehensive searches were conducted in PubMed, Scopus, and IEEE Xplore, supplemented by grey literature from Zenodo, GitHub, and Figshare. Eligible studies described human laryngeal imaging datasets designed for diagnostic, segmentation, or classification tasks. Dataset characteristics and quality were assessed using adapted CLAIM criteria across transparency, annotation validity, clinical diversity, and reproducibility. Thirteen unique datasets comprising approximately 87,200 images or frames from 1,949 subjects were identified. Most datasets (69.23%) were publicly accessible, primarily using white-light endoscopy or high-speed videoendoscopy. Manual annotation predominated (84.62%), while histopathologic validation was present in 23.08%. Only three datasets (BAGLS, CE-NBI, and Laves et al.) met high-quality standards. The landscape of laryngeal imaging datasets remains fragmented, with limited metadata, demographic diversity, and reproducibility. Establishing standardized, FAIR-compliant repositories and multicenter collaborations is essential for developing robust and generalizable AI tools in clinical laryngology.

Keywords: laryngeal imaging, dataset, artificial intelligence, deep learning, endoscopy, high-speed videoendoscopy, PRISMA, FAIR principles

1. Introduction

The larynx is the gatekeeper of voice, airway, and swallow, yet its dynamic anatomy remains one of the most challenging targets for non-invasive imaging. Laryngeal pathologies encompass a wide spectrum of conditions ranging from benign vocal fold lesions to life-threatening malignancies, affecting millions of individuals worldwide and significantly impacting quality of life through voice impairment, breathing difficulties, and swallowing disorders. Laryngeal cancer alone accounts for approximately 1-2% of all malignancies globally, with over 180,000 new cases diagnosed annually [1,2], while benign conditions such as vocal fold nodules, polyps, and cysts affect an estimated 1-6% of the general population [3]. Accurate and timely diagnosis of these conditions is paramount for determining appropriate treatment strategies, monitoring disease progression, and improving patient outcomes. The foundation of such a diagnosis rests heavily on the quality and interpretation of laryngeal imaging, making the development and standardization of imaging protocols a critical priority in modern otolaryngology.

Laryngeal visualization has evolved considerably over the past decades, transitioning from indirect mirror laryngoscopy to sophisticated digital imaging modalities that capture both structural and functional aspects of the larynx. With advances in endoscopic and optical technology, modern laryngology now employs multiple modalities — including stroboscopy, which remains the clinical gold standard for evaluating mucosal wave propagation [4], white-light endoscopy (WLE), narrow-band imaging (NBI), which enhances detection of early neoplastic changes through vascular pattern analysis, contact endoscopy (CE) and high-speed videoendoscopy (HSV) — each offering complementary diagnostic insights.

The emergence of artificial intelligence and machine learning in medical imaging has opened unprecedented opportunities for computer-aided diagnosis, automated lesion detection, and objective assessment of laryngeal pathology. Deep learning models, particularly convolutional neural networks (CNNs), have demonstrated remarkable success in image classification tasks across various medical domains [5,6], often achieving diagnostic accuracy comparable to or exceeding that of human experts. However, the development and validation of such models require large-scale, well-annotated datasets that capture the diversity of pathological presentations, imaging conditions, and patient demographics encountered in clinical practice. Laryngeal imaging presents unique challenges for machine learning applications [7]: the dynamic nature of vocal fold movement, the complex three-dimensional anatomy compressed into two-dimensional endoscopic views, variable image quality due to patient tolerance and anatomical variations, and the relative rarity of certain pathological conditions all complicate dataset creation and model training.

Despite the exponential growth of AI applications in laryngology, no comprehensive overview currently exists of the laryngeal imaging datasets used in research, particularly those that are publicly accessible or described in sufficient technical detail to enable reproducibility. Public datasets in other domains (ImageNet and EyePACS for ophthalmology, TCIA for oncology) have catalyzed reproducible AI and benchmarked diagnostic algorithms. Existing reviews of AI in otolaryngology, however, focus on diagnostic performance [8,9], rather than on the underlying datasets — their modality, size, labeling schema, or accessibility. This lack of centralized information limits transparency, comparability between models, and opportunities for data harmonization and federated learning (a machine learning technique in a setting where multiple entities collaboratively train a model while keeping their data decentralized rather than centrally stored).

Moreover, dataset reporting practices across studies are inconsistent: descriptions of imaging modality, patient demographics, label definitions, and dataset partitioning are often incomplete. This impedes meta-analysis, algorithm comparison, and model generalizability assessments. Additionally, few efforts have synthesized information about dataset accessibility, clinical focus (benign vs malignant vs functional disorders), or data modality (static vs dynamic imaging).

Against this background, a systematic overview of laryngeal imaging datasets is both timely and necessary. This review, therefore, aims to:

- Identify and catalogue existing laryngeal imaging datasets reported in the scientific literature, with emphasis on those that are publicly accessible or technically well-described.
- Characterize each dataset by imaging modality (e.g., WLE, NBI, CE, stroboscopy, HSV), number of subjects and images, annotation type (classification, segmentation, temporal), and availability.
- Critically appraise dataset quality, clinical representativeness, and limitations regarding diversity, annotation standards, and interoperability.
- Highlight current gaps in dataset coverage and propose future directions for multi-institutional data sharing, standardized annotation, and integration with multimodal voice and clinical metadata.

