Artificial intelligence in otolaryngology: redefining automatic laryngeal paralysis assessment for optimal healthcare

Artificial Intelligence in Otolaryngology: Redefining Automatic Laryngeal Paralysis Assessment for Optimal Care

Emanuele Agrimi[a,b], Francesco Pietrogiacomi[†a], Linda Fiorini[†a], Francesca Mularoni[c], Isabel Vilaseca[d], Giorgio Peretti[f], Stefano Taboni[c], Marco Ferrari[c], Andrea Luigi Camillo Carobbio[c], Piero Nicolai[c], Claudio Sampieri[d,e] and Giorgio Gnecco[a,*]

[a]*IMT, School for Advanced Studies, Lucca, Italy*
[b]*Korteweg de Vries Institute for Mathematics, University of Amsterdam, The Netherlands*
[c]*Department of Neurosciences, Otorhinolaryngology, University of Padua, Padua, Italy*
[d]*Department of Otolaryngology, Hospital Clínic de Barcelona, Barcelona, Spain*
[e]*Department of Experimental Medicine (DIMES), University of Genoa, Genoa, Italy*
[f]*Department of Surgical Sciences and Integrated Diagnostics (DISC), University of Genoa, Genoa, Italy*

ARTICLE INFO

Keywords:
Artificial Intelligence Applied to Medicine
Computer Vision
Otolaryngology
Laryngeal Paralysis Assessment

ABSTRACT

Background and Objective: Laryngeal motility assessment is essential for diagnosing and managing laryngeal disorders. However, paralysis evaluations suffer from high inter-rater variability, necessitating a more objective and quantitative approach. This study introduces a novel AI-driven pipeline that leverages computer vision techniques to classify 155 video-laryngoscopies into unilateral paralysis (n=68), bilateral paralysis (n=50), and healthy laryngeal function (n=37).
Methods: Our approach includes several advancements over existing literature. We extract the vocal fold positions from each video and automatically identify the most informative, noise-cleaned video segments for classification. We define novel movement-based features that quantitatively capture the restricted mobility characteristics of laryngeal paralysis. These features are used to train two classification models using a 5-fold cross-validation strategy: one model for binary classification (healthy vs. paralyzed) and the other for multi-class classification (healthy vs. unilateral paralysis vs. bilateral paralysis). To assess the importance of these features, we conduct an ablation study using Shapley values.
Results: Our method achieved a precision of 0.83, sensitivity (recall) of 0.85, F1-score of 0.84, and balanced accuracy of 0.85 for distinguishing between healthy and paralyzed individuals. For multi-class classification (healthy vs unilateral paralysis vs bilateral paralysis), our model achieves a precision of 0.80, sensitivity of 0.83, F1-score of 0.81, and a balanced accuracy of 0.83. These results highlight the effectiveness of our method and underscore the relevance of our features, further validated by the ablation study.
Conclusions: Our AI-grounded approach enhances the accuracy and reliability of automatic laryngeal motility assessment. By introducing novel metrics to quantify paralysis severity, we provide a more objective, reproducible, and clinically valuable evaluation tool.

1. Introduction

Human larynx plays a pivotal role in our ability to communicate, swallow, and breathe [1]. Its function, however, can be severely compromised by various disorders, including Parkinson's disease, laryngeal nerves injury, stroke, and Laryngeal Cancer (LC) [2]. One of the key indicators of laryngeal health is laryngeal motility, which allows to close and open the airway while swallowing, breathing, and speaking [3]. A correct motility assessment is crucial for the diagnosis, management, and treatment of laryngeal disorders. For instance, in LC patients, reduced motility significantly impacts prognosis and treatment decisions [4]. Considering that laryngeal motility can also be impaired by tumors' vocal muscle or vocal cord-arytenoid unit invasion or compression, not even laryngeal electromyography can be considered a gold standard for motility assessment and therefore, most of the time, laryngeal paralysis is diagnosed based on endoscopic findings. However, the assessment of laryngeal motility requires time and expertise

[†]These authors contributed equally to this work
*Corresponding author
giorgio.gnecco@imtlucca.it (G. Gnecco)
ORCID(s): 0000-0002-5427-4328 (G. Gnecco)

and is characterized by considerable variability both between and within raters [5, 6]. This often leads to subjective assessments that rely heavily on the clinician's expertise and experience. Naturally, this approach has limitations, including potential variability and errors. To address these challenges, researchers have begun to explore the potential of Artificial Intelligence (AI) techniques in providing a more objective and comprehensive assessment of laryngeal motility. Some of these applications, for instance, have focused on developing classification models to diagnose laryngeal paralysis based on the patient's voice [7]. However, voice can be impaired by many factors, thereby limiting the utility of voice analysis for screening purposes and pushing researchers to focus on endoscopic findings. In this setting, to extract more tangible information, the majority of the literature in this field has focused on computer vision techniques to automatically extract quantifiable features that describe the reduced patterns of movements shown by impaired vocal folds in video-laryngoscopies [8, 9, 10, 11]. Among these, the Automated Glottic Action Tracking by Artificial Intelligence (AGATI) toolbox, developed in [8], enables the tracking of the vocal folds' free edge position in each frame of the video-laryngoscopy. This position is used to measure the value of the Anterior Glottic Angle (AGA) and the angle between one of the two vocal folds and a reference vertical half-line. Recent works [9, 12] leverage similar mechanisms, enabling the tracking of the AGA value in real-time. However, these methods can fail when vocal folds are not aligned as expected or are not well-framed. Another work [10] instead allows the segmentation of the area comprised between the two vocal folds through the application of a Gabor filter [13] and a Chan-Vese segmentation [14] to a single frame. Similarly, deep learning was recently applied to vocal-fold segmentation from laryngoscopy images [15]. These methods, allowing for feature extraction, led to the development of novel classification models. Kuo and colleagues [16], for instance, used a decision tree trained on the output of another work [10] to categorize subjects into healthy, paralyzed, and patients with nodules. These works helped identify the most straightforward and crucial features for assessing laryngeal paralysis. As observed by clinicians in daily practice, previous works [17, 18] identified the minimum and maximum AGA values observed in each video-laryngoscopy as a robust, generalizable, and reliable feature for paralysis assessment. However, the AGA value is not always easily measurable, as in the case of LC patients or when dealing with noisy video recordings. This issue becomes even more critical when comparing patients with unilateral paralysis to healthy subjects, as inaccurate estimates of AGA values can result in misclassification. This question is addressed in recent works, which apply deep learning to automatically estimate vocal fold pose from video-laryngoscopy frames [19]. Lately, there was a growing interest in AI-assisted endoscopy, where automated pipelines can reduce subjectivity and increase throughput [20, 21, 22]. Notable examples of recent and relevant studies that also address vocal fold paralysis classification using video-laryngoscopy are the study by Villani et al. [23] and by Zhang et al. [24]. Both focus on the recognition of landmarks from frames of video-laryngoscopies, but while the first uses machine learning models to classify vocal fold fixation based on manual landmark annotations and image features, the second leverages a multimodal approach that combines video and audio data to classify vocal fold paralysis and distinguish between unilateral and bilateral paralysis. While Villani et al. [23] require manual intervention for feature extraction and landmark annotation, Zhang et al. [24] benefit from automated segmentation and audio-assisted frame classification, which adds complexity but also increases the robustness of their system under controlled conditions.

Building on the outlined framework, this work presents an innovative and quantitative approach to computer-assisted laryngeal paralysis assessment, aiming to enhance diagnostic reliability with potential extension to more complex cases, such as LC patients. It mainly builts upon the application and the analysis of AGATI [8], contributing to establishing robust frameworks for computer-assisted evaluation, advancing clinical outcomes in laryngology, and offering the following novel contributions to the current state-of-the-art:

- Providing an innovative technique to automatically select and extract the important video segments for a correct assessment, i.e., those showing both abduction and adduction of vocal folds.
- Introducing an innovative set of multimodal (i.e., combining images and statistical data) features to describe quantitatively and objectively vocal folds motility.
- Defining a model combining a Convolutional Neural Network (CNN) and a Multi Layer Perceptron (MLP), exploiting transfer learning and fine-tuning to classify patients based on their laryngeal motility. This classification model achieved an accuracy of 0.85 for binary patient-wise classification and 0.84 for multi-class patient-wise classification.

