



A Multimodal Approach for Deep-Learning Classification of Vocal Fold Pathologies in Stroboscopy

Sruthi Surapaneni, BS^{1,2,*}, Rachel B. Kutler, BA^{1,3,*}, Sean A. Setzen, MD^{1,4,*}, Yeo Eun Kim, BA¹, Peter Yao, MD¹, Sana H. Siddiqui, MD^{1,3}, Michael J. Pitman, MD³, Lucian Sulica, MD¹, Olivier Elemento, PhD⁵, Pegah Khosravi, PhD^{6,7,**}, Anaïs Rameau, MD MSc MPhil MS FACS^{1,**}

¹Sean Parker Institute for the Voice, Department of Otolaryngology-Head and Neck Surgery, Weill Cornell Medical College, New York, New York, USA

²Michigan State University College of Human Medicine, Grand Rapids, MI, USA

³The Center for Voice and Swallowing, Department of Otolaryngology-Head and Neck Surgery, Columbia University Irving Medical Center, New York, NY, USA

⁴Department of Otolaryngology-Head and Neck Surgery, St. Luke's University Health Network, Bethlehem, Pennsylvania, USA

⁵Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY, USA

⁶Department of Clinical Sciences, College of Medicine, University of Central Florida, Orlando, FL, USA

⁷Institute for Artificial Intelligence, Department of Computer Science, University of Central Florida, Orlando, FL, USA

Abstract

Objective: To develop and validate a multimodal deep-learning classifier trained on stroboscopic image, voice, and clinicodemographic data, differentiating between three different vocal fold (VF) states: healthy (HVF), unilateral paralysis (UVFP), and VF lesions, including benign and malignant pathologies.

Methods: Patients with UVFP (n = 54), VF lesions (n = 42), and HVF (n = 41) were retrospectively identified. Image frames and voice samples were extracted from stroboscopic videos. Clinicodemographic variables were collected from the electronic health record. Patient-level data was independently divided into training (80%) and testing (20%). Visual features were extracted using a transformer DINOv2 and acoustic features were extracted using Librosa. All three feature modalities were combined using a custom multilayer perceptron. Unimodality

Corresponding Author: Anaïs Rameau, MD, MPhil, MS, Sean Parker Institute for the Voice, Department of Otolaryngology-Head and Neck Surgery, Weill Cornell Medical College, 240 E 59th Street, New York, NY 10022, anr2783@med.cornell.edu.

*Co-first authors

**Co-last authors

Conflict of Interest: Anaïs Rameau is an advisor for Sound Health Inc. and Pentax Medical.

Level of Evidence: 4

models using only image or only voice data were trained for comparison. Accuracy and F-1 scores were used to validate the models.

Results: On a hold-out test set, the multimodal classifier demonstrated stronger performance (76.9% accuracy) compared to the image classifier (61.5% accuracy) and audio classifier (65.4% accuracy). On an external dataset, the multimodal classifier accuracy dropped to 45%, though still an improvement compared to accuracies of 42% and 31% for the video-only and audio-only modalities, respectively.

Conclusions: In this proof-of-concept study, we successfully developed a multimodal dataset and classifier for VF pathology, demonstrating the potential of combining stroboscopic frames, voice and text data. The multimodal classifier achieved higher accuracy than the image-only model and audio-only models. Future models should validate these findings on larger datasets.

Keywords

artificial intelligence; transformer; multimodal

INTRODUCTION

Clinical evaluation requires the integration of multiple data sources, including a patient's history, physical examination findings, imaging studies, and laboratory results.^{1,2} Physicians must synthesize these diverse data to arrive at accurate diagnoses and guide treatment decisions. To streamline and standardize this process, artificial intelligence (AI) has been increasingly applied to support clinical decision-making. However, most current AI tools are designed for single tasks and rely on isolated data types.^{3,4} This narrow focus limits their applicability in the multifaceted context of real-world clinical care.