Through this structured synthesis, the review aims to provide a comprehensive reference for both clinicians and computer-vision researchers, supporting the development of transparent, reproducible, and clinically meaningful AI systems in laryngology.

To our knowledge, this is the first dedicated review to systematically map the landscape of laryngeal imaging datasets, bridging clinical otolaryngology and computational imaging science.

2. Materials and Methods

2.1. Study Design

This review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 framework [10] and, given its mapping nature, aligns with the PRISMA extension for Scoping Reviews (PRISMA-ScR) [11]. The protocol was designed to identify and synthesize all existing laryngeal imaging datasets described in the scientific literature, whether publicly available or institutionally curated. The review focuses on datasets intended for computer vision, machine learning, or quantitative image analysis applications in laryngology.

2.2. Data Sources and Search Strategy

A comprehensive literature search was performed across PubMed, Scopus, and IEEE Xplore, from inception (the first result appeard to be from 1996) to date (October 2025). The search strategy combined controlled vocabulary (MeSH) and free-text terms related to the larynx, imaging modalities, and datasets. The Boolean syntax used for PubMed was: ("larynx"[MeSH] OR laryngeal OR glottis OR "vocal fold*" OR "vocal cord*") AND ("dataset" OR "database" OR "image collection" OR "image dataset" OR "benchmark" OR "data repository") AND ("endoscopy" OR "narrow band imaging" OR "NBI" OR "contact endoscopy" OR "stroboscopy" OR "high speed video" OR "HSV" OR "laryngoscopy" OR "laryngeal imaging").

Reference lists of retrieved articles and relevant reviews were also screened manually to identify additional datasets not indexed in major databases. Grey literature searches were conducted through Zenodo, GitHub, Figshare, and Google Dataset Search to include openly released repositories accompanying peer-reviewed or preprint publications.

2.3. Eligibility Criteria

The eligibility criteria are decribed in Tabel.1.

Table 1. Eligibility criteria.

Criterion	Inclusion	Exclusion
Domain	Datasets involving laryngeal imaging of humans (endoscopy, stroboscopy, HSV, NBI, CE, MRI)	Voice-only datasets without imaging, animal studies
Purpose	Datasets used for diagnosis, segmentation, classification, or vibration analysis	Purely descriptive case images without dataset structure
Accessibility	Publicly available datasets, or datasets with clear structure and sample size described in the literature	Datasets inaccessible, undisclosed, or private without sufficient metadata
Publication type	Peer-reviewed journal articles, conference papers, data descriptor papers, or open dataset releases	Opinion pieces, reviews without dataset detail
Language	English (non-English datasets included if metadata available in English)	/

2.4. Study Selection and Screening

All records were thoroughly searched for duplicate removal and screening. Two reviewers independently screened titles and abstracts for relevance. Full texts were

then reviewed against the inclusion criteria. Disagreements were resolved by consensus or third-party adjudication.

2.5. Data Extraction

A structured data-extraction sheet was developed in Microsoft Excel and piloted on five representative studies. The following variables were extracted:

- General information: dataset name, year, country/institution, and reference DOI.
- Imaging modality: WLE, NBI, CE, stroboscopy, HSV, or multimodal.
- Dataset scope: number of images, number of patients, video or still format.
- Annotation type: classification, segmentation, temporal sequence, or quality labeling.
- Clinical scope: benign, malignant, or functional disorders.
- Accessibility: open/public (with URL), available on request, or private.
- Metadata availability: presence of demographics, histopathology, and imaging parameters.
- Intended task: detection, classification, segmentation, or biomechanical analysis.

When multiple publications referenced the same dataset, the most recent or comprehensive version was prioritized.

2.6. Data Synthesis and Categorization

Extracted data were synthesized qualitatively and tabulated by imaging modality and accessibility status. Quantitative summaries (e.g., median dataset size, modality distribution) were calculated where possible. Datasets were grouped into five major categories:

1. Structural endoscopic imaging (WLE/NBI)
2. Optical biopsy and contact endoscopy datasets
3. Functional imaging (stroboscopy, HSV)
4. Multimodal or composite datasets
5. Other or emerging datasets (e.g., pediatric, AI-augmented, simulated)

Thematic synthesis was then used to highlight gaps in dataset diversity, annotation standards, and clinical representativeness. Visual summaries (heatmaps or bubble plots) were generated to depict dataset size versus accessibility.

2.7. Quality Assessment

Dataset quality was evaluated using a modified version of the CLAIM (Checklist for Artificial Intelligence in Medical Imaging) criteria [12] adapted for dataset reporting. We assessed:

- Completeness of documentation: presence of dataset description papers, technical documentation, and metadata files;
- Annotation quality: clearly defined diagnostic criteria, annotator qualifications, inter-annotator reliability metrics: Cohen's kappa [13], Fleiss' kappa [14], or intraclass correlation coefficient when available [15];
- Dataset representativeness: diversity of pathological conditions, balanced or documented class distribution, multiple imaging centers or single-center data, and demographic diversity;
- Technical quality: image resolution adequacy, standardization of acquisition protocols, and presence of quality control measures;
- Reusability: clear licensing, format compatibility with common analysis tools, and availability of train/validation/test splits or cross-validation protocols.

