ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11463621

# AUTOMATED DYSPHAGIA SCREENING USING NONINVASIVE NECK ACOUSTIC SENSING

*Jade Chng*[1,2,*], *Rong Xing*[1,*], *Yunfei Luo*[3,*],
*Kristen Linnemeyer-Risser*[4], *Tauhidur Rahman*[1,3], *Andrew Yousef*[4,†], *Philip A Weissbrod*[4,†]

[1] Jacobs School of Engineering, University of California San Diego
[2] Department of Biomedical Engineering, Duke University
[3] Halıcıoğlu Data Science Institute, University of California San Diego
[4] Department of Otolaryngology-Head and Neck Surgery, University of California San Diego

## ABSTRACT

Pharyngeal health plays a vital role in essential human functions such as breathing, swallowing, and vocalization. Early detection of swallowing abnormalities, also known as dysphagia, is crucial for timely intervention. However, current diagnostic methods often rely on radiographic imaging or invasive procedures. In this study, we propose an automated framework for detecting dysphagia using portable and noninvasive acoustic sensing coupled with applied machine learning. By capturing subtle acoustic signals from the neck during swallowing tasks, we aim to identify patterns associated with abnormal physiological conditions. Our approach achieves promising test-time abnormality detection performance, with an AUC-ROC of 0.904 under 5 independent train–test splits. This work demonstrates the feasibility of using noninvasive acoustic sensing as a practical and scalable tool for pharyngeal health monitoring.

***Index Terms*—** Applied Machine Learning, Signal Processing, Pharyngeal Health, Digital Health.

## 1. INTRODUCTION

Dysphagia, or difficulty swallowing, continues to pose a significant public health concern, affecting 10-20% of adults over the age of 50 and up to two-thirds of patients with conditions such as Parkinson's disease, stroke, or head and neck cancer [1, 2]. It is estimated that around $4.3 to $7.1 billion go towards hospitalization costs associated with dysphagia annually [3]. Currently, swallow evaluations are performed through video-based modalities such as flexible endoscopic evaluation of swallowing (FEES), videofluoroscopic swallowing studies (VFSS) or high-resolution manometry, which, while informative, are either invasive, require radiation exposure, necessitate a trained practitioner, or are limited in cost efficiency [4, 5]. Given the limitations of current gold-standard swallow evaluations, hospitalized patients at risk for aspiration are typically screened using a clinical swallow evaluation. However, these assessments are performed without instrumentation, demonstrate limited diagnostic sensitivity and specificity, and often leave providers with uncertainty in clinical decision-making. This highlights the pressing need for an objective swallow assessment tool capable of efficiently and accurately identifying dysphagia risk.

To address this gap, we collected audiometric data during fiberoptic endoscopic evaluation of swallowing (FEES) to investigate associations between swallow dysfunction and acoustic features. We then used these data to develop a machine learning model for predicting swallow dysfunction. We present a system that achieved promising classification performance, suggesting that surface digital auscultation captures acoustic signatures with diagnostic value reflective of underlying oropharyngeal dysfunction.

## 2. RELATED WORK

To address these challenges imposed by dysphagia assessments, a growing body of research has also explored the use of machine learning techniques to analyze swallow mechanisms and detect related abnormalities using a mix of accelometry, acoustic signals, and electromyography (EMG) [6, 7, 8, 9]. While preliminary studies suggest the potential of machine learning for detecting swallow dysfunction, they remain significantly limited. First, and most notably none have conducted real-time assessments of swallow sounds during a gold-standard evaluation such as FEES, restricting their ability to characterize the severity and nature of dysfunction. Second, many studies validated their models using only a single bolus consistency (e.g., a sip of water) [6, 7, 8, 9], despite the fact that dysphagia and aspiration may manifest differently across consistencies- thereby limiting external validity. Finally, most prior work has been constrained to proof-of-concept designs with small sample sizes [10]. To overcome these limitations, we developed a machine learning model trained on more than 600 swallow events, using acoustic data collected during comprehensive swallow evaluations.