Multimodal AI refers to machine learning (ML) models capable of processing and integrating information from multiple modalities or types of data.⁵ These modalities can include text, images, audio, video, and other forms of sensory input. By combining complementary information from different sources, these models can compensate for noise, variability, or missing data in any single modality, thereby enhancing both robustness and predictive accuracy. In various medical specialties, multimodal AI has demonstrated promising results in both initial diagnosis and longer-term patient management.⁶⁻⁸

In otolaryngology, many prior studies have developed AI models based on convolutional neural networks (CNNs) in the analysis of laryngoscopic and endoscopic images and videos, including the classification of various VF pathologies.⁹⁻¹¹ This application of deep learning to endoscopic data, referred to as "videomics" as defined by Paderno et al., offers a promising opportunity to expand access to diagnostic evaluations, especially in low-resource settings.¹² Voice-based algorithms have also been explored, though with generally less robust performance.^{13,14} However, few prior studies have evaluated the integration of multiple modalities or compared the performance of multimodal models to unimodal classifiers.^{15,16}

This study aims to develop a prediction model for VF pathologies by integrating 3 distinct data modalities: stroboscopic video, audio recordings of voice samples extracted from

these videos, and clinicodemographic data from patient electronic health records. To our knowledge, this is a novel application of multimodal classification using 3 modalities to classify 3 types of VF pathology. By leveraging the complementary strengths of each data type, our multimodal approach seeks to enhance the classification of 3 categories of VF states, healthy vocal folds (HVF), unilateral vocal fold paralysis (UVFP), and VF lesions, including both benign and malignant pathologies, as compared to single modality models. This novel integrated approach more closely approximates the multifaceted evaluation performed by clinicians, potentially offering improved diagnostic accuracy with potential for clinical utility in laryngological practice. The objective was to explore the utility and benefit of multimodal classification performance compared to audio-only or video-only models in video-stroboscopy.

METHODS

Data Collection

This study was approved by the Institutional Review Boards of Weill Cornell Medicine and Columbia University Irving Medical Center (CUIMC). Videolaryngostroboscopy (VLS) recordings were collected from adult patients seen at the Sean Parker Institute for Voice at Weill Cornell Medicine and at the Center for Voice and Swallowing at CUIMC. Patients included those with HVF, UVFP, and VF lesions, and who had audible pre-treatment VLS recordings available. Those with multiple diagnoses (e.g., both UVFP and VF lesion) were excluded. Patients in the VF malignancy group from both institutions were confirmed to have squamous cell carcinoma (SCC) based on histopathology.

To ensure consistency in visualization and reduce variability related to recording modality and image quality differences, only VLS examinations performed using flexible laryngoscopy were included. Recordings obtained via rigid laryngoscopy and those with special filtering systems such as i-scan were excluded to standardize the dataset and minimize differences in acoustic and visual capture that could arise from differences in video-endoscopy technique or patient phonation posture. Only pre-treatment examinations were included, ensuring patients had not received medical, surgical, or behavioral voice therapy prior to their VLS exam. Finally, the senior laryngologist on the team (AR) reviewed the data collection methods throughout the course of the project.

Video and Audio Data Acquisition

Voice samples were recorded during the stroboscopic examination using the Kay-Pentax Stroboscopy System (Laryngeal Strobe: Model 9400; New Jersey, USA), in standard outpatient clinic rooms with sound proofing. At the Sean Parker Institute for the Voice, audio was captured using a lapel microphone positioned on the patient's shirt collar (Model 7175-6000, Pentax Medical; New Jersey, USA). Mouth-to-microphone distance was not standardized. At CUIMC, audio was recorded via a microphone mounted on the camera head (Audio-Technica; Tokyo, Japan), approximately 2 to 3 feet from the patient's mouth. All recordings were captured at a sampling rate of 44 kHz with 16-bit resolution and embedded in MP4 video files at both Sean Parker and CUIMC, which were downloaded from the Pentax EndoPortal at both clinical sites, for downstream data processing.

Video Data Preprocessing

Exams and diagnosis were performed by fellowship trained senior laryngologists on adult patients in clinic setting with a Pentax VNL-1590sTi naso-pharyngo-laryngoscope. All MP4 video files downloaded from the Pentax EndoPortal at both Sean Parker and CUIMC were then trimmed using Apple iMovie (Mac, version 10.4.3). Export settings were standardized across all clips, including MP4 format, 720p resolution, quality: high, and compression speed: faster. Each video was trimmed to approximately five seconds in length and selected to ensure that at least one full cycle of VF abduction and adduction was visible. Additionally, VLS recording MP4 files from CUIMC contained surrounding black frame areas, including metadata overlays such as timestamps and system settings, which extended beyond the region of clinical interest. These non-imaging regions were removed using Apple Photos (Mac, version 10.0), resulting in tightly cropped clips focused solely on the laryngeal anatomy.