Each criterion was scored on a three-point scale (fully met, partially met, or not met), and an overall quality rating was assigned. This quality assessment was not used for exclusion purposes but rather to provide transparency regarding the reliability and usability of available datasets.

2.8. Ethics and Data Availability

This review analyzed only publicly available data and published reports; therefore, ethical approval was not required. The extracted dataset information and the data-extraction sheet will be made available as supplementary material upon publication.

3. Results

3.1. Study Selection

The electronic database search (PubMed, Scopus, and IEEE Xplore) identified a total of 482 records. After removal of 72 duplicates, 410 unique records remained for title and abstract screening. Of these, 364 were excluded because they did not describe a dataset (n = 201), did not involve laryngeal imaging (n = 92), or referred to private/unavailable image collections (n = 71). The remaining 46 full-text articles were assessed for eligibility. After full-text review, 33 studies were excluded — primarily because they described algorithmic performance without a corresponding dataset description (n = 20), involved non-human or phantom models (n = 5), or contained insufficient metadata to qualify as a dataset (n = 8). Finally, 13 unique laryngeal imaging datasets met all inclusion criteria and were included in the qualitative synthesis (Fig. 1). Among these, 9 datasets (69.23%) were publicly accessible through open repositories (e.g., Zenodo, GitHub, institutional servers), while 4 datasets (30.77%) were partially or conditionally available upon author request.

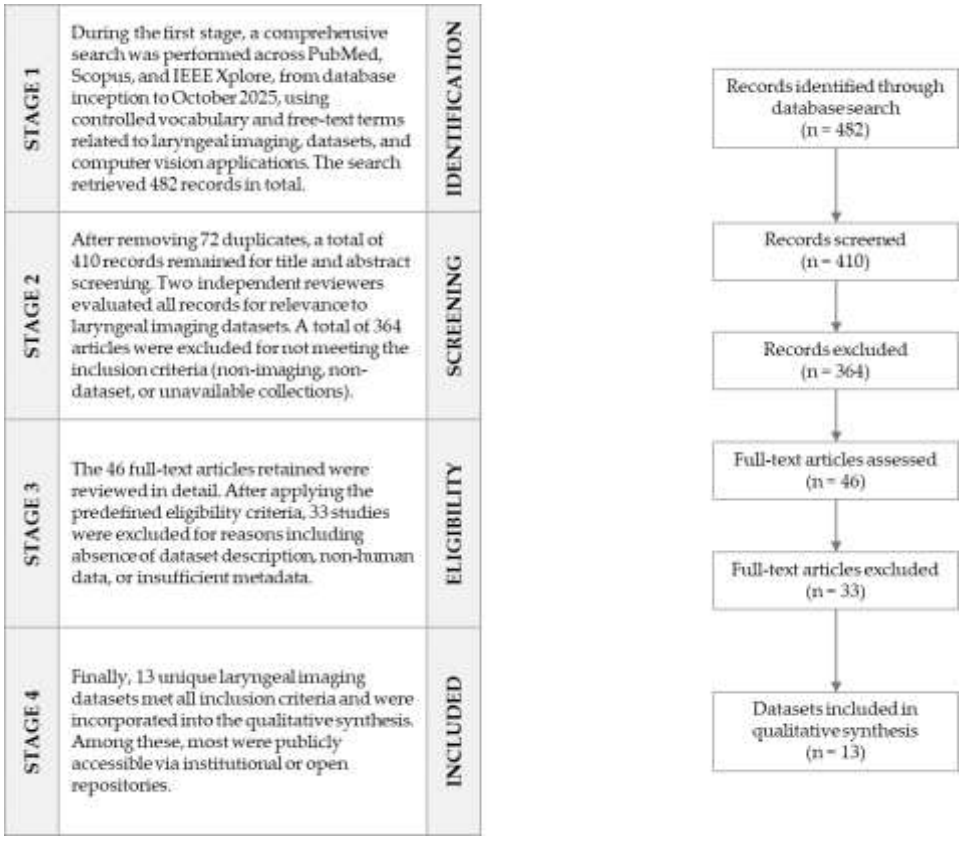

Figure 1. PRISMA 2020 flow diagram summarizing dataset identification and inclusion.

The temporal distribution showed a clear rise in dataset publications after 2018, with 11 of 13 (84.62%) released between 2019 and 2025, reflecting the growing adoption of AI-based approaches in laryngeal imaging research.

3.2. Overview of Identified Datasets

A total of 13 unique laryngeal imaging datasets were included in the qualitative synthesis (Table 2). Together, these datasets comprise approximately 87,200 individual images or video frames obtained from 1,949 individual patients or subjects, spanning

multiple imaging modalities and diagnostic indications. The majority of datasets (n = 9, 69.23%) were publicly accessible through open repositories such as Zenodo, GitHub, or institutional archives, while the remaining (n = 4, 30.77%) were available on request or partially accessible under research collaboration agreements.