[*] These authors contributed equally to this work.
[†] Equal Senior Authors

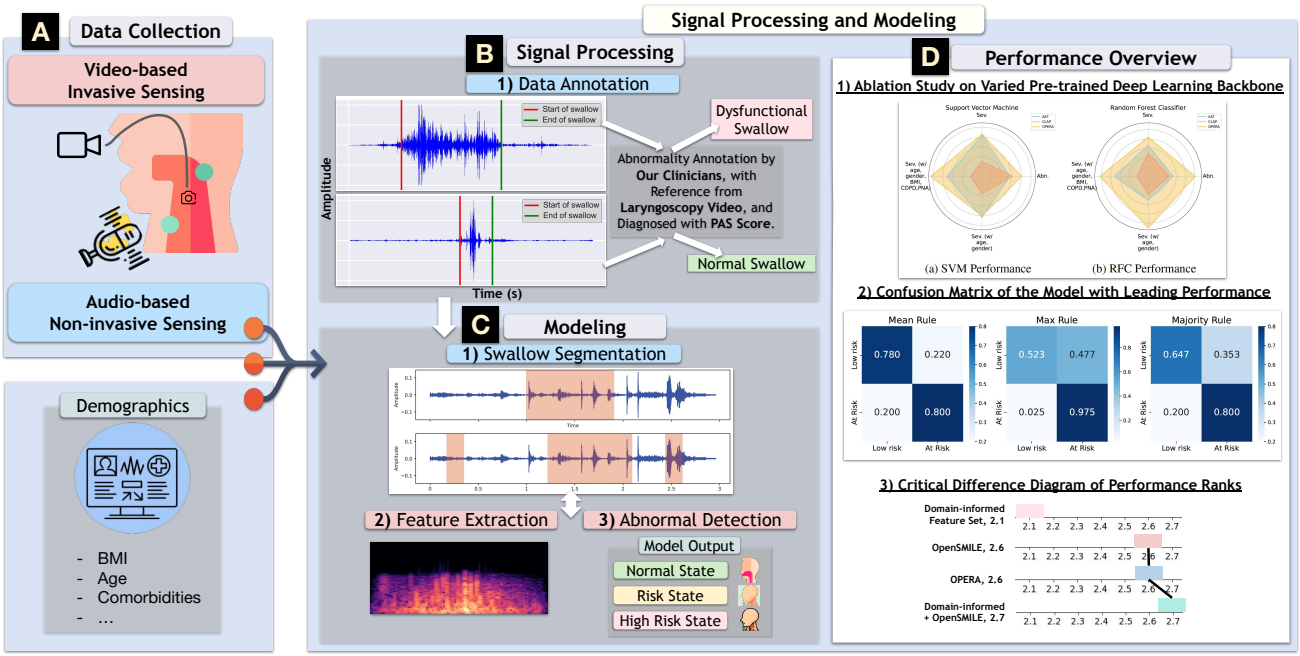


**Fig. 1**: Overview of our proposed automated system. (A) Data collected from participants during this study. (B) Demonstration of our data annotation process. (C) Modeling procedure to explore the relationship between acoustic signal and the swallow abnormalities. (D) Presentation of the empirical results.

## 3. METHOD

### 3.1. Preliminary Experiments

Initial preliminary experiments were conducted to explore the performance of established baseline audio embedding models: AST, CLAP and OPERA [11, 12, 13] (Figure 1 section D.1.). This was done with our existing dataset, both with and without demographic features across 5 randomized train–test splits. "Sev." and "Abn." standing for severity (3-class) and abnormality (2-class) respectively. It is observed that the OPERA model achieve the best performance, hence it was included in our main experiments as a baseline comparison.

Although age and gender did not have substantial improvement to the model's performance, we decided to include it due to prior research indicating that age and gender may change the acoustic characteristics of the swallow[14]. Furthermore, the performance between RFC and SVM was comparable, but for consistency we chose to use RFC for our main evaluation.

### 3.2. Data Annotation

This study was approved by University of California San Diego, where we recruited a total of 49 participants from Center for Airway, Voice and Swallowing, who self-reported symptoms of dysphagia. These participants underwent swallow evaluations using FEES. A standard FEES includes 8-10 trials of oral intake with different consistencies and bolus sizes. During this, a flexible laryngoscope is placed through the nose and each swallow trial is observed. Speech and language pathologists and fellowship-trained laryngologists rated each FEES using the penetration-aspiration scale (PAS) to assess the severity of swallow dysfunction. The PAS is a categorical assessment rating, ranging from 1–8, assessing the degree of dysfunction and evidence of oral intake invading the airway with higher scores indicating higher levels of swallow dysfunction. Based on this scale, PAS of 1-2 are relatively normal swallows, PAS of 3-5 indicate evidence of penetration, and PAS of 6 and above refers to evidence of aspiration. During the FEES, each participant also had a digital stethoscope (3M TM Littmann Core Digital Stethoscope) placed lateral to the thyroid cartilage to collect audiometric data in real time during the FEES.