Frames at 30 frames per second were extracted from each video and processed using a pre-trained DINOv2 transformer to obtain high-level feature embeddings. DINOv2, a self-supervised model, was chosen for its self-attention mechanism, which captures long-range dependencies across image patches and facilitates the detection of spatially distributed patterns. Recent literature has demonstrated that pre-trained transformer-based models like DINOv2 consistently outperform traditional CNNs in medical image classification tasks.^{12,17} To maximize the model's ability to generalize across varying recording conditions and prevent overfitting, data augmentation techniques, including random cropping, rotation, and horizontal flipping were used.^{15,16} All video processing, including trimming and frame extraction, was performed by authors SAS, RBK, and SS.

Audio Data Extraction

Audio files were extracted from the MP4 video files in WAV format using the Adapter tool (Macroplant; Massachusetts, USA). Recordings from both institutions included low-level background noise, primarily produced from the Pentax Endoscopy System equipment, including noise produced by the flickering strobe light of the laryngoscope. Because these recordings were retrospectively acquired, some clips contained different voice tasks and/or examiner voice overlap; all such non-patient speech segments were therefore identified and excised during our trimming process.

Audio Segment Isolation

Targeted voice segments were isolated from the WAV files produced in the prior extraction step. This task was performed using Audacity (version 3.7.1; Audacity Team, <https://www.audacityteam.org/>), a free, open-source, cross-platform audio editing software. Trained researchers (YK, SAS, RK) utilized Audacity's waveform visualization, real-time playback, and precise selection tools to manually identify and extract specific voice tasks elicited during the VLS exams. The isolated tasks included sustained /i/ phonation, pitch glides, and alternating /i/ with sniff exercises. Only segments that featured the patient's voice were retained for analysis. Audio segments were excluded if they contained non-patient speech or excessive background interference in the form of background noise generated by the

stroboscopy tower. The extracted segments were also saved in WAV format to maintain high temporal resolution and audio quality for downstream acoustic and ML analysis.

Audio Data Preprocessing

Audio signals were resampled to 22,050 Hz. A comprehensive set of acoustic features was extracted from each recording using Librosa, including Mel-frequency cepstral coefficients (MFCCs) with their first and second derivatives, spectral features (centroid, bandwidth, contrast, rolloff), cepstral peak prominence (CPP), acoustic perturbation measures approximating jitter and shimmer, and harmonics-to-noise ratio (HNR) with a total of 51 features.

Principal Component Analysis (PCA) from Python's scikit-learn library was applied to the extracted acoustic features for dimensionality reduction. By evaluating baseline classifier performance, the smallest number of components that retained at least 95% of the variance and maintained stable classification accuracy was selected to optimize the feature set while preserving most of the important information.

Clinicodemographic Data Collection

Patient clinicodemographic variables were retrieved from the electronic health record (EHR). Data collected included age, sex, smoking status, history of neck surgery, and level of voice use according to the classification scheme proposed by Behrman et al.¹⁸ The level of voice use refers to the extent and intensity of vocal activity the patient engages in on a regular basis, categorized into "professional" use (singer, actor, radio personnel), "high" use (including teachers, sales, tech support, mothers of young children, individuals who work with high background noise), and "routine" use.¹⁸ Categorical variables, such as smoking history, were converted into binary format for analysis. Missing values were assigned "unknown."

Multimodal Architecture and Fusion

The multimodal model comprised three parallel pathways. A pre-trained DINOv2 transformer was employed to extract visual features from video frames, with selective fine-tuning via progressive unfreezing of its top layers. The extracted features were then fed into a custom multilayer perceptron (MLP) for classification. In parallel, an MLP was trained on acoustic features extracted from the audio signal using Librosa, and another MLP processed clinicodemographic data. Finally, the features from the video, audio, and metadata branches were integrated using a gated attention mechanism, which employs an uncertainty-based weighting approach to dynamically amplify and reduce the influence of modalities based on certainty, thereby optimizing the fusion of information. (Figure 1).