2. Methods

2.1. Methodology Overview

The methodology employed in this study encompasses an integrated approach to assess vocal folds motility through a reproducible pipeline with possible applications in cases of laryngeal disorders.

In consideration of the limitations highlighted in a recent research [25] on the scarce heterogeneity characterizing the datasets used in previous studies and the lack of available public repositories, in this work, we exploited a multi-centric dataset comprising 155 video-laryngoscopies that we collected in the Hospital Clínic de Barcelona and in the Azienda Ospedale – Università di Padova. Each of these video recordings, representing a different subject, was retrospectively analyzed and categorized as a healthy subject or a patient with laryngeal paralysis, making further distinctions between unilateral and bilateral paralysis [26]. To support accurate laryngeal motility assessment, we developed a robust pipeline that incorporates advanced feature extraction and classification techniques. At first, we extracted the vocal fold positions and the AGA values over frames using the AGATI toolbox [8]. A Gaussian Mixture Model (GMM) [27] was then used to filter each video for salient information, automatically extracting video segments (referred to as trials in this work) that show the full excursion of AGA values, from abduction to adduction. This resulted in a larger and less noisy dataset, as it was typically possible to extract more than one trial per patient while removing, for instance, frames where the vocal folds were not visible.

Each of these trials was then analyzed and characterized through a newly defined set of features, providing a quantitative and robust description of the patterns of vocal folds' movements represented. These features were then processed by a composite model combining a CNN and an MLP. To enhance performance, generalizability, and robustness, the backbone of the CNN is based on the ResNet-18, pre-trained on the ImageNet dataset provided by PyTorch [28]. This model was used to perform two different classifications: a binary classification (healthy vs paralyzed) and a multi-class classification (healthy vs unilateral paralysis vs bilateral paralysis). Evaluations were conducted on both trials and patients, with trial results aggregated to provide patient-level insights. An ablation study [29], which applied the Shapley value [30], enabled us to isolate the impact of each group of features on the model's performance, thereby ensuring the tool's efficiency and robustness.

2.2. Dataset

Our dataset comprised 155 in-office video-laryngoscopies, each representing a different subject with either normal or reduced laryngeal motility, recorded at a frame rate of 35 frames per second. We retrospectively collected this dataset across two different countries and clinical settings, including academic medical centers and community hospitals: the Hospital Clínic de Barcelona and the Azienda Ospedale – Università di Padova. After the preprocessing steps described below, this dataset yielded a total of 404 analyzable trials. This allowed us to enlarge the heterogeneity of our dataset with respect to the ones used in previous studies, capturing variations in patient demographics and clinical practices, thereby enhancing the generalizability and robustness of our findings. Inclusion criteria were: 1) patients in-office endoscopically assessed for laryngeal motility; 2) availability of recorded video-laryngoscopies 3) transnasal flexible video-laryngoscopic assessment. For this pilot study, for homogeneity reasons, LC patients were excluded. The decision to exclude such patients stems from the need to proceed step by step and develop expertise within our working group, starting with simpler cases in which the mucosal morphology of the larynx is not altered by a neoplasm. Similar to other studies [17], patients were asked to make the "ee" sound (/i:/) followed by a deep inhalation to allow the assessment of maximal adduction and maximal abduction of their vocal folds. This led to a bimodal distribution of the values assumed by the AGA. Stroboscopic examinations were not included in the dataset for higher homogeneity, as the vocal folds movements are sampled with low frequency in those cases. Based on the collected videos, patients were categorized into three classes: unilateral laryngeal paralysis (n=68), bilateral laryngeal paralysis (n=50), and normal healthy laryngeal function (n=37). In particular, the first two classes can be further grouped together in a broader class of patients with paralysis. To assess laryngeal paralysis and establish the ground truth labels, clinicians reviewed the video-laryngoscopies and reached a consensus on whether to classify patients as healthy, unilaterally paralyzed, or bilaterally paralyzed. When a consensus was not found, the mode among all the different ratings was used to define the final category. Data were acquired with informed consent following the principles of the Helsinki Declaration, and ethical approval was obtained by the local ethical committee of each of the hospitals involved[1]. Patients' identification data were anonymized.

[1] Hospital Clínic de Barcelona Reg. HCB/2023/0897 and Azienda Ospedale – Università di Padova 190n/AO/21.

2.3. AGATI Toolbox

We began by extracting the first features from video-laryngoscopies using the AGATI toolbox to track the position of each vocal fold across frames. The AGATI toolbox, built upon the DeepLabCut toolbox [31], implements a deep learning-based object detection and tracking algorithm for markerless pose estimation, which enables the tracking of vocal fold positions in video-laryngoscopies [8]. Furthermore, always based on the positions of vocal folds, the AGATI toolbox allows the extraction of the values of the following features: anterior glottic angle, left and right vocal fold angles, and velocity and acceleration of adduction and abduction of the vocal folds.

The most robust feature provided by the AGATI toolbox was the AGA value, which effectively handles intricate patterns of noise, including rotational and translational blends, as well as scale-varying distortions. Indeed, in principle, the AGA value is not influenced by rotational and translational noise, unlike left and right angles, which are defined in the framework of the AGATI toolbox as the angles that extend from each vocal fold to an imaginary vertical half-line. However, in some cases, distinguishing between unilateral paralyzed patients and normal healthy controls using the AGA value was insufficient, especially when measurement was affected by noise. That's why we needed more robust and reliable features together with a method to increase the "signal-to-noise" ratio, focusing on important parts of the videos while filtering out the noise.

2.4. Automatic Trial Selection

In a similar perspective to the selection of salient frames performed manually in the work by Villani and colleagues [23] and by an AI classification model in the work by Yao and colleagues [32], Baldini and colleagues [33] and Zhang and colleagues [24], we automatically extracted trials from each video recording. In this way, we extracted the salient part of the videos, removing all those parts where, for instance, the vocal folds were not framed.

The idea behind this was as follows: to perform an accurate assessment of laryngeal motility, only a few seconds per video are relevant. Still, these seconds have to show the entire excursion of the vocal folds, from complete abduction (first condition) to complete adduction (second condition). These two conditions can be represented quantitatively by examining the lowest and largest AGA values obtained using the AGATI toolbox. Specifically, complete adduction is represented by the lowest values of the AGA, whereas complete abduction is represented by its largest values.

To automatically identify informative segments of the video-laryngoscopies, hereafter referred to as "trials", we used the frame-wise AGA time series extracted by the AGATI toolbox. A video segment was considered informative if it contained at least one full opening–closing cycle of the vocal folds, i.e., a complete excursion of AGA values from abduction to adduction.

For each video, we modeled the empirical distribution of AGA values using a two-component Gaussian Mixture Model (GMM) [34, 35], which approximated the two extreme states of the vocal folds (abducted vs. adducted), as illustrated in Fig. 1. Candidate trials were defined around transitions between the two components of the GMM in the AGA time series, corresponding to the opening–closing cycles of the vocal folds.

To ensure that the trials captured full excursions and were not dominated by noise, we retained only those segments whose AGA values spanned from below the mean of the first Gaussian (abducted position) to above the mean of the second Gaussian (adducted position). Additionally, to enhance robustness, we required each trial to meet a minimum duration threshold of 100 frames (approximately 4 seconds), with at least 10% of the frames below the first Gaussian mean and at least 10% above the second Gaussian mean. This ensured that only full cycles of motion were included, preventing the inclusion of overlapping or incomplete trials.

By applying these criteria, we ensured that the trials automatically extracted from each video comprehensively represented the full range of vocal fold motion, from complete abduction to adduction, thus enhancing the reliability of our findings. These criteria were predefined and applied consistently across all subjects, enabling the extraction of multiple informative trials from each video-laryngoscopy. This approach ultimately increased the effective dataset size while focusing on the most relevant segments for laryngeal motility assessment.