In total, 392 audio recordings were collected for this study. Corrupt audio files that contained talking or non-swallow sounds that our parameters could not filter out were removed. The swallows in each recording were then cleaned in preparation for feature extraction with an average duration of 0.64 seconds. After processing, there were 617 individual

**Table 1**: Background Distribution of Participants

| Factors | Sample Size (n) | Percentage (%) |
|---|---|---|
| **Gender** | | |
| Female | 25 | 51.02 |
| Male | 24 | 48.98 |
| **PAS** | | |
| Normal | | |
| PAS[1-2] | 22 | 44.90 |
| Abnormal | | |
| PAS[3-5] | 20 | 40.82 |
| PAS[6-8] | 7 | 14.29 |
| Total | 27 | 55.10 |
| **BMI** | | |
| $<$ 18.5 (underweight) | 2 | 4.08 |
| $\leq$ 24.9 (normal) | 17 | 34.69 |
| $>$ 24.9 (overweight) | 30 | 61.22 |
| **Age** | | |
| 30–60 | 13 | 26.53 |
| 62–69 | 16 | 32.65 |
| 70–96 | 20 | 0.41 |
| **Chronic Obstructive Pulmonary Disease (COPD)** | | |
| No | 41 | 83.67 |
| Yes | 8 | 16.33 |
| **History of Pneumonia caused by Aspiration (PNA)** | | |
| No | 38 | 77.55 |
| Yes | 11 | 22.45 |

swallow events (24 participants contributed to 10-15 swallow events each, 10 participants contributed to 15-20 events each, 8 participants contributed $\leq$10 events each and 3 participants contributed $\geq$20 events each). We aimed to use our domain-informed features compared with features extracted from OpenSMILE and embeddings from OPERA to differentiate patients with swallow dysfunction from those without and possibly reflect the severity of the patients' condition. The distribution of patients' background data is shown in Table 1.

We used the Python Librosa [15] to process and "clean" the audio recordings collected for this study and used several parameters to properly identify the start and end (a single segment) of all swallows in each audio file. The parameters consist of 4 elements: (i) A threshold of amplitudes with unit in dB to determine the silence period; (ii) The gap time in seconds for which segments that are within this range are merged into a single segment; (iii) The minimum amplitude for which segments with amplitudes less than this value are discarded; and (iv) A maximum amplitude for which segments with amplitudes greater than this value are discarded. By cross referencing with the visual data of the swallows, we tuned the parameters for each audio file individually such that all swallows were segmented correctly (Figure 1 section B.1.).

### 3.3. Feature Extraction and Modeling

In addition to extracting OPERA embeddings, we also analyzed a series of domain-informed features that are heuristically associated with swallow dysfunction [16].

#### *3.3.1. Frequency*

To calculate the maximum five frequencies of each audio file, we used Fast Fourier Transform. This algorithm efficiently implements the discrete Fourier Transform (DFT) equation, which transforms signals from a time domain $x[n]$ into its frequency domain $X[k]$. DFT Equation:

$$X[k] = \sum_{n=0}^{N-1} x[n]\, e^{-j2\pi kn/N} \tag{1}$$

Here, $x[n]$ is the input signal at time index $n$, $j$ is the imaginary unit, $k$ is the index of the frequency bin, $N$ is the total number of samples and $X[k]$ is the transformed signal in the frequency domain.

Then we used Librosa's Short-Time Fourier Transform (STFT) function, which repeatedly applies this formula to overlapping windows, to get the entire frequency content of the audio file. Given this frequency content over time, we calculated the average and median frequency over time for each audio file. Finally, we extracted the maximum five frequency values: $freq_i$ where $i = 1, ..., 5$.

#### *3.3.2. Amplitude*

To get the amplitude- related values, we used Librosa's load function which converts the .WAV audio files into a time series represented as a 1-dimensional array of floating point values. These values are the respective amplitudes of all samples of the file. Hence, using this array, we calculated the peak (maximum) and average amplitude over time.

#### *3.3.3. Area Under the Curve*

To calculate the area under the curve, we integrated the absolute waveform using the composite trapezoidal rule provided by the Numpy library. The explicit formula of the composite trapezoidal rule is:

$$\int_a^b f(x)\, dx \approx \frac{1}{2} \sum_{j=1}^{n} (x_j - x_{j-1}) \left[ f(x_{j-1}) + f(x_j) \right] \tag{2}$$

where $a, b$ is the interval of integration and $n$ is the number of partitions.