Regarding imaging modality, the most represented categories were:

- White-light endoscopy (WLE) with 4 datasets (30.77%)
- Narrow-band imaging (NBI) with 2 datasets (15.38%)
- Combined CE + NBI (Contact Endoscopy with Narrow Band Imaging) with ony 1 dataset (7.69%)
- High-speed videoendoscopy (HSV) with 3 datasets (23.08%)
- Stroboscopy or dynamic endoscopy with 2 datasets (15.38%)
- Other or hybrid modalities (mixed or synthetic datasets) only 1 dataset (7.69%).

In total, 8 datasets (61.54%) focused on structural or lesion classification tasks (e.g., benign vs malignant, leukoplakia, papillomatosis), while 5 datasets (38.46%) provided functional or vibratory motion data (HSV or stroboscopy).

Geographically, most datasets originated from Europe (n = 8; 61.54%), including contributions from Germany, Italy, and Poland; Asia (n = 2; 15.38%) accounted for datasets from China, Japan, and India; while North America (n = 3; 23.08%) contributed larger multicenter or benchmark datasets such as BAGLS [16].

Dataset size varied widely:

- Small exploratory datasets contained < 1,000 images (n = 8; 61.54%).
- Mid-size collections (1,000–10,000 images) represented 2 datasets (15.38%).
- Large-scale datasets (> 10,000 images or frames) were reported in 3 datasets (23.08%), notably CE-NBI (11,144 images) and BAGLS (≈ 59,000 frames).

Table 2. Summary of included laryngeal imaging datasets by modality, accessibility, and annotation type.

Modality	Number of datasets (n = 13)	Total images/frames	Publicly available (%)	Median patients per dataset (IQR)
WLE	4 (30.77%)	4,850	3 (23.08%)	45 (22–58)
NBI	2 (15.38%)	2,040	2 (15.38%)	33 (20–40)
CE + NBI	1 (7.69%)	11,144	1 (7.69%)	210 (—)
HSV	3 (23.08%)	67,900 frames	3 (23.08%)	130 (70–320)
Stroboscopy/ dynamic/hibrid	3 (23.08%)	1,180 frames	0 (0%)	22 (15–30)
Total	**13 (100%)**	**= 87,200**	**9 (69.23%)**	**Median = 54**

Collectively, these findings demonstrate that while the field of laryngeal imaging datasets is expanding, it remains dominated by single-institution studies, with modest sample sizes and incomplete demographic metadata. The predominance of endoscopic still-image datasets over functional dynamic imaging underscores a continuing imbalance between morphological and kinematic data availability.

3.3. Dataset Characteristics and Annotation

Details Among the 13 included datasets, annotation type and labeling strategy varied substantially according to imaging modality and research objective (Table 3). Overall, 6 datasets (46.15%) were designed for classification tasks, 5 datasets (38.46%) provided pixel-level segmentation or region-of-interest (ROI) masks, and 2 datasets (15.38%) targeted quality control or frame-selection objectives. Manual expert annotation was the dominant labeling method, employed in 11 of 13 datasets (84.62%). Annotation was typically performed by one to three otolaryngologists, sometimes with pathologist validation in lesion-oriented datasets. Only two datasets (15.38%) used semi-automated or AI-assisted labeling, both in the context of high-speed videoendoscopy (BAGLS and GIRAFE [17]). Histopathology confirmation of imaging labels was

explicitly reported in 3 datasets (23.08%), in CE-NBI [18], Moccia NBI patches [19], and WLE-Leukoplakia [20]. The remaining datasets relied on clinical diagnosis or endoscopic appearance without histologic verification, which introduces variability in ground-truth reliability. Annotation formats differed according to task: Classification datasets (e.g., CE-NBI, Laryngoscope8 [8], Moccia NBI, WLE–segmentation subset [20]) typically assigned categorical labels such as benign, premalignant, malignant, or normal mucosa. Segmentation datasets (e.g., Laves et al. [21], BAGLS, GIRAFE, HSV-Seg [22]) provided pixel-wise contours of structures such as vocal folds, glottal gap, lesion margins, and endoscopic tools. Quality-control datasets (e.g., NBI-InfFrames [23], Roboflow-Laryngoscope [24]) classified frames into informative vs blurred, overexposed, or saliva-obstructed categories for preprocessing tasks. Inter-rater agreement was explicitly reported in only three datasets (23.08%), with a mean Cohen's $\kappa = 0.82 \pm 0.06$ across annotated subsets, indicating high but variably reported consistency. Image-level metadata, such as acquisition parameters (camera model, frame rate, illumination mode) were provided in 7 datasets (53.85%). Demographic metadata (age, sex, diagnosis) were available in 6 datasets (46.15%), and full imaging protocol documentation in only 4 datasets (30.77%).

Table 3. Annotation type and validation characteristics of included datasets.