Through this model, we were able to identify the frames where the vocal folds were at maximum abduction or adduction. This kind of feature selection was conducted for each patient independently of their class. On average, 4.21 trials were extracted per video, giving a larger dataset comprising 404 trials (135 referring to subjects with normal healthy motility, 198 referring to patients with unilateral paralysis, and 71 referring to patients with bilateral paralysis).

2.5. Definition and Extraction of Innovative Features

After defining the trials, we conducted feature extraction to provide a quantitative description of vocal fold movements, enabling differentiation between classes. With this aim, we considered the maximum and the minimum

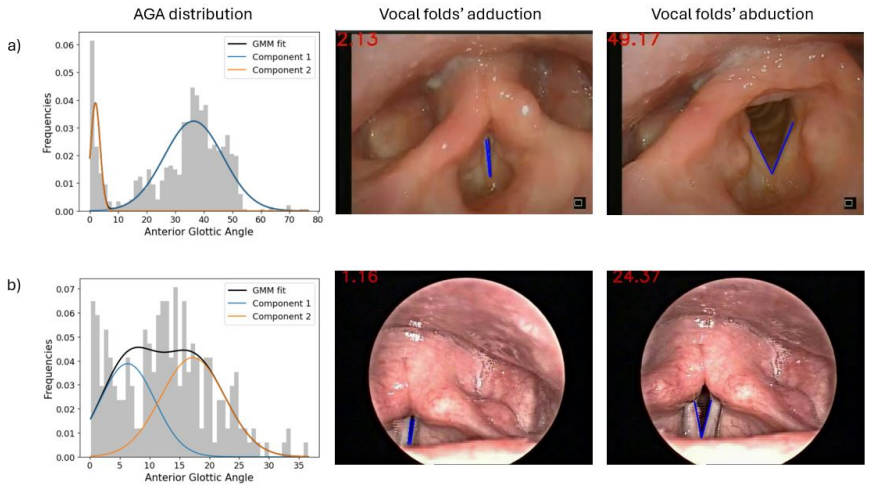

Figure 1: Fit of the distribution of the Anterior Glottic Angle (AGA) using a Gaussian Mixture Model (GMM) consisting of a mixture of two Gaussian distributions. Row (a): subject with healthy motility; row (b): patient affected by laryngeal paralysis, where abduction and adduction are inevitably more similar. AGA values are shown in red in the top-left part of each frame.

AGA values, i.e., those characterizing maximal adduction and abduction of the vocal folds. The AGA feature was suggested by doctors in our research team. Furthermore, it was already highlighted as important for laryngeal motility assessment in [8, 17], as it provides important information when distinguishing between healthy subjects and patients with paralysis. Then, we computed its standard deviation, and, in a similar manner as described in the previous subsection, the difference of the two means extracted through the GMM, which was applied in this case to trials. In order to give a suitable description that could allow the identification of the asymmetrical movements of vocal folds in cases of unilateral paralysis, we computed Kendall's correlation [36] and Granger's causality [37] between left and right angles. Here, Kendall's correlation measures the similarity of the rankings characterizing the values assumed by right and left angles, while Granger's causality quantifies whether one of the two time series (the left angle or the right angle) anticipates the other. Additionally, we compared the dispersion of the values assumed by left and right angles in each trial, examining their difference and the lowest values of their standard deviations. Although some of these features are computed based on noisy estimates of left and right angles, the way we combine these estimates makes our features more reliable. For instance, if the vocal folds are not aligned with the vertical half-lines, the estimates of right and left angles given by AGATI are wrong and noisy, while the standard deviation of the values assumed by the two is way more reliable. For a more mathematical definition of each of these features, we refer the reader to the Appendix.

Finally, similarly to what was done in the context of facial paralysis [38], an image feature accounting for the difference in the positions of the entire anatomical region surrounding and including the vocal folds was extracted. Specifically, from each trial, we extracted the frames associated with maximal abduction and adduction. Then, we standardized the intensity distribution across pixels. After subtracting the frames and computing the absolute difference, we rescaled the pixel intensities to a range of 0 to 255. As an example, in Fig. 2, we sketch the definition of one of those image features obtained starting from the images of maximal abduction and adduction of the same trial.

2.6. Classification Model

The feature extraction process yielded a multimodal set of features that needed to be input into a classification model to categorize patients based on their laryngeal motility.

Since our set of features included both images and numerical statistical features, we employed a model combining a CNN, commonly used in medical image analysis [39, 40, 41], and an MLP.

Although the dataset size was limited with respect to typical deep-learning requirements, several design choices were adopted to mitigate overfitting and improve generalization. Specifically, we employed transfer learning, fine-tuning a ResNet-18 backbone pre-trained on ImageNet, a trial-based data augmentation strategy through automatic segmentation of each video into multiple informative trials, and subject-wise cross-validation with a weighted loss function.

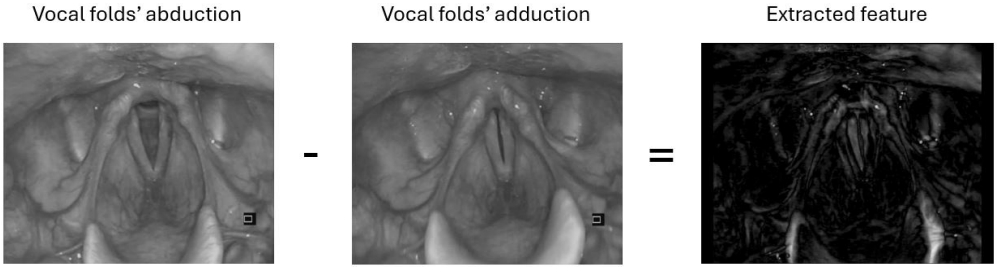

Figure 2: In the right image, we show the image feature obtained by comparing the two left images representing the vocal folds in maximal abduction and maximal adduction. The resulting feature is an image representing the motion and changes in anatomical configuration between the two poses, where white pixels indicate the areas with the highest difference between the two frames.

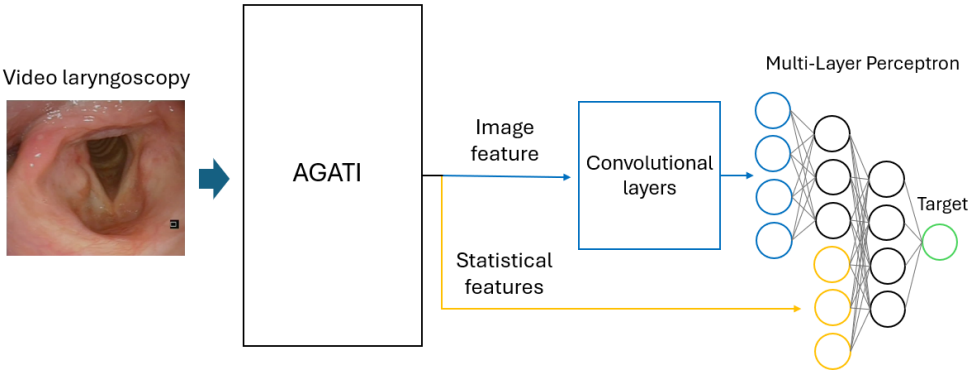

Figure 3: The figure illustrates the classification pipeline utilized for patient categorization. Initially, video-laryngoscopies were processed through the AGATI toolbox, followed by trial definition and feature extraction. Subsequently, a combined CNN/MLP model was used to optimize predictions by leveraging the extracted image features, along with other statistical features, in an end-to-end model. This model passed the image feature to the Resnet-18 convolutional layers [43]. The output of the CNN was then passed to an MLP layer for dimensionality reduction. Finally, after concatenating all the remaining features, the model produced the prediction using a few additional MLP layers.

Since, as shown in Fig. 2, the image features exhibited relatively simple patterns, we chose to build the model on one of the simplest yet effective pre-trained CNNs available in PyTorch, namely ResNet-18, pre-trained on the ImageNet dataset. More specifically, we fed the image feature into the convolutional layers, followed by an MLP layer, aimed at achieving a proper non-linear dimensionality reduction of the 512 features output by the ResNet. The output of this MLP layer was then concatenated with all the other features. Finally, a second set of MLP layers, whose number was optimized through a cross-validation described in the following subsection, was applied to obtain the final prediction. This resulted in an end-to-end model that we used for both binary and multi-class classification, where the difference between the models was only in the number of output neurons, which was 1 for binary classification and 3 for multi-class classification. The entire pipeline, including the model we described, is sketched in Fig. 3.