#### *3.3.4. Representations from Pretrained Audio Models.*

Finally, to stay consistent with literature, we also collected all features available from OpenSMILE's [17] toolkit. (For all three methods of extraction, participants' age and gender was included.)

### 3.4. Segmentation and Aggregation Analysis

In clinical settings, audio recordings often contain multiple swallows within a single trial. Hence, the model is required to detect and classify individual swallows from continuous input. To simulate this, we further evaluate model performance by using both *fixed-parameter* and *sliding window automatic segmentation.* These methods simulate automated preprocessing in real-world deployment. After generating per-swallow predictions, we computed patient-level classifications by aggregating results across all swallows for that patient using three strategies: Mean-risk (average prediction), Max-risk (highest risk across swallows), and Mode-risk (most frequent prediction).

#### *3.4.1. Fixed Parameter Segmentation*

Fixed parameters were selected by testing threshold combinations and maximising intersection over union (IoU) with ground-truth masks. The best overlap (IoU = 0.4775) was obtained with top_db = 20, gap_time = 0.6, allowed_max_amplitude = 2, and allowed_min_amplitude = 0, yielding 65.8% sensitivity and 87.6% specificity. An example segmentation is shown in Figure 1 section C.1. with ground truth on top of the predicted segmentation.

#### *3.4.2. Sliding Window Segmentation*

The sliding window Segments the audio into 1-second windows with 50% overlap, ensuring full coverage of the audio timeline. This strategy minimizes the risk of missing swallows and is also algorithmically simple, making it suitable for real-time use.

### 3.5. Evaluation Protocol

All experiments were carried out using train-test splits at the patient level to better reflect clinical use scenarios. This ensures that the model does not overfit to patients already seen

**Table 2**: Main Results (Patient-Level Splits). "Sev." and "Abn." stand for severity and abnormality respectively.

| Task | Method | AUC ROC | AUC PRC | Balanced Accuracy |
|---|---|---|---|---|
| Sev. | OPERA | $0.557 \pm 0.159$ | $0.434 \pm 0.130$ | $0.542 \pm 0.047$ |
| Sev. | OpenSMILE (OpSL) | $0.583 \pm 0.120$ | $0.503 \pm 0.145$ | $0.606 \pm 0.079$ |
| Sev. | Domain-Informed | $\mathbf{0.611 \pm 0.055}$ | $\mathbf{0.519 \pm 0.061}$ | $\mathbf{0.659 \pm 0.028}$ |
| Sev. | Domain-Informed w/ OpSL | $0.561 \pm 0.135$ | $0.493 \pm 0.120$ | $0.610 \pm 0.080$ |
| Abn. | OPERA | $0.651 \pm 0.176$ | $0.718 \pm 0.140$ | $0.579 \pm 0.080$ |
| Abn. | OpenSMILE | $0.778 \pm 0.144$ | $0.850 \pm 0.094$ | $0.665 \pm 0.152$ |
| Abn. | Domain-Informed | $\mathbf{0.904 \pm 0.015}$ | $\mathbf{0.913 \pm 0.075}$ | $\mathbf{0.755 \pm 0.061}$ |
| Abn. | Domain-Informed w/ OpSL | $0.804 \pm 0.183$ | $0.862 \pm 0.081$ | $0.710 \pm 0.159$ |

**Table 3**: Evaluation with Audio Segmentation, Aggregated by Patient (AUC-ROC Scores)

| Method | Mean | Max | Mode |
|---|---|---|---|
| Sliding Window | $0.893 \pm 0.103$ | $0.856 \pm 0.106$ | $0.884 \pm 0.104$ |
| Fixed-Parameters | $0.868 \pm 0.142$ | $\mathbf{0.942} \pm 0.051$ | $0.842 \pm 0.141$ |
| Human Segmented Swallows | $\mathbf{0.967} \pm 0.054$ | $0.918 \pm 0.079$ | $\mathbf{0.971} \pm 0.041$ |

during training. This emulates our target clinical use case, such that the model is expected to validate and make inferences on entirely unseen patients. We performed five independent patient-level splits, with patient ID used as the primary unit of data partitioning. Each split was stratified to maintain class and swallow distribution. We report AUC-ROC scores as the primary metric for comparing model performance.