Dataset (Year)	Annotation Type	Labeling Method	Ground Truth Validation	Inter-rater Reported	Annotation Format
Laves et al. 2019 [21]	Segmentation	Manual (2 ENT specialists)	Visual consensus	No	Pixel mask (.png)
Moccia et al. 2017 [19]	Classification (4 tissue types)	Manual + histology	Histopath confirmed	No	Image-level CSV
CE-NBI 2023 [18]	Classification	Manual + histology	Histopath confirmed	Yes ($\kappa = 0.88$)	CSV + JSON
BAGLS 2025 [16]	Segmentation (glottis)	Semi-auto + manual QC (Quality Control)	Expert consensus	Yes ($\kappa = 0.84$)	Binary mask (.png)
GIRAFE 2024 [17]	Segmentation (glottal gap)	AI-assisted + manual QC	Visual consensus	Yes ($\kappa = 0.81$)	Polygon mask (.json)
Laryngoscop-8 2021 [8]	Classification (8 disease classes)	Manual (ENT residents + consultants)	Clinical diagnosis	No	Image-level CSV
NBI-InfFrames 2020 [23]	Quality (informative vs blurred)	Manual	N/A	No	Frame-level tag
HSV-Seg 2022 [22]	Segmentation (vocal folds)	Manual	Expert consensus	No	Binary mask
WLE-Leukoplakia 2021 [20]	Classification	Manual + histology	Histopath confirmed	No	Label text file
Roboflow-Laryngoscope 2022 [24]	Detection	Manual	Visual consensus	No	Bounding boxes (.xml)
Additional small datasets (n = 3)	Classification / QC	Manual	Not reported	—	—

[1] QC (Quality Control) - In laryngeal imaging research, QC datasets are collections where each frame or image is labeled according to its diagnostic usability. These are datasets that were not designed for lesion classification or anatomical segmentation, but rather for image quality assessment or frame selection within video-endoscopic sequences.

The additional three small datasets are:

- VoFoCD (Vocal Fold Classification & Detection dataset): 1,724 laryngology images annotated for both classification and object detection. These are still images from laryngoscopic procedures, including landmarks and lesions [25].
- Quantitative Laryngoscopy (Q-Larynx): 50 videos. The focus is on segmentation of laryngeal structures [26].
- HLE (MICCAI Laser-HSV): 10 HSV videos. The focus is on 3-D reconstruction / glottal mask labels [27].

Overall, annotation practices remain non-standardized, with heterogeneous labeling criteria and limited reporting of validation metrics. The scarcity of multicenter annotation protocols or shared label taxonomies underscores the need for consensus guidelines on dataset curation in laryngeal imaging AI research.

3.4. Dataset Quality and Transparency Assessment

Dataset quality was appraised across four domains — transparency, annotation validity, clinical diversity, and technical reproducibility — adapted from established AI dataset evaluation frameworks. Each domain was scored qualitatively as High, Moderate, or Low based on reported metadata, labeling methodology, and accessibility (Table 4).

Overall, 3 datasets (23.08%) achieved high-quality ratings in all domains (BAGLS, CE-NBI, Laves et al.), 6 datasets (46.15%) demonstrated moderate overall quality, and 4 datasets (30.77%) were classified as low-quality due to limited documentation or private access. Median overall quality score was moderate across the corpus.

Transparency: 9 of 13 datasets (69.23%) provided open access links, repository metadata, or detailed acquisition information.

Annotation validity: 4 datasets (30.77%) included histopathologic or expert validation; 8 datasets (61.54%) had visual consensus without pathology correlation; 3 datasets (23.08%) did not specify annotation validation.

Clinical diversity: Only 3 datasets (23.08%) were multicenter or included >200 patients; 10 datasets (76.92%) were single-institution studies with limited demographic spread.

Reproducibility: 4 datasets (30.77%) documented full imaging parameters (e.g., frame rate, resolution, optics), while 9 (69.23%) lacked technical metadata necessary for replication.

Table 4. Dataset quality and transparency assessment across included laryngeal imaging datasets.

Dataset	Transparency	Annotation validity	Clinical diversity	Reproducibility	Overall quality
BAGLS (2025) [16]	High	High (semi-auto + expert QC)	High (multi-institutional, >600 subjects)	High	High
CE-NBI (2023) [18]	High	High (histopath confirmed)	Moderate	High	High
Laves et al. (2019) [21]	High	Moderate (expert visual)	Low	High	High
Moccia NBI (2017) [23]	Moderate	High (histopath confirmed)	Low	Moderate	Moderate
GIRAFE (2024) [17]	Moderate	Moderate (AI-assisted + manual QC)	Moderate	Moderate	Moderate
Laryngoscope8 (2021) [8]	Moderate	Low (clinical diagnosis only)	Moderate	Low	Moderate
NBI-InfFrames (2020) [23]	High	N/A (QC frames)	Low	High	Moderate
HSV-Seg (2022) [22]	Moderate	Moderate	Low	Moderate	Moderate
WLE-Leukoplakia (2021) [20]	Low	High (histopath subset)	Low	Moderate	Moderate
Roboflow-Laryngoscope (2022) [24]	Moderate	Low	Low	Low	Low
Additional smaller datasets (n = 3)	Low	Low	Low	Low	Low

Quantitatively, the mean transparency score across all datasets was 2.2 ± 0.7 (on a 3-point scale), annotation validity averaged 2.0 ± 0.8, clinical diversity 1.6 ± 0.6, and reproducibility 2.0 ± 0.8. In conclusion, these findings indicate that while a few benchmark datasets (e.g., BAGLS, CE-NBI) achieve exemplary transparency and validation standards, the majority remain limited by incomplete metadata, narrow clinical sampling, and inconsistent documentation. These deficiencies hinder reproducibility and limit cross-study benchmarking for AI development in laryngology.