In alignment with recent research demonstrating the substantial benefits of transfer learning for enhancing the robustness and stability of medical imaging models [42], we applied this approach to our model. Specifically, to further evaluate its impact, we also trained the same model using random weight initialization (i.e., without applying transfer learning) under identical hyperparameter tuning settings. This allowed us to directly compare the performance improvements facilitated by transfer learning (through pretraining). The results of this comparison can be found in the Appendix.

Hyperparameter	Values
Batch Size	{4, **8**, 16}
Learning Rate	{0.01, **0.001**}
Number of MLP Neurons	{[8], [32], **[32, 8, 3]**, [8, 32, 1], [8, 32]}
CNN Output Dimension	{1, **8**}

Table 1
Hyperparameter search space with optimization of the following hyperparameters: batch size, learning rate, number of neurons in the second MLP, and dimension of features outputted by the first MLP. The selected values are shown in bold for multi-class classification and underlined for binary classification.

2.7. Training and Evaluation Strategies

First, subjects were partitioned into a training/validation set (80%) and an independent held-out test set (20%) using stratified sampling to preserve class proportions. The test set was used exclusively for the final evaluation and was never involved in model selection or hyperparameter tuning. Then, a 5-fold subject-wise cross-validation was employed to partition the training/validation sets, avoiding data leakage and ensuring that all trials belonging to the same patient were always assigned to the same subset, thereby enhancing generalizability. A grid search on a predefined set of hyperparameters (Table 1) was subsequently applied. Besides batch size and learning rate, our grid search allowed us to optimize the number of neurons in the second MLP, as well as the number of features output by the first MLP, which was applied to reduce the dimensions of the CNN output features.

At the start of each iteration of the cross-validation, a combination of the above hyperparameters was selected, and the five models were trained. We employed Adam optimization and used the cross-entropy loss for multi-class classification and the binary cross-entropy loss for binary classification, both with class-specific weights to mitigate the imbalanced influence of each class. Moreover, to address potential class imbalances, we maintained the same proportions of classes' occurrences in every set. Furthermore, we assigned a weight equal to the ratio of the total number of examples to the total number of positive examples for each class. In this way, a weighted loss function was defined to balance the training process. The combined model was trained using an optimization algorithm with a learning rate specified by one of the hyperparameters, for up to 1000 epochs. An early stopping criterion was used to terminate training when the model failed to show improvements in the validation loss for more than 20 epochs, while ensuring the model was trained for a minimum of 50 epochs.

Patient-level predictions were derived by averaging trials' (belonging to the same patient) assignment probabilities. The class with the highest probability was then assigned to each patient. Finally, results were evaluated both at the patient and trial level. Evaluations were performed on the test set across all five models obtained through model optimization on the five folds, ultimately yielding a distribution of performance metrics from which we computed the median and the Median Absolute Deviation (MAD).

2.8. Ablation Study

In order to study the importance of the newly introduced features, we performed an ablation study by systematically removing either individual features or combinations of features. This allowed us to assess the contribution of each component to the overall model performance, thereby identifying the most critical subsets of features that drive accurate predictions in both binary and multi-class classification tasks.

Following the Shapley value approach [30], a widely adopted explainability approach in medical imaging used to quantify feature contribution and improve model transparency [44], we defined 4 players, defined as groups of one or more features, each one allowing to assess the relevance of the newly defined features to the predictions of the models. We defined the four players as follows:

- The first player includes only one feature: the image fed into the CNN. This allows us to check whether the inclusion of this feature, which led us to increase the complexity of the model, using the CNN, is indeed necessary.
- The second player includes those features that were already found as relevant for laryngeal motility assessment in other studies in the literature. This player includes the minimum and maximum values of the AGA for each sample. This allows isolating the contribution of the traditionally used features while, at the same time, determining if the newly defined features led to an improvement in the model's performance.

Metric	Trials	Patients
Precision	0.73±0.03	0.83±0.05
Sensitivity	0.73±0.01	0.85±0.07
F1-score	0.73±0.01	0.84±0.05
Balanced Accuracy	0.73±0.01	0.85±0.07

Table 2
Median and Median Absolute Value (MAD) of the performance obtained by the binary classification model across the 5 folds. Here we report the values of Balanced Accuracy, F1-score, Sensitivity, and Precision. We do this for both trials and patients, where predictions from several trials related to specific subjects are aggregated.

- The third player includes the difference between the two means of the two Gaussians found by the GMM model, the standard deviation of the AGA value distribution, and the minimum standard deviation of the two angles, left and right. These features should allow to distinguish between healthy control and patients with bilateral paralysis.
- The fourth player includes the correlation and Granger's causality values between the values assumed by the right and left angles, together with the difference of the standard deviations of left and right angles. These features should be critical to distinguish subjects with unilateral paralysis.

In the following, we will refer to these groups of features as player 1, 2 , 3, and 4.

Following the Shapley value approach, our model was trained on all possible coalitions (subsets) of these players, including those comprising only one player. As described above, for each coalition, the model was trained through a combination of 5-fold cross-validation and grid search, with the hyperparameters considered to be the same as those shown in Table 1. These hyperparameters were chosen by balancing the optimization of model performance while minimizing changes to the model. The results obtained by the model for each coalition were then combined to obtain the average marginal contribution, namely the Shapley value, of each player i, expressed by the following formula:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left(v(S \cup \{i\}) - v(S)\right) , \tag{1}$$

where N is the number of players, S is any coalition that does not include the player i, and $v(S)$ is the balanced accuracy obtained by the model on the coalition S [45].

3. Results

3.1. Model Performance

Standard metrics were used to evaluate the outcome as previously reported for this field [46]. After automatic trial identification and extracting features, we proceeded to train the classification model. The dataset splitting was based on subject IDs, as multiple trials could be extracted from a single patient. We first trained a binary model to classify healthy subjects and patients affected by laryngeal paralysis, the latter group including both unilateral and bilateral paralysis. We assessed the variability of the performance of the model on the test set for both the cases of trial and patient classification computing the median and the MAD of the performance of the model resulting from each fold[2] to assess the impact of performance outliers. Our binary model reached a median balanced accuracy of 0.73 on trials and 0.85 on patient classification. The metrics obtained are shown in Table 2, while in Fig. 4 we show the median confusion matrices related to trials and patients.

Then, we employed the same model introduced in Subsection 2.6 to perform a multi-class classification of unilateral laryngeal paralysis, bilateral laryngeal paralysis, and healthy laryngeal function. For the multi-class classification, we achieved a balanced accuracy of 0.70 for trials and 0.83 for patients, based on the median across the 5 folds. The detailed

[2]I.e., given the optimal choice of the hyperparameters, for each fold the model was trained on the whole dataset excluding the test set and the validation set associated with that fold.

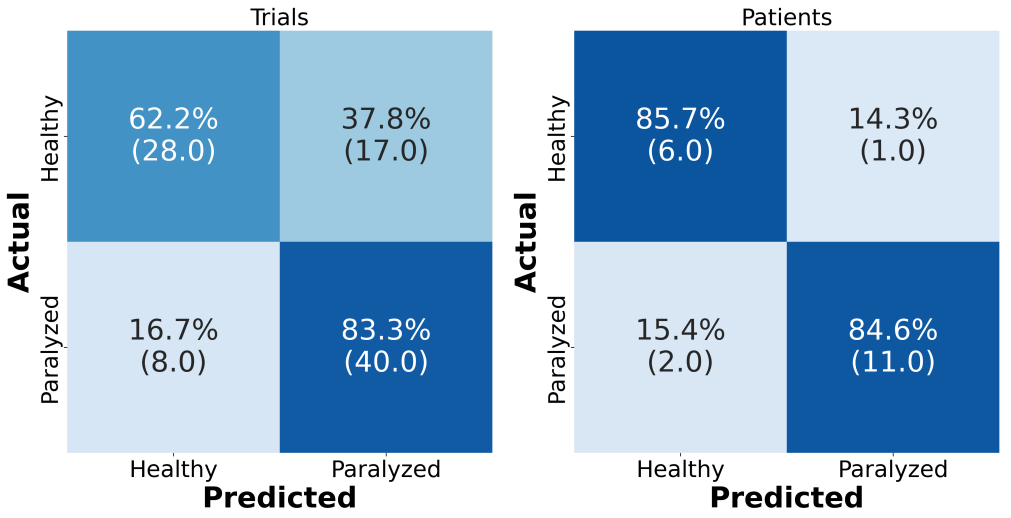

Figure 4: Median confusion matrices obtained by the binary classification model across folds. Here, each entry shows the percentage of subjects actually categorized in the class characterizing its row and predicted as the class assigned to its column. The median is computed between the different percentages obtained across folds.