Training and Evaluation

The video-only, audio-only and multimodal frameworks were trained using stratified patient-level split sampling with 80% of data used for training and 20% held out for testing the final model, with no patient overlap between sets. To mitigate class imbalance, weighted random sampling was employed. All models were trained for up to 20 epochs using a batch size of 12 and an AdamW optimizer (initial learning rate = 1×10^{-4} , weight decay = 5

$\times 10^{-4}$). This means that the learning rate gradually decreased over time, which helped stabilize training and convergence during training. To prevent overfitting, we monitored patient-level validation accuracy and stopped training early if there was no improvement for 3 consecutive epochs. These parameters were chosen a priori based preliminary experiments to reflect commonly effective values and ensure consistency across model comparisons.

The models were then evaluated using accuracies and F1-scores on the same held-out test set. All source code for data processing and model training is available at <https://github.com/Rameau-Lab/Multimodal-.git>

RESULTS

Cohort Characteristics

A total of 136 videos were included in the dataset, with 41 HVF, 54 UVFP, and 42 VF lesions (12 unilateral VF polyps, 16 cysts, 14 SCC) with a mean duration of five seconds. Diagnoses were made by the senior laryngologist in the team with over 15 years of experience. A second retrospective dataset of 40 VLS recordings was collected at CUIMC. This dataset included 15 videos of HVF, 12 with UVFP, and 13 with VF lesions (7 unilateral VF polyps, one cyst, 2 SCC). Full cohort characteristics are presented in Table 1.

Performance of Unimodality and Multimodality Classifiers

The overall accuracy achieved by the multimodal classifier was 76.92% (95% confidence interval [CI]: 57.60% – 88.46%), whereas the video modality classifier achieved an accuracy of 61.54% (95% CI: 38.46% – 76.92%) and the audio modality classifier achieved an accuracy of 65.38% (95% CI: 42.31% – 80.77%). Precision, recall, and F1 score of each model are presented in Table 2.

Confusion matrices were used to visualize classification patterns across modalities (Figure 2). The audio modality correctly identified healthy VF cases in 83% of instances (5 out of 6 cases), but demonstrated considerable misclassification when differentiating between UVFP and VF lesions. In contrast, the video modality correctly classified 100% of UVFP cases (11 out of 11), but performed poorly (50%) when distinguishing healthy VFs from VF lesions. The multimodal approach effectively combined the strengths of the individual modalities, correctly identifying 67% of healthy VF cases (4 out of 6), 100% of UVFP cases, and 70% of VF lesion cases.

When evaluated on the external Columbia dataset, the multimodal classifier achieved an accuracy of 45%, marginally outperforming both the video-only (42%) and audio-only (31%) models.

DISCUSSION

This work demonstrated that integrating stroboscopic video, audio, and clinicodemographic data in a multimodal DL framework improves the diagnosis of VF pathologies compared to relying on unimodal approaches. The multimodal classifier achieved an absolute accuracy gain of 15.38% over the video-only model (76.92% vs. 61.54%) and 11.54% over the

audio-only model (76.92% vs. 65.38%). The observed trend highlights the advantage of combining heterogeneous data sources, as each modality captures different aspects of VF pathology that, when fused, yield more robust diagnostic predictions leveraging cross-modality patterns. This multimodal approach is more representative of the clinical decision-making process by human experts and therefore holds significant promise in providing more reliable results in low validity settings.⁵

Few studies to date have implemented multimodal approaches in otolaryngology, and this is the first study to combine stroboscopic video and audio data with EHR text data for VF pathology classification. Prior work has typically focused on unimodal classifiers, reporting mean accuracies of 89.0% for acoustic studies and 91.3% for imaging studies.¹⁹ Similar to our approach, multimodal application by Kwon et al. and Fang et al. achieved improved classification accuracy with image and voice data, and voice and medical record data, respectively.^{15,16} The comparatively lower accuracy of our unimodal models may be attributed to differences in data preprocessing pipelines, modeling architectures for video and audio inputs, and potential institutional variability in data quality, including differences in recording protocols and patient populations. The model's ability to perform reasonably well despite equipment and mouth-to-microphone distance variability suggests better real-world applicability to diverse clinical settings, as strict standardization, while ideal for model training, can be impractical. Importantly, despite the lower baseline performance of our individual modalities, our multimodal framework achieved a better performance over the best-performing unimodal model than was observed in prior multimodal studies.¹⁵