## 4. RESULTS

### 4.1. Feature Performance Comparison

The 2-class (abnormality detection) model trained with our domain-informed features performed the best with an **AUC-ROC of 0.904** compared to the pretrained audio embeddings (Table 2). Combining OpenSMILE with our domain-informed features decreases the model performance to a AUC-ROC score of 0.804 in a binary-class setting. This makes sense as the acoustic features of OpenSMILE may capture irrelevant or noisy characteristics. Overall performance in the 3-class (severity classification) setting is notably lower than in the binary-classification setting. This is largely due to the smaller training samples per-class which hinders the model's ability to learn discriminative patterns for each category. Further exploration with larger and more balanced datasets may help unlock the potential of multiclass swallowing classification.

### 4.2. Segmentation and Patient Aggregation Results

The Human Segmented Swallows condition serves as a benchmark, representing the scenario where all swallows are accurately segmented (Table 3). The Fixed parameter segmentation method produces promising results, with the highest AUC-ROC score of 0.942 observed under the Max-risk aggregation strategy. However, we note that under Mean-risk and Mode-risk aggregation, the Sliding-window segmentation outperforms the fixed-parameter approach. This may suggest that while the fixed-parameter & max-risk result is strong, it may be a dataset-specific artifact rather than a consistently superior configuration.

The confusion matrices uses Human Segmented Swallows, illustrating that the model is effective in identifying at-risk patients with minimal false negatives (Figure 1 section D.2). This is clinically significant, as misclassifying high-risk patients could result in missed patients with unsafe swallows being cleared for an oral diet putting them at risk for aspiration pneumonia.

### 4.3. Feature importance and statistical analysis

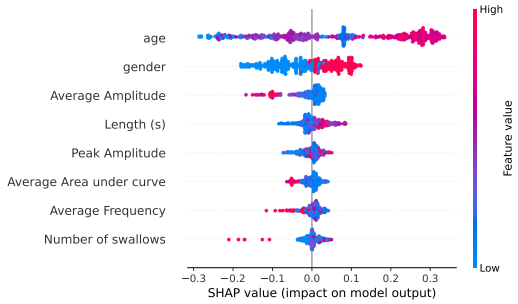


**Fig. 2**: SHAP Summary Plot of Top 8 Features from Performance on Human Segmented Swallows

SHAP analysis (Fig.2) shows that age and gender are influencial predictors of dysphagia, with older age and male sex associated with higher risk, consistent with prior studies.[18, 19] Signal-based features, average amplitude, frequency, area under the curve, and swallow count also strongly influenced classification, with weaker swallows linked to dysphagia. These findings support integrating both demographics and acoustic characteristics maximizes predictive accuracy.

Lastly, we compared the accuracy of using our domain-informed features to previously created acoustic datasets (OPERA & OpenSMILE) and found that our domain-informed method outperformed all other approaches, as shown by the critical difference test (Figure 1 section D.3).

## 5. CONCLUSION

This study presents a machine learning framework for detecting dysphagia using noninvasive acoustic sensing. By capturing neck-region acoustic signals and leveraging time-frequency analysis, our system demonstrates the potential for identifying abnormal physiological patterns of swallowing with high accuracy. The proposed approach offers a low-cost, scalable alternative to traditional diagnostic methods and represents a step toward accessible, real-time pharyngeal health monitoring.

**Limitation and Future Work.** While promising, the dataset size remains moderate, limiting generalization across diverse populations and conditions. Expanding the dataset, improving the automatic swallow segmentation algorithm and validating in diverse scenario, such as at-home settings, will be essential next steps.

## 6. ACKNOWLEDGMENTS

We thank medical students Sacha Moufarrej, Bao Luu, and David Zeng from the University of California San Diego School of Medicine, for their assistance with splicing the swallow audio files. No external funding was received for this study. The authors declare no relevant financial or nonfinancial conflicts of interest. All data were collected from patients as part of an approved study and were fully anonymized prior to analysis.

## 7. REFERENCES

[1] Z. W. Seo, J. H. Min, S. Huh, Y.-I. Shin, H.-Y. Ko, and S.-H. Ko, "Prevalence and severity of dysphagia using videofluoroscopic swallowing study in patients with aspiration pneumonia," *Lung*, vol. 199, pp. 55–61, 2021.

[2] M. C. Rivelsrud, L. Hartelius, L. Bergström, M. Løvstad, and R. Speyer, "Prevalence of oropharyngeal dysphagia in adults in different healthcare settings: a systematic review and meta-analyses," *Dysphagia*, vol. 38, no. 1, pp. 76–121, 2023.