	Trials			Patients		
Metric	**Healthy**	**Unilateral**	**Bilateral**	**Healthy**	**Unilateral**	**Bilateral**
Precision	0.78±0.02	0.59±0.01	0.55±0.05	0.75±0.05	0.83±0.08	0.83±0.05
Sensitivity	0.64±0.04	0.67±0.04	0.89±0.12	0.86±0.00	0.62±0.08	1.00±0.10
F1-score	0.69±0.02	0.60±0.02	0.67±0.07	0.80±0.03	0.71±0.07	0.91±0.07
General Performance	**Balanced Accuracy:** 0.70±0.03			**Balanced Accuracy:** 0.83±0.06		

Table 3
Median and Median Absolute Value (MAD) of the performance obtained by the multi-class classification model across the 5 folds. Here, we report the values of the one-vs-rest Balanced Accuracy, F1-score, Sensitivity, and Precision obtained in a one-vs-all (micro) approach. This is done both for trials and patients, where the predictions on the several trials related to specific subjects are aggregated.

class-wise performance is presented in Table 3, while in Fig. 5 we show the median confusion matrices related to trials and patients.

For more detailed results on the model's performance for each fold, please refer to the Appendix.

3.2. Results of the Ablation Study

As described in Subsection 2.8, we performed an ablation study through a Shapley value approach both for the binary and the multi-class classification. This allowed us to assess whether the newly defined group of features was actually informative for predicting laryngeal paralysis.

These results were combined to compute, first for binary classification and then for multi-class classification, the Shapley value of each player using Equation (1). The values obtained (all of which turned out to be positive) are depicted in Fig. 6.

For more detailed results about the performance of the models trained during the ablation study, considering the different players, please refer to the Appendix.

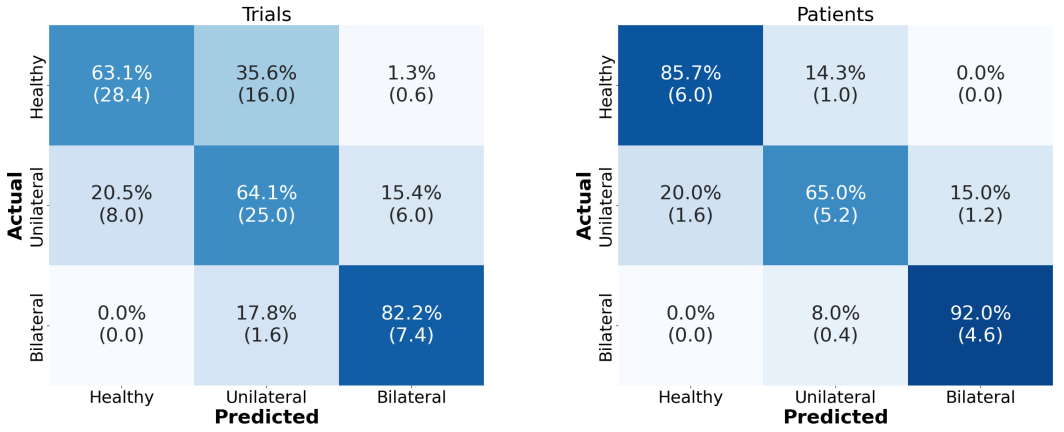

Figure 5: Median confusion matrices obtained by the multi-class classification model across folds. Here, each entry shows the percentage of subjects actually categorized in the class characterizing its row and predicted as the class assigned to its column. The median is computed between the different percentages obtained across folds.

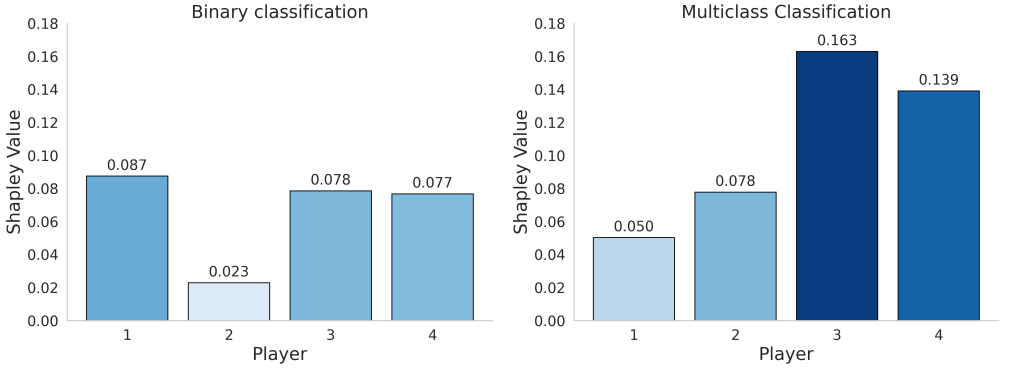

Figure 6: Here we show, for both binary and multi-class classification, the Shapley value of each player, computed through Equation (1). In the left plot, we show the Shapley values of the players in the binary classification. In the right plot, we show the Shapley values of the same players in the multi-class classification. Color intensities are defined to be proportional to Shapley values, ranging from 0, the value that would be assumed by a player that does not contribute to the classification, to the maximum Shapley value observed in the two bar plots.

4. Discussion

4.1. Interpretation of Results and Implications

In this paper, we presented an innovative pipeline to increase the reliability and transparency of automatic laryngeal endoscopic motility assessment using AI-based techniques. By leveraging a dataset of 155 video-laryngoscopies and the AGATI toolbox, we were able to extract meaningful features from each video, such as the trajectories of the vocal folds. These features provided a quantitative description of vocal folds' motion, which was used in combination with a Gaussian mixture model to extract trials representing the full excursion of the vocal folds. In principle, these trials contained all the necessary information for a correct assessment of each subject. However, this step made it possible to reduce noise, focusing only on relevant segments of the video, while augmenting the size of the available dataset, allowing for the extraction of multiple trials from each video-laryngoscopy. From these trials, by combining the outputs of the AGATI toolbox in a convenient manner to reduce noise and represent the three classes we considered, we extracted new, additional features that quantitatively described subjects' laryngeal motility. These features were fed into a model that combined a convolutional neural network based on ResNet-18, pre-trained on the ImageNet dataset, and a multi-layer perceptron to leverage the multimodal nature of the feature set. Transfer learning was performed to improve

generalizability, given the limited amount of available data for our specific classification tasks, in line with recent work showing that transfer learning can substantially improve robustness and stability of medical imaging models [42]. In the multi-class classification task, the pre-trained model consistently outperformed the non-pre-trained model (see the Appendix). Specifically, the pre-trained model showed a higher balanced accuracy across all classes: Healthy, Unilateral, and Bilateral. These results highlight the advantage of transfer learning in improving classification. The non-pretrained model, due to its limited data, struggled to generalize effectively, resulting in slightly lower performance. The binary classification model achieved better performance than other similar studies in the literature. While our performance is slightly better than that reported in [23], it is somewhat worse than the results in [24]. This difference could be attributed primarily to the much larger sample size in [24], as well as differences in trial extraction methods. While this work is undoubtedly valuable and introduces an interesting trial selection procedure, we believe our approach offers novel and useful techniques that go beyond mere performance comparison. Specifically, in [24], trials were extracted using audio data, which is a useful method but not always reliable, especially in cases of vocal fold paralysis. In such cases, the sound-related information is often so compromised that patients require specialized treatments to regain their ability to speak [47], a particular challenge in bilateral fold paralysis. Our approach, which combines AGATI [8] with automatic trial extraction, circumvents this limitation.