Several key methodological decisions guided the development of the multimodal model. First, the gated attention mechanism employed an uncertainty-based weighting approach similar to that described by Kendall et. al, allowing the model to dynamically adjust the contribution of each modality based on the estimated uncertainty of its predictions for each sample.²⁰ When analyzing the confusion matrices, this adaptive weighting approach proved particularly valuable for VF lesion cases, where the audio modality often showed higher uncertainty. In these instances, the fusion mechanism appropriately increased the influence of the video modality, which demonstrated stronger discrimination capabilities for lesions. Conversely, when classifying healthy cases, where the video modality showed lower confidence, the model adjusted to leverage the audio features more heavily. This dynamic balancing contributed to the improved performance observed in the multimodal approach. In contrast, clinical decision-making by human experts often relies on fixed or subjective weighting of disparate inputs and may not fully account for complex, non-linear relationships between modalities.⁵ This highlights the potential for multimodal AI to enhance diagnostic accuracy by systematically integrating and weighting diverse sources of information in ways that may be difficult for human cognition to replicate.

Second, raw VLS video recordings, without manual frame selection or extensive preprocessing, were utilized. This approach contrasts with current research that employs careful manual selection of video frames or time-intensive labeling of individual frames to train automated informative frame classifiers.²¹ While our approach yielded lower performance for the video-only classifier relative to other approaches with a mean accuracy of 91%, minimal pre-processing was chosen to enhance clinical applicability

by streamlining the data preparation process and reduce reliance on labor-intensive curation.^{22,23} To address the limitations of minimal video preprocessing, we employed a transformer-based architecture (pretrained DINOv2), which is effective in modeling long-range dependencies and spatiotemporal patterns. DINOv2's self-supervised learning capabilities allowed it to learn rich, high-level feature representations from the video data without the need for extensive manual labeling.²⁴ This capability leverages the inherent structure in the data to improve model performance. As delineated by prior uses of DINOv2, transformers can streamline the process of generating high-quality pseudo-labels, which can then be used to train models more effectively.^{12,25}

Across all classification tasks, the multimodal approach demonstrated the best overall performance. However, in line with existing literature, both the unimodal and multimodal models exhibited reduced performance when evaluated on an external dataset.^{14,26,27} Compared to unimodal classifiers, multimodal models are particularly vulnerable to single-modality variations, as discrepancies in any one input can disrupt the integration of cross-modal information. This sensitivity degrades overall performance and limits generalizability across clinical settings. Nonetheless, through the integration of varying data streams, multimodal models can support more accurate, comprehensive decision-making, particularly for non-expert providers who may lack subspecialty expertise. With ongoing efforts to improve external validation and model robustness, multimodal VLS automation holds significant potential to expand access to high-quality diagnostic support in resource-limited settings, where lack of access to specialty referral often leads to delays in the diagnosis and management of laryngeal disorders.⁹

While this proof-of-concept study provides a novel multimodal approach for the integration of video, audio, and clinicodemographic data, it is not without limitations. First, the small sample size reduces statistical power, contributing to wide confidence intervals and limiting the reliability of performance estimates. In addition, the observed accuracy drop in the external validation set highlights the challenges of generalizability across diverse clinical settings. Expanding the size and diversity of validation cohorts, including through prospective data collection and potentially synthetic data generation, will be critical to improving robustness and external validity.

Further, training a clinicodemographic-only model was avoided because the limited variables would likely cause underfitting, preventing the capture of meaningful patterns and rendering it an uninformative benchmark. This contrasts Fang et al.'s successful analysis of a demographics-only model due to a richer dataset containing over 34 detailed features. However, future work should include sensitivity studies to assess the relative contribution of each modality, including clinico-demographic data, to overall model performance, further clarifying the added value of multimodal integration. Additionally, the audio samples were trimmed; however, truncating voice tasks can misrepresent the phonation time and quality, which are important clinical features in UVFP and VF lesions. While our preprocessing pipeline is less labor-intensive than those used in prior studies, it remains time-consuming and should be further optimized in future work. Lastly, although transformer-based video models provide strong feature representations, they lack inherent interpretability; future

studies applying model explainability techniques and uncertainty quantification can help clinicians understand predictions and gauge confidence in clinical decision support.