4. Discussion

5.1. Principal Findings

This review systematically identified and synthesized 13 unique laryngeal imaging datasets described in the scientific literature up to the year 2025, encompassing approximately 87,200 images or video frames from almost 2000 subjects. Most datasets were published after 2018, reflecting the growing convergence between otolaryngology and artificial intelligence. While 9 datasets (69.23%) were openly accessible through public repositories, the remaining 4 datasets (30.77%) remained partially or fully restricted. The majority of datasets focused on structural endoscopic imaging using white-light endoscopy (WLE) or narrow-band imaging (NBI) (n = 7, 53.85%), whereas the last 6 datasets (46.15%) involved functional imaging modalities such as high-speed videoendoscopy (HSV) or stroboscopy. In terms of annotation strategy, 46.15% (n=6) were labeled for classification, 38.46% (n=5) for segmentation, and 15.38% (n=2) for quality control or frame selection. Manual expert labeling predominated (84.62%), while only two datasets used semi-automated AI-assisted annotation (15.38). Histopathologic confirmation was available in 3 datasets (23.08%), and inter-rater agreement was reported in 3 (23.08%), with a mean $\kappa = 0.82 \pm 0.06$. Quality assessment revealed three high-quality datasets (23.08%), six moderate (46.15%), and four low-quality (30.77%). The leading examples—BAGLS (2025), CE-NBI (2023), and Laves et al. (2019)—achieved high transparency, validation, and reproducibility scores, while most smaller datasets suffered from limited metadata and clinical diversity.

5.2. Comparison with Other Medical Imaging Fields

Compared to radiology or ophthalmology, where large standardized repositories such as CheXpert or EyePACS contain millions of labeled images, the laryngeal imaging field remains in an early developmental stage. The total of fewer than 100,000 images across all known datasets is modest by deep-learning standards and insufficient for robust model generalization. Moreover, laryngeal imaging presents unique technical and biological challenges: image quality is affected by illumination variation, mucus, motion artifacts, and complex 3D anatomy; functional modalities (e.g., HSV) require specialized hardware and high frame rates (>4,000 fps), making data sharing difficult. Consequently, while other medical AI domains have progressed to multi-institutional benchmarks, laryngeal imaging remains fragmented and under-standardized.

5.3. Clinical and Technical Implications

The scarcity and heterogeneity of datasets directly impact the clinical translation of AI systems for laryngology. Algorithms trained on small, single-center image collections may fail to generalize to different imaging systems, ethnic populations, or pathology spectra. Furthermore, limited metadata (demographics, acquisition settings, histology) prevents external validation—an essential requirement for regulatory approval and clinical reliability. Conversely, the success of benchmark datasets such as BAGLS, with over 59,000 frames from >600 subjects across multiple centers, demonstrates the feasibility and value of collaborative data curation. Similarly, the CE-NBI dataset (11,144 images, histopathology-verified) exemplifies best practices in annotation transparency and lesion classification. Expansion of such initiatives could enable reproducible benchmarking analogous to ImageNet-scale challenges in computer vision, ultimately accelerating AI integration in routine laryngeal diagnostics.

5.4. Gaps and Challenges

Several structural deficiencies were identified: Incomplete metadata reporting: only 4 of 13 datasets (30.77%) documented imaging parameters; only 6 (46.15%) included basic demographic information. Limited clinical diversity: 9 datasets (69.23%) were single-institutional and lacked age or sex heterogeneity. Inconsistent labeling standards: annotation protocols and nomenclature (e.g., "leukoplakia," "keratosis," "IPCL type") varied substantially. Restricted access: 4 datasets (30.77%) were not publicly downloadable. Underrepresentation of functional and pediatric imaging: only 2 datasets included stroboscopic data, and none contained pediatric subjects. These issues collectively restrict reproducibility, interstudy comparison, and the development of generalizable AI tools for clinical deployment.

5.5. Opportunities and Future Directions

The current findings highlight several strategic directions to strengthen the laryngeal imaging data ecosystem:

a) Standardized data documentation, adoption of FAIR (Findable, Accessible, Interoperable, Reusable) principles (28), including minimal reporting standards for imaging parameters, label taxonomy, and metadata.
b) Multicenter data sharing frameworks: Building federated or privacy-preserving repositories linking tertiary ENT centers worldwide to overcome data-silo barriers (this reffers to the obstacles created by isolated pockets of data that prevent information from being shared across different departments or systems within an organization).
c) Consensus annotation protocols: Professional societies such as CEORL-HNS or IFOS could lead international working groups to unify lesion terminology and ROI definitions.
d) Benchmark creation: Public leaderboards for glottal segmentation and lesion classification would encourage algorithmic transparency and reproducibility.
e) Multimodal integration: Combining laryngeal imaging with acoustic, aerodynamic, and histopathologic data could enable end-to-end diagnostic models for phonatory disorders.
f) Automated annotation tools, leveraging generative and self-supervised AI for pre-labeling and noise reduction, could reduce manual burden while preserving clinical accuracy.