Furthermore, our multi-class classification model enables finer granularity in classifying laryngeal paralysis by distinguishing between monolateral and bilateral paralysis, which is an important contribution not addressed, for instance, by Villani et al. [23]. Notably, this added level of detail was achieved without a significant drop in performance: at a patient level, the binary classification model reached a balanced accuracy of 0.85, while the multi-class classification model achieved 0.83. These findings indicate that our newly defined features are highly effective for achieving more granular classification.

More insights on these results were obtained through a Shapley value approach. In both classifications, all the players defined are characterized by Shapley values greater than 0, indicating that all the groups of features defined are relevant and contribute to the classification. In the left plot, related to binary classification, the most relevant group of features appears to be the first player, i.e., the image feature. Indeed, in patients with laryngeal paralysis, the whole area surrounding and including the vocal folds hardly moves compared to patients with healthy laryngeal motility. This results in very different images. At the same time, in the right plot, related to multi-class classification, the highest Shapley value is assigned to the third player, a group of newly defined features that include correlation, Granger causality, and the difference between the standard deviations of the left and right angles. This group of features was indeed designed with the goal of helping in the multi-class classification to recognize subjects with unilateral paralysis. Interestingly, the features identified as relevant in the literature, namely the minimum and maximum AGA values [17, 18], characterizing the second player, were never found to be the most relevant. Probably, the correlation between some of the features belonging to the second and fourth players played a role in this. However, the higher Shapley value obtained in both classifications by the fourth player with respect to the second suggests a higher reliability of the fourth player's features compared to those highlighted in the literature. Overall, our study's results demonstrate the efficacy of the introduced features in quantitatively describing laryngeal motility in patients with normal superficial anatomy, as well as our model's capability to discriminate between the various classes.

The contribution of this paper is not limited to improving the automatic laryngeal motility assessment compared to the state-of-the-art, as it also aims to provide reliable features for describing laryngeal motility, potentially inspiring new studies and suggesting a more objective way to diagnose laryngeal motility disorders. This is essential since traditional laryngoscopic evaluations are subject to inter-observer variability, which can lead to a tremendous rate of inconsistent diagnoses, e.g., in the case of LC patients. In the paper by Ferrari et al. [6], 22 clinicians evaluated 366 videos of laryngeal cancer patients, for a total of 2170 evaluations. The concordance of clinical rating was excellent in only 22.7% of cases, with an overall weak inter- and intra-rater agreement. In this context, AI-driven systems may offer standardized assessments, thereby reducing subjectivity and improving diagnostic precision. Building on the intuition by DeVore and colleagues [48] that AGA is statistically different between patients with laryngeal paralysis and healthy individuals, our work has the merit of developing an automatic pipeline for direct motility assessment and disease classification. Particularly in the work by Villani et al. [23], the authors addressed as a potential limitation of their method the time-consuming manual annotation of laryngeal anatomical key points and suggested that, as future work, to support clinicians in the actual clinical practice, a classification model could be included within other computer-assisted algorithms for frame selection and automatic key-points regression. This is what we have achieved in our work: a pipeline capable of automatically identifying laryngeal abduction and adduction poses, extracting trials, and classifying paralysis, all without requiring any manual input from physicians. This is particularly relevant in light of a

streamlined workflow and efficiency view, where automating the analysis of laryngeal motility could reduce the time required for manual assessments, allowing clinicians to allocate more time to patient care. This efficiency could lead to increased throughput in clinical settings and potentially reduce healthcare costs, likewise in other medical applications of AI to medicine, such as breast cancer detection [49, 50, 51].

4.2. Limitations

As already highlighted, video laryngoscopies are significantly affected by various and complex patterns of noise. During registration, neither the patient nor the doctor relies on fixed supports to stabilize their positions. Furthermore, the camera is mounted on flexible supports and can capture vocal folds from various angles. This gives rise to very complex patterns of noise that standard correction methods are not able to correct. This significantly impacts the reliability of the results obtained by each of the papers analyzed that used this kind of data. AGATI and its features are no exception in this regard. We attempted to mitigate the impact of noise on the data by defining the features in a convenient manner; however, much remains to be done. Our assumption, anyway, was that the noise patterns did not affect the classification model's performance, as noise should not be related to specific classes; hence, its influence should be evenly distributed.

Another limitation of the study can be attributed to the technique we used to extract the trials. Indeed, while this appears to perform very well on subjects with healthy motility and monolateral paralysis, it does not perform as effectively on patients with bilateral paralysis, resulting in the lack of usable trials from some of the video laryngoscopies. Those patients are indeed characterized by a distribution of AGA values that is more approximable by a uniform distribution than by a bimodal distribution. Moreover, in this case, the values within the uniform distribution are assigned in a more random manner, since even AGATI works worst for patients with bilateral paralysis, and the extreme values typically do not correspond to frames representing maximal abduction and adduction. A possible improvement to the work can be achieved by retraining the AGATI model to make it more robust and less sensitive to noise. It is essential to emphasize that this technique is designed to aid physicians in selecting relevant trials for evaluation. In this, we were conservative, opting for a more reliable and reproducible across-classes algorithm that retains only the best trials, i.e., those that we could expect to actually contain the information required. This inevitably excluded those trials that were more subtle. Moreover, additional features, possibly less sensitive to noise, can be introduced to describe laryngeal motility. Specifically, our model performs well if the movements of the vocal folds can be extracted from the videos, while disregarding other movements of the anatomical region surrounding the vocal folds that are not considered important for clinical decision-making. However, clear visibility of the vocal folds is not always possible since laryngeal cancer patients typically have tumors that cover partially or completely this area. This is the reason why we introduced the image feature described above, which, to our knowledge, is the first in the literature to allow for the consideration of broader clinically significant movements in the area surrounding the vocal folds, potentially enabling applications even in cases of laryngeal cancer patients. However, the definition of additional robust features, intended to go in this direction, is still required. Hence, the adoption of AI techniques to classify laryngeal motility in such patients remains an unmet need.

4.3. Future Directions

We intend to focus future works on validating our approach in larger and more diverse patient populations while exploring its applicability to additional laryngeal disorders where the image feature could turn out to be crucial. Another important validation we plan to conduct is a further comparison between AI outputs and clinician evaluations, with the goal of identifying cases where uncertainty is higher and determining which method yields the best performance. Indeed, as highlighted in recent reviews on AI in upper aerodigestive tract endoscopy and beyond, this is a necessary step for any real-world deployment [52, 53, 54]. Hence, we believe that in the future, AI models will be crucial for detecting subtle changes in laryngeal motility that may be imperceptible to the human eye, facilitating the early identification of disorders such as vocal fold paralysis or spasmodic dysphonia, where early detection is essential for timely intervention and improved patient outcomes. In oncology, the impairment of laryngeal motility is a well-established risk factor with a pivotal role in influencing the treatment strategy and patient's prognosis [55, 56]. By providing quantifiable metrics of laryngeal function, AI systems would enable clinicians to precisely tailor the treatment for each case and monitor disease progression or response to therapy with greater objectivity. In the future, the application of AI in analyzing laryngeal motility via laryngoscopy might offer a transformative approach to otolaryngological practice, enhancing diagnostic accuracy, enabling early detection, providing objective monitoring, supporting education, improving efficiency, and ensuring standardization. These advancements collectively could

significantly contribute to improved patient care and outcomes. We believe that our work represents a significant step forward in this. Moreover, it is the intention of this working group to proceed further with the use of AI for the classification of laryngeal motility and/or posterior laryngeal extension (which can cause an impairment of motility but is clinically more relevant than the latter) in the case of laryngeal cancer patients, possibly through a multi-omics approach (in the sense that the information provided to AI would come not only from videos but also from other sources such as audio, clinical notes, and imaging).