CONCLUSION

This study successfully developed a multimodal dataset and classifier for diagnosing VF pathologies, demonstrating the potential benefits of integrating stroboscopic video, audio, and clinicodemographic data for automated laryngeal disease classification. The multimodal classifier achieved higher accuracy than the video-only and audio-only models on both the test and external datasets. Addressing variability in external datasets and increasing sample size remain crucial for advancing the clinical applicability of these models. Nevertheless, our findings demonstrate the potential superiority of multimodal approaches over unimodal AI models for voice pathology classification.

Funding:

Anaïs Rameau was supported by a Paul B. Beeson Emerging Leaders Career Development Award in Aging (K76 AG079040) from the National Institute on Aging and by the Bridge2AI award (OT2 OD032720) from the NIH Common Fund.

REFERENCES

1. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med.* 2022;28(1):31–38. doi:10.1038/s41591-021-01614-0 [PubMed: 35058619]
2. Sisson JC, Schoomaker EB, Ross JC. Clinical decision analysis. The hazard of using additional data. *JAMA.* 1976;236(11):1259–1263. doi:10.1001/jama.236.11.1259 [PubMed: 989070]
3. Alowais SA, Alghamdi SS, Alsuhbany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ.* 2023;23(1):689. doi:10.1186/s12909-023-04698-z [PubMed: 37740191]
4. Chang C, Shi W, Wang Y, Zhang Z, Huang X, Jiao Y. The path from task-specific to general purpose artificial intelligence for medical diagnostics: A bibliometric analysis. *Comput Biol Med.* 2024;172:108258. doi:10.1016/j.combiomed.2024.108258 [PubMed: 38467093]
5. Kline A, Wang H, Li Y, et al. Multimodal machine learning in precision health: A scoping review. *NPJ Digit Med.* 2022;5(1):171. doi:10.1038/s41746-022-00712-8 [PubMed: 36344814]
6. Shao J, Ma J, Yu Y, et al. A multimodal integration pipeline for accurate diagnosis, pathogen identification, and prognosis prediction of pulmonary infections. *The Innovation.* 2024;5(4):100648. doi:10.1016/j.xinn.2024.100648 [PubMed: 39021525]
7. Zhou J, He X, Sun L, et al. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nat Commun.* 2024;15(1):5649. doi:10.1038/s41467-024-50043-3 [PubMed: 38969632]
8. Judge CS, Krewer F, O'Donnell MJ, et al. Multimodal Artificial Intelligence in Medicine. *Kidney360.* 2024;5(11):1771. doi:10.34067/KID.0000000000000556 [PubMed: 39167446]
9. Ren J, Jing X, Wang J, et al. Automatic Recognition of Laryngoscopic Images Using a Deep-Learning Technique. *The Laryngoscope.* 2020;130(11):E686–E693. doi:10.1002/lary.28539 [PubMed: 32068890]
10. Tran BA, Dao TTP, Dung HDQ, et al. Support of deep learning to classify vocal fold images in flexible laryngoscopy. *Am J Otolaryngol.* 2023;44(3):103800. doi:10.1016/j.amjoto.2023.103800 [PubMed: 36905912]
11. Cho WK, Choi SH. Comparison of Convolutional Neural Network Models for Determination of Vocal Fold Normality in Laryngoscopic Images. *J Voice Off J Voice Found.* 2022;36(5):590–598. doi:10.1016/j.jvoice.2020.08.003