[3] D.A. Patel, S. Krishnaswami, E. Steger, E. Conover, M. F Vaezi, M. R. Ciucci, and D. O. Francis, "Economic and survival burden of dysphagia among inpatients in the united states," *Diseases of the Esophagus*, vol. 31, no. 1, pp. dox131, 2018.

[4] G. Sanjeevi, U. Gopalakrishnan, R. Pathinarupothi, and K. Iyer, "Artificial intelligence in videofluoroscopy swallow study analysis: A comprehensive review," *Dysphagia*, pp. 1–14, 2025.

[5] D. Wong, J. Wang, S. Cheung, D. Lai, A. Chiu, D. Pu, J. Cheung, and T. Kwok, "Current technological advances in dysphagia screening: Systematic scoping review," *Journal of Medical Internet Research*, vol. 27, pp. e65551, 2025.

[6] M. Nikjoo, C. Steele, E. Sejdić, and T. Chau, "Automatic discrimination between safe and unsafe swallowing using a reputation-based classifier," *Biomedical engineering online*, vol. 10, pp. 1–18, 2011.

[7] C. Steele, R. Mukherjee, J. Kortelainen, H. Pölönen, M. Jedwab, S. Brady, K. Theimer, S. Langmore, L. Riquelme, N. Swigert, et al., "Development of a noninvasive device for swallow screening in patients at risk of oropharyngeal dysphagia: results from a prospective exploratory study," *Dysphagia*, vol. 34, pp. 698–707, 2019.

[8] C. Steele, E. Sejdić, and T. Chau, "Noninvasive detection of thin-liquid aspiration using dual-axis swallowing accelerometry," *Dysphagia*, vol. 28, pp. 105–112, 2013.

[9] H. Mohammadi, A. Samadani, C. Steele, and T. Chau, "Automatic discrimination between cough and non-cough accelerometry signal artefacts," *Biomedical Signal Processing and Control*, vol. 52, pp. 394–402, 2019.

[10] D. Jayatilake, T. Ueno, Y. Teramoto, K. Nakai, S. Hidaka, K.and Ayuzawa, K. Eguchi, and A. Matsumura, "Smartphone-based real-time assessment of swallowing ability from the swallowing sound," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 3, 11 2015.

[11] Y. Gong, Y. Chung, and J. Glass, "Ast: Audio spectrogram transformer," 2021.

[12] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap: Learning audio concepts from natural language supervision," 2022.

[13] Y. Zhang, T. Xia, J. Han, Y. Wu, G. Rizos, Y. Liu, M. Mosuily, J. Chauhan, and C. Mascolo, "Towards open respiratory acoustic foundation models: Pretraining and benchmarking," 2024.

[14] S. Youmans and J. Stierwalt, "Normal swallowing acoustics across age, gender, bolus viscosity, and bolus volume," *Dysphagia*, vol. 26, no. 4, pp. 374–384, Jan 2011.

[15] B. McFee, M. McVicar, D. Faronbi, I. Roman, M. Gover, S. Balke, S. Seyfarth, A. Malek, C. Raffel, V. Lostanlen, B. van Niekerk, D. Lee, F. Cwitkowitz, F. Zalkow, O. Nieto, D. Ellis, J. Mason, K. Lee, B. Steers, and D. Südholt, "librosa/librosa: v0.11.0," 2025, Audio and music signal processing in Python.

[16] A. Yousef, S. Moufarrej, D. Zeng, B. Luu, R. Xing, J. Crook, Y. Luo, K. Linnemeyer-Risser, T. Rahman, and P. Weissbrod, "Sounding out dysphagia: Using a digital stethoscope to assess audiometric differences in swallowing," *The Laryngoscope*, 2025.

[17] Oana Pascu, Diana Oneaţă, Horia Cucu, and Nikolaus Müller, "Easy, interpretable, effective: opensmile for voice deepfake detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[18] C. Bolinger, "Pneumonia: Does age or gender relate to the presence of dysphagia in geriatric patients?," *Pneumonia*, vol. 12, no. 1, pp. 14, 2020.

[19] I. Tomonaga, H. Koseki, C. Imai, T. Shida, Y. Nishiyama, D. Yoshida, S. Yokoo, and M. Osaki, "Incidence and characteristics of aspiration pneumonia in the nagasaki prefecture from 2005 to 2019," *BMC Pulmonary Medicine*, vol. 24, no. 191, 2024.