5. Statements and Declarations

5.1. Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

5.2. Ethics Statement

Data were acquired with informed consent following the principles of the Helsinki Declaration, and ethical approval was obtained by the local ethical committee of each of the hospitals involved[3]. Patients' identification data were anonymized.

5.3. Funding Sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. G. Gnecco was partially supported by the project "THE – Tuscany Health Ecosystem" (CUP: D63C22000400001), funded by the European Union – Next Generation EU program, in the context of the Italian National Recovery and Resilience Plan, Investment 1.5: Ecosystems of Innovation, and by the PRIN PNRR 2022 project "MOTUS" (CUP: D53D23017470001), funded by the European Union – Next Generation EU program.

5.4. Statement on Why Research Data Cannot be Published

The raw data (patients' videos) cannot be published due to privacy concerns.

[3]Hospital Clínic de Barcelona Reg. HCB/2023/0897 and Azienda Ospedale – Università di Padova 190n/AO/21.

Appendix A. Mathematical Formulation of Features

Here, we provide additional details regarding the mathematical formulation of the features used to quantitatively describe vocal folds' motility. In the following subsections we denote the values at time t of the Anterior Glottic Angle (AGA), the left angle, and the right angle as $x_{\mathrm{AGA},t}$, $x_{\mathrm{left},t}$, and $x_{\mathrm{right},t}$, respectively.

These features were considered also in light of the increments of the left and right angle values from frame t to frame $t+1$:

$$\dot{x}_{\mathrm{left},t} = x_{\mathrm{left},t+1} - x_{\mathrm{left},t}\,, \quad \text{and} \quad \dot{x}_{\mathrm{right},t} = x_{\mathrm{right},t+1} - x_{\mathrm{right},t}\,. \tag{A.1}$$

These derived features, in the formulas that follow, were also used in place, respectively, of the values of right and left angles. However, such derived features were not included in the final version of the main manuscript, as being them strongly affected by noise.

Appendix A.1 Statistical Features Based on the Anterior Glottic Angle

The first features we included in the study (in consideration of other works in the literature) are the maximum and the minimum AGA values that represent respectively the angles of maximal abduction and adduction:

$$\max_t x_{\mathrm{AGA},t}\,, \quad \text{and} \quad \min_t x_{\mathrm{AGA},t}\,. \tag{A.2}$$

Also, we took into consideration the standard deviation of the AGA values distribution:

$$\sqrt{\frac{1}{N}\sum_t (x_{\mathrm{AGA},t} - \bar{x}_{\mathrm{AGA}})^2}\,, \tag{A.3}$$

where N is the length of the AGA values time series extracted from the trial, and $\bar{x}_{\mathrm{AGA}}$ is the mean value of the AGA values distribution across time.

Finally, we fitted the AGA values distribution by means of a Gaussian Mixture Model (GMM) with two Gaussians, then we computed the absolute value of the difference between the two means characterizing the two fitted Gaussian distributions:

$$\left| \frac{\sum_t \gamma_1(x_{\mathrm{AGA},t}) \cdot x_{\mathrm{AGA},t}}{\sum_t \gamma_1(x_{\mathrm{AGA},t})} - \frac{\sum_t \gamma_2(x_{\mathrm{AGA},t}) \cdot x_{\mathrm{AGA},t}}{\sum_t \gamma_2(x_{\mathrm{AGA},t})} \right|. \tag{A.4}$$

In this case, $\gamma_1(x_{\mathrm{AGA},t})$ and $\gamma_2(x_{\mathrm{AGA},t})$ are the posterior probabilities that the data point $x_{\mathrm{AGA},t}$ belongs, respectively, to the first or to the second of the two Gaussian distributions. In general, $\gamma_k(x)$, i.e., the posterior probability of a data point x to belong to the k-th component, can be expressed in the following manner:

$$\gamma_k(x) = \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)}\,, \tag{A.5}$$

where π_k is the mixing coefficient for the k-th Gaussian, representing the prior probability that a general observation x belongs to it, $\mathcal{N}(x|\mu_k, \Sigma_k)$ is the probability density function of the Gaussian distribution with mean μ_k and covariance Σ_k evaluated at the data point x, and K is the total number of components in the mixture model.

Appendix A.2 Statistical Features Based on the Left and Right Angles

To assess the capability of patients to move in a symmetric and synchronized manner the left and right vocal folds, we extracted features related to left and right angles. We started by computing the standard deviations of the left and right angles. Then, we considered the minimum of these two standard deviations and the absolute value of their

difference:

$$\min\left(\sqrt{\frac{1}{N}\sum_{t}(x_{\text{left},t}-\bar{x}_{\text{left}})^2},\sqrt{\frac{1}{N}\sum_{t}(x_{\text{right},t}-\bar{x}_{\text{right}})^2}\right), \tag{A.6}$$

$$\left|\sqrt{\frac{1}{N}\sum_{t}(x_{\text{left},t}-\bar{x}_{\text{left}})^2}-\sqrt{\frac{1}{N}\sum_{t}(x_{\text{right},t}-\bar{x}_{\text{right}})^2}\right|. \tag{A.7}$$

Finally, we computed the Kendall's correlation and Granger's causality between the absolute values of the left and right angles [36, 37], as well as between the increments in time of the two.

Specifically, Kendall's correlation coefficient τ is computed as follows:

$$\tau = \frac{\text{number of concordant pairs} - \text{number of discordant pairs}}{\text{total number of pairs}}, \tag{A.8}$$

where, for each pair of variables (say, X and Y), the number of concordant pairs refers to pairs of observations that have the same ordering with respect to both variables (X and Y), while the number of discordant pairs refers to pairs of observations that have opposite ordering with respect to the two variables.

Finally, Granger's causality refers to the fact that, in a linear regression model, past values of X have a statistically significant effect on the current value of Y, taking into account also past values of Y as regressors. The null hypothesis for Granger's causality test is that the first time series, X, does not Granger cause the second time series, Y. Such a null hypothesis is rejected if the p-value of Granger's causality test is smaller than a desired α level. Specifically, we took into consideration the maximum value of the χ^2 statistics used for Granger's causality test, resulting from its application to the left and right angles and vice versa, with a maximum time lag equal to 1. This was because the two angles were expected to change in an approximate synchronous manner. For this part of the analysis, we used the Python function "statsmodels.tsa.stattools.grangercausalitytests".

Appendix B. Detailed Results

Appendix B.1 Results in Terms of Confusion Matrices, Receiver Operating Characteristic (ROC) Curves and Area Under the Curve (AUC) Scores

We performed a training involving a combination of a 5-fold cross-validation approach and a grid search on a predefined set of hyperparameters. We also computed patient-wise and trial-wise performance metrics. The confusion matrices shown in the main paper are the aggregate results of the confusion matrices shown in the following pages for both binary classification (see Figs. B.1 and B.2) and multiclass classification (see Figs. B.4 and B.5). Mean Receiver Operating Characteristic (ROC) curves and median Area Under the Curve (AUC) scores of the 5 folds for binary classification (Fig. B.3) and multi-class classification (Fig. B.6) are also shown.

Appendix B.2 Comparison with the Results Obtained Using a Non-Pretrained ResNet

To evaluate the effectiveness of using transfer learning to increase the performance of our model, we conducted a comparison between our ResNet-18-based model (pre-trained on ImageNet) and a version of the same architecture without pre-training. Same models and hyperparameter tuning settings were considered here, though CNN's weights were randomly initialized.

For the multiclass classification, we obtained a median balanced accuracy of 0.68 across the five folds for the trials-wise classification, with a MAD of 0.01. The patient-wise classification yielded a median balanced accuracy of 0.82 across the five folds and a MAD of 0.02. Overall, we observed a slight improvement for both the trials and patient-wise classification when using transfer learning. For more detailed performance metrics, please see the results provided below in Figures B.7 and B.8, and in Table B.1.

These results highlight the advantages of transfer learning, particularly when working with relatively small datasets, such as the one used in the present work, confirming what was already highlighted by other studies in the literature [42].