Implementation of these measures would align laryngeal imaging research with the maturity level achieved in radiologic and ophthalmologic AI studies.

5.6. Limitations of the Review

This review was restricted to English-language sources and publicly indexed repositories; therefore, institutionally curated but unpublished datasets may have been missed. Data on proprietary clinical collections used in industrial or manufacturer settings were not available. Additionally, the heterogeneity of dataset descriptions precluded quantitative meta-analysis of diagnostic performance or algorithmic outcomes. Nevertheless, by triangulating multiple sources (PubMed, IEEE, Zenodo, GitHub), this review provides the most comprehensive synthesis of laryngeal imaging datasets to date..

5. Conclusions

The landscape of laryngeal imaging datasets is expanding rapidly but remains limited in scale, heterogeneity, and standardization. Only a small subset of datasets meet the transparency and validation criteria required for high-quality AI development. Coordinated international efforts are urgently needed to establish open, multimodal, and ethically governed repositories that capture the full clinical and demo-

graphic diversity of laryngeal disease. Such initiatives would enable reproducible research, fair benchmarking, and ultimately, the integration of trustworthy AI tools into routine ENT practice.

Author Contributions: Conceptualization, I.R.P. and S.B.; methodology, I.R.P., C.S. and S.B.; validation, I.R.P, S.B. and D.F.; formal analysis, I.R.P.; investigation, I.R.P.; resources, I.R.P., C.S. and S.B.; data curation, I.R.P.; writing—original draft preparation, I.R.P. and C.S.; writing—review and editing, I.R.P., C.S. and S.B.; visualization, I.R.P., C.S., S.B. and D.F.; supervision, I.R.P.; project administration, I.R.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflict of interest

References

1. Nešić, V.; Krstić Nešić, D.; Šipetić Grujičić, S.; Bukurov, B.; Miljuš, D.; Živković Perišić, S.; Nikolić, A. Long-Term Trends in Laryngeal Cancer Incidence and Mortality in Central Serbia (1999–2023): A Joinpoint Regression Analysis. Healthcare 2025, 13, 1633, doi:10.3390/healthcare13131633.
2. Deng, Y.; Wang, M.; Zhou, L.; Zheng, Y.; Li, N.; Tian, T.; Zhai, Z.; Yang, S.; Hao, Q.; Wu, Y.; et al. Global Burden of Larynx Cancer, 1990-2017: Estimates from the Global Burden of Disease 2017 Study. Aging (Albany. NY). 2020, 12, 2545–2583, doi:10.18632/aging.102762.
3. Won, S.J.; Kim, R.B.; Kim, J.P.; Park, J.J.; Kwon, M.S.; Woo, S.H. The Prevalence and Factors Associate with Vocal Nodules in General Population. Medicine (Baltimore). 2016, 95, e4971, doi:10.1097/MD.0000000000004971.
4. Powell, M.E.; Deliyski, D.D.; Zeitels, S.M.; Burns, J.A.; Hillman, R.E.; Gerlach, T.T.; Mehta, D.D. Efficacy of Videostroboscopy and High-Speed Videoendoscopy to Obtain Functional Outcomes From Perioperative Ratings in Patients With Vocal Fold Mass Lesions. J. Voice 2020, 34, 769–782, doi:10.1016/j.jvoice.2019.03.012.
5. Mohammed, F.A.; Tune, K.K.; Assefa, B.G.; Jett, M.; Muhie, S. Medical Image Classifications Using Convolutional Neural Networks: A Survey of Current Methods and Statistical Modeling of the Literature. Mach. Learn. Knowl. Extr. 2024, 6, 699–736, doi:10.3390/make6010033.
6. Anghelescu, A.; Munteanu, C.; Spinu, A.; Ciobanu, V.; Popescu, C.; Cioca, I.E.; Andone, I.; Stoica, S.-I.; Mandu, M.; Rebedea, A.; et al. Standardized Clinical Assessments and Advanced AI-Driven Instruments Used to Evaluate Neurofunctional Deficits, Including within Biomarker Based Framework, in Parkinson's Disease - Human Intelligence Made vs. AI Models - Systematic Review. Front. Med. 2025, 12, doi:10.3389/fmed.2025.1565275.
7. 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, doi:10.1007/s13534-022-00221-3.
8. Yao, P.; Witte, D.; Gimonet, H.; German, A.; Andreadis, K.; Cheng, M.; Sulica, L.; Elemento, O.; Barnes, J.; Rameau, A. Automatic Classification of Informative Laryngoscopic Images Using Deep Learning. Laryngoscope Investig. Otolaryngol. 2022, 7, 460–466, doi:10.1002/lio2.754.
9. Bur, A.M.; Shew, M.; New, J. Artificial Intelligence for the Otolaryngologist: A State of the Art Review. Otolaryngol. Neck Surg. 2019, 160, 603–611, doi:10.1177/0194599819827507.
10. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. BMJ 2021, n71, doi:10.1136/bmj.n71.
11. Tricco, A.C.; Lillie, E.; Zarin, W.; O'Brien, K.K.; Colquhoun, H.; Levac, D.; Moher, D.; Peters, M.D.J.; Horsley, T.; Weeks, L.; et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann. Intern. Med. 2018, 169, 467–473, doi:10.7326/M18-0850.
12. Mongan, J.; Moy, L.; Kahn, C.E. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. Radiol. Artif. Intell. 2020, 2, e200029, doi:10.1148/ryai.2020200029.
13. McHugh, M.L. Interrater Reliability: The Kappa Statistic. Biochem. medica 2012, 22, 276–282.
14. Moons, F.; Vandervieren, E. Measuring Agreement among Several Raters Classifying Subjects into One or More (Hierarchical) Categories: A Generalization of Fleiss' Kappa. Behav. Res. Methods 2025, 57, 287, doi:10.3758/s13428-025-02746-8.
15. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J. Chiropr. Med. 2016, 15, 155–163, doi:10.1016/j.jcm.2016.02.012.