12. Paderno A, Rau A, Bedi N, et al. Computer Vision Foundation Models in Endoscopy: Proof of Concept in Oropharyngeal Cancer. *The Laryngoscope*. 2024;134(11):4535–4541. doi:10.1002/lary.31534 [PubMed: 38850247]
13. Bensoussan Y, Elemento O, Rameau A. Voice as an AI Biomarker of Health-Introducing Audiomics. *JAMA Otolaryngol-- Head Neck Surg*. 2024;150(4):283–284. doi:10.1001/jamaoto.2023.4807 [PubMed: 38386315]
14. Hu HC, Chang SY, Wang CH, et al. Deep Learning Application for Vocal Fold Disease Prediction Through Voice Recognition: Preliminary Development Study. *J Med Internet Res*. 2021;23(6):e25247. doi:10.2196/25247 [PubMed: 34100770]
15. Kwon I, Wang SG, Shin SC, et al. Diagnosis of Early Glottic Cancer Using Laryngeal Image and Voice Based on Ensemble Learning of Convolutional Neural Network Classifiers. *J Voice*. 2025;39(1):245–257. doi:10.1016/j.jvoice.2022.07.007 [PubMed: 36075802]
16. Fang SH, Wang CT, Chen JY, Tsao Y, Lin FC. Combining acoustic signals and medical records to improve pathological voice classification. *APSIPA Trans Signal Inf Process*. 2019;8:e14. doi:10.1017/ATSIP.2019.7
17. Takahashi S, Sakaguchi Y, Kouno N, et al. Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review. *J Med Syst*. 2024;48(1):84. doi:10.1007/s10916-024-02105-8 [PubMed: 39264388]
18. Behrman A, Sulica L, He T. Factors predicting patient perception of dysphonia caused by benign vocal fold lesions. *The Laryngoscope*. 2004;114(10):1693–1700. doi:10.1097/00005537-200410000-00004 [PubMed: 15454756]
19. Barlow J, Sragi Z, Rivera-Rivera G, et al. The Use of Deep Learning Software in the Detection of Voice Disorders: A Systematic Review. *Otolaryngol--Head Neck Surg Off J Am Acad Otolaryngol-Head Neck Surg*. 2024;170(6):1531–1543. doi:10.1002/ohn.636
20. Kendall A, Gal Y, Cipolla R. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. Published online April 24, 2018. doi:10.48550/arXiv.1705.07115
21. Baldini C, Azam MA, Sampieri C, et al. An automated approach for real-time informative frames classification in laryngeal endoscopy using deep learning. *Eur Arch Otorhinolaryngol*. 2024;281(8):4255–4264. doi:10.1007/s00405-024-08676-z [PubMed: 38698163]
22. Liu GS, Jovanovic N, Sung CK, Doyle PC. A Scoping Review of Artificial Intelligence Detection of Voice Pathology: Challenges and Opportunities. *Otolaryngol--Head Neck Surg Off J Am Acad Otolaryngol-Head Neck Surg*. 2024;171(3):658–666. doi:10.1002/ohn.809
23. Pelletier ED, Jeffries SD, Song K, Hemmerling TM. Comparative Analysis of Machine-Learning Model Performance in Image Analysis: The Impact of Dataset Diversity and Size. *Anesth Analg*. 2024;139(6):1332–1339. doi:10.1213/ANE.0000000000007088 [PubMed: 39116018]
24. DINOv2: Learning Robust Visual Features without Supervision. Accessed April 16, 2025. <https://arxiv.org/html/2304.07193v2>
25. Zhang X, Zhao J, Zong D, Ren H, Gao C. Taming vision transformers for clinical laryngoscopy assessment. *J Biomed Inform*. 2025;162:104766. doi:10.1016/j.jbi.2024.104766 [PubMed: 39827999]
26. Zhao Q, He Y, Wu Y, et al. Vocal cord lesions classification based on deep convolutional neural network and transfer learning. *Med Phys*. 2022;49(1):432–442. doi:10.1002/mp.15371 [PubMed: 34813114]
27. Kim YE, Dobko M, Li H, et al. A Deep-Learning Model for Multi-class Audio Classification of Vocal Fold Pathologies in Office Stroboscopy. *The Laryngoscope*. Published online February 5, 2025. doi:10.1002/lary.32036