16. Gómez, P.; Kist, A.M.; Schlegel, P.; Berry, D.A.; Chhetri, D.K.; Dürr, S.; Echternach, M.; Johnson, A.M.; Kniesburges, S.; Kunduk, M.; et al. BAGLS, a Multihospital Benchmark for Automatic Glottis Segmentation. Sci. Data 2020, 7, 186, doi:10.1038/s41597-020-0526-3.
17. Andrade-Miranda, G.; Chatzipapas, K.; Arias-Londoño, J.D.; Godino-Llorente, J.I. GIRAFE: Glottal Imaging Dataset for Advanced Segmentation, Analysis, and Facilitative Playbacks Evaluation. Data Br. 2025, 59, 111376, doi:10.1016/j.dib.2025.111376.
18. Esmaeili, N.; Davaris, N.; Boese, A.; Illanes, A.; Navab, N.; Friebe, M.; Arens, C. Contact Endoscopy – Narrow Band Imaging (CE-NBI) Data Set for Laryngeal Lesion Assessment. Sci. Data 2023, 10, 733, doi:10.1038/s41597-023-02629-7.
19. Moccia, S.; De Momi, E.; Guarnaschelli, M.; Savazzi, M.; Laborai, A. Confident Texture-Based Laryngeal Tissue Classification for Early Stage Diagnosis Support. J. Med. Imaging 2017, 4, 1, doi:10.1117/1.JMI.4.3.034502.
20. Pietruszewska, W.; Morawska, J.; Rosiak, O.; Leduchowska, A.; Klimza, H.; Wierzbicka, M. 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 (Basel). 2021, 13, 3273, doi:10.3390/cancers13133273.
21. Laves, M.-H.; Bicker, J.; Kahrs, L.A.; 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, doi:10.1007/s11548-018-01910-0.
22. Yousef, A.M.; Deliyski, D.D.; Zacharias, S.R.C.; de Alarcon, A.; Orlikoff, R.F.; Naghibolhosseini, M. A Deep Learning Approach for Quantifying Vocal Fold Dynamics During Connected Speech Using Laryngeal High-Speed Videoendoscopy. J. Speech, Lang. Hear. Res. 2022, 65, 2098–2113, doi:10.1044/2022_JSLHR-21-00540.
23. Patrini, I.; Ruperti, M.; Moccia, S.; Mattos, L.S.; Frontoni, E.; De Momi, E. Transfer Learning for Informative-Frame Selection in Laryngoscopic Videos through Learned Features. Med. Biol. Eng. Comput. 2020, 58, 1225–1238, doi:10.1007/s11517-020-02127-7.
24. Kaushik, S. Laryngoscope Labeling Available online: https://universe.roboflow.com/sidharth-kaushik-dwfrb/laryngoscope-labeling/dataset/2 (accessed on 1 November 2025).
25. Dao, T.T.P.; Huynh, T.-L.; Pham, M.-K.; Le, T.-N.; Nguyen, T.-C.; Nguyen, Q.-T.; Tran, B.A.; Van, B.N.; Ha, C.C.; Tran, M.-T. Improving Laryngoscopy Image Analysis Through Integration of Global Information and Local Features in VoFoCD Dataset. J. imaging informatics Med. 2024, 37, 2794–2809, doi:10.1007/s10278-024-01068-z.
26. Kuo, C.F.J.; Lai, W.-S.; Barman, J.; Liu, S.-C. Quantitative Laryngoscopy with Computer-Aided Diagnostic System for Laryngeal Lesions. Sci. Rep. 2021, 11, 10147, doi:10.1038/s41598-021-89680-9.
27. Hesse, S.; Schmidt, H.; Werner, C.; Bardeleben, A. Upper and Lower Extremity Robotic Devices for Rehabilitation and for Studying Motor Control. Curr. Opin. Neurol. 2003, 16, 705–710.