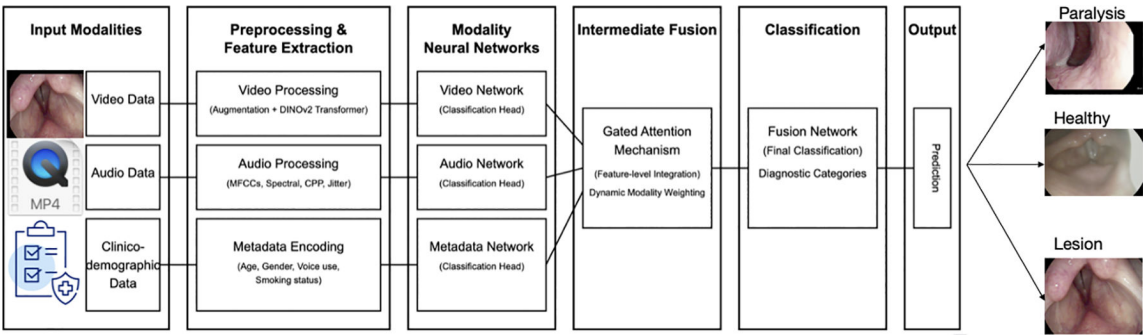


Figure 1:
Multimodal Approach Methodology

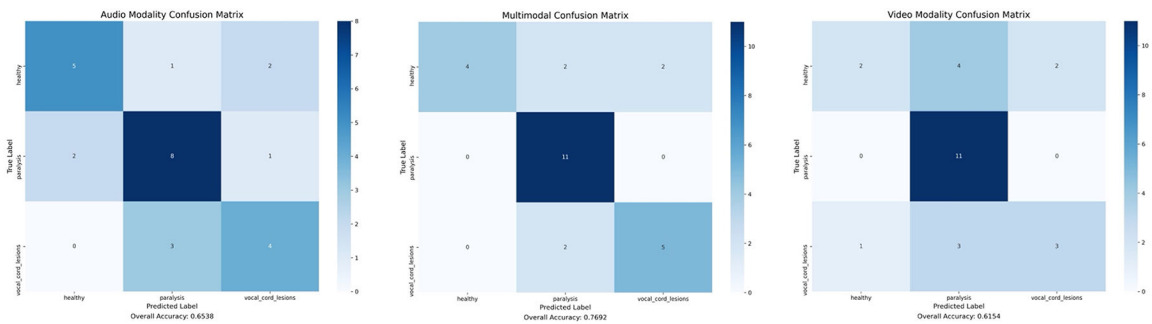


Figure 2:
Audio-Only, Video-Only, and Multimodal Confusion Matrices

Table 1:**Demographics of Patients in Training Data Set**

		Vocal Fold Patient Cohort				
		All Patients (N=137)	Healthy (N=41)	Polyp(s) (N=12)	Cyst(s) (N=16)	Paralysis (N=54)
Age ± Standard Deviation (Yrs)		56.18 ± 19.41	45.56 ± 20.48	44.92 ± 19.44	56.76 ± 18.22	62.61 ± 16.89
Sex	Male	55 (40.15%)	8 (19.51%)	6 (50.00%)	6 (37.5%)	21 (38.89%)
	Female	82 (59.85%)	33 (80.49%)	6 (50.00%)	10 (62.50%)	33 (61.11%)
Level of Voice Use	Routine	92 (67.15%)	19 (46.34%)	6 (50.00%)	10 (62.50%)	52 (96.30%)
	High Intensity	31 (22.65%)	22 (53.66%)	2 (16.67%)	4 (25.00%)	1 (1.85%)
Vocal Performer	Vocal Performer	7 (5.10%)	0 (0.00%)	4 (33.33%)	2 (12.50%)	0 (0.00%)
	Unknown	7 (5.10%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (1.85%)
Smoking History?	Yes	37 (27.01%)	6 (14.63%)	2 (16.67%)	0 (0.00%)	21 (38.89%)
	No	98 (71.53%)	33 (80.49%)	10 (83.33%)	16 (100.00%)	33 (61.11%)
H&N Surgery History?	Unknown	2 (1.46%)	2 (4.88%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
	Yes	18 (13.14%)	3 (7.32%)	0 (0.00%)	1 (6.25%)	14 (25.93%)
	No	119 (86.86%)	38 (92.68%)	12 (100.00%)	15 (93.75%)	50 (74.07%)
						14 (100.00%)

Table 2:
Precision, Recall, and F1 Score of Audio-Only, Video-Only, and Multimodal Models.

Classification	Model	Precision	Recall	F1-score	Test Set (N)
Healthy Vocal Folds	Audio	0.62	0.62	0.62	8
	Video	0.29	0.25	0.27	8
	Multimodal	1.00	0.5	0.67	8
Unilateral Vocal Fold Paralysis	Audio	0.62	0.73	0.67	11
	Video	0.55	1.00	0.71	11
	Multimodal	0.73	1.00	0.85	11
Vocal Fold Lesions	Audio	0.56	0.57	0.57	7
	Video	0.5	0.43	0.46	7
	Multimodal	0.71	0.71	0.71	7
Overall Accuracy	Audio	0.65	N/A	N/A	26
	Video	0.62	N/A	N/A	26
	Multimodal	0.77	N/A	N/A	26