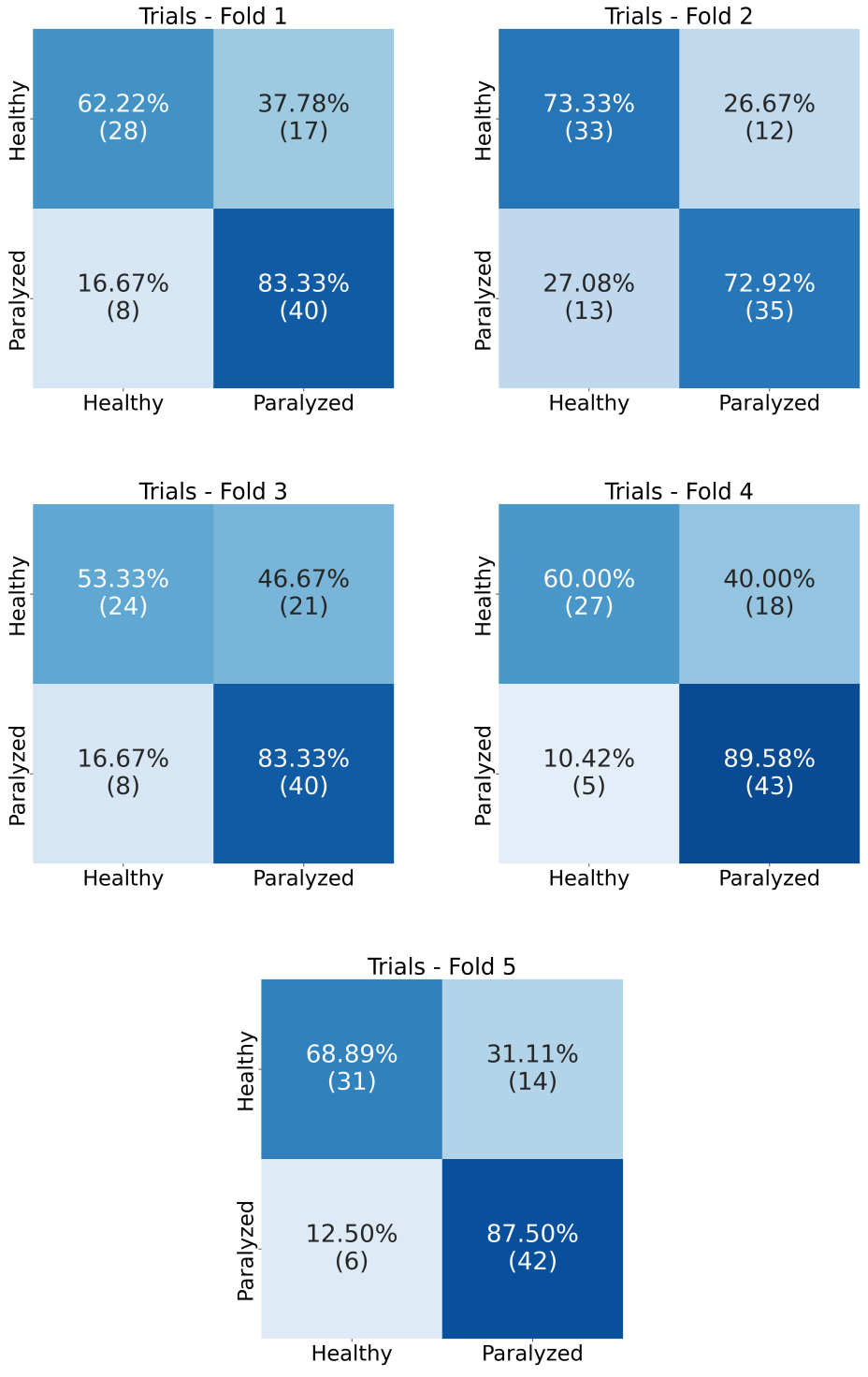

Figure B.1: Confusion matrices for trials (folds 1 to 5) - binary classification.

Appendix B.3 Ablation Study

In the following, we report more detailed information about the results obtained by the binary and multi-class classification models trained during the ablation analysis. As written in the main text, ablation was performed using

Figure B.2: Confusion matrices for patients (folds 1 to 5) - binary classification.

a Shapley value approach. To see the meaning of each player, please refer to the main text. Tables B.2 and B.3 report the balanced accuracy obtained by each coalition for the binary and multi-class classification, respectively.

	Trials (median ± MAD)			Patients (median ± MAD)		
Metric	**Healthy**	**Unilateral**	**Bilateral**	**Healthy**	**Unilateral**	**Bilateral**
Precision	0.82±0.05	0.56±0.01	0.57±0.11	0.85±0.06	0.80±0.02	0.80±0.07
Recall	0.60±0.04	0.64±0.08	0.77±0.11	0.85±0.04	0.62±0.14	0.80±0.09
F1-score	0.67±0.01	0.60±0.04	0.66±0.08	0.85±0.02	0.71±0.06	0.83±0.024
General Performance	**Balanced Accuracy:** 0.68±0.02			**Balanced Accuracy:** 0.82±0.02		

Table B.1
Median and Median Absolute Value (MAD) of the performance obtained by the non-pretrained ResNet multi-class classification model across the 5 folds. Here we report the values of Balanced Accuracy, F1-score, Sensitivity, and Precision obtained in a one-vs-all (micro) approach. This is done both for trials and patients, where the predictions on the several trials related to specific subjects are aggregated.

Coalitions	Trials (median ± MAD)	Patients (median ± MAD)
1	0.69 ± 0.09	0.74 ± 0.07
2	0.76 ± 0.04	0.81 ± 0.07
3	0.77 ± 0.05	0.85 ± 0.06
4	0.76 ± 0.04	0.85 ± 0.09
1, 2	0.73 ± 0.08	0.85 ± 0.08
1, 3	0.76 ± 0.06	0.85 ± 0.03
1, 4	0.77 ± 0.05	0.92 ± 0.05
2, 3	0.64 ± 0.09	0.78 ± 0.11
2, 4	0.73 ± 0.06	0.88 ± 0.07
3, 4	0.73 ± 0.05	0.85 ± 0.08
1, 2, 3	0.70 ± 0.04	0.81 ± 0.06
1, 2, 4	0.64 ± 0.06	0.77 ± 0.06
1, 3, 4	0.76 ± 0.02	0.85 ± 0.03
2, 3, 4	0.73 ± 0.03	0.89 ± 0.03

Table B.2
Balanced accuracy (median ± MAD) for trials and patients across different combinations for binary classification. The participation of each player in each coalition is indicated in the left column through the number of each player. See the Methods for more details.

Appendix C. Code Availability

The complete pipeline for the automatic assessment of laryngeal paralysis from video-laryngoscopy data, including feature extraction from glottic angle time series, supervised learning with deep neural networks and multimodal classifiers, and ablation studies for model interpretability and performance attribution, is available at the following URL: https://github.com/Emaagr/Automatic_Laryngeal_Paralysis_Assessment.

Coalitions	Trials (median ± MAD)	Patients (median ± MAD)
1	0.37 ± 0.08	0.41 ± 0.06
2	0.55 ± 0.03	0.57 ± 0.07
3	0.66 ± 0.04	0.67 ± 0.02
4	0.68 ± 0.08	0.62 ± 0.08
1, 2	0.53 ± 0.03	0.50 ± 0.07
1, 3	0.63 ± 0.06	0.69 ± 0.10
1, 4	0.61 ± 0.04	0.61 ± 0.03
2, 3	0.63 ± 0.04	0.65 ± 0.04
2, 4	0.61 ± 0.05	0.62 ± 0.05
3, 4	0.61 ± 0.07	0.67 ± 0.11
1, 2, 3	0.67 ± 0.06	0.75 ± 0.08
1, 2, 4	0.61 ± 0.05	0.57 ± 0.05
1, 3, 4	0.70 ± 0.05	0.73 ± 0.10
2, 3, 4	0.61 ± 0.06	0.63 ± 0.10

Table B.3
Balanced accuracy (median ± MAD) for trials and patients across different combinations for multi-class classification. The participation of each player in each coalition is indicated in the left column through the number of each player. See the Methods for more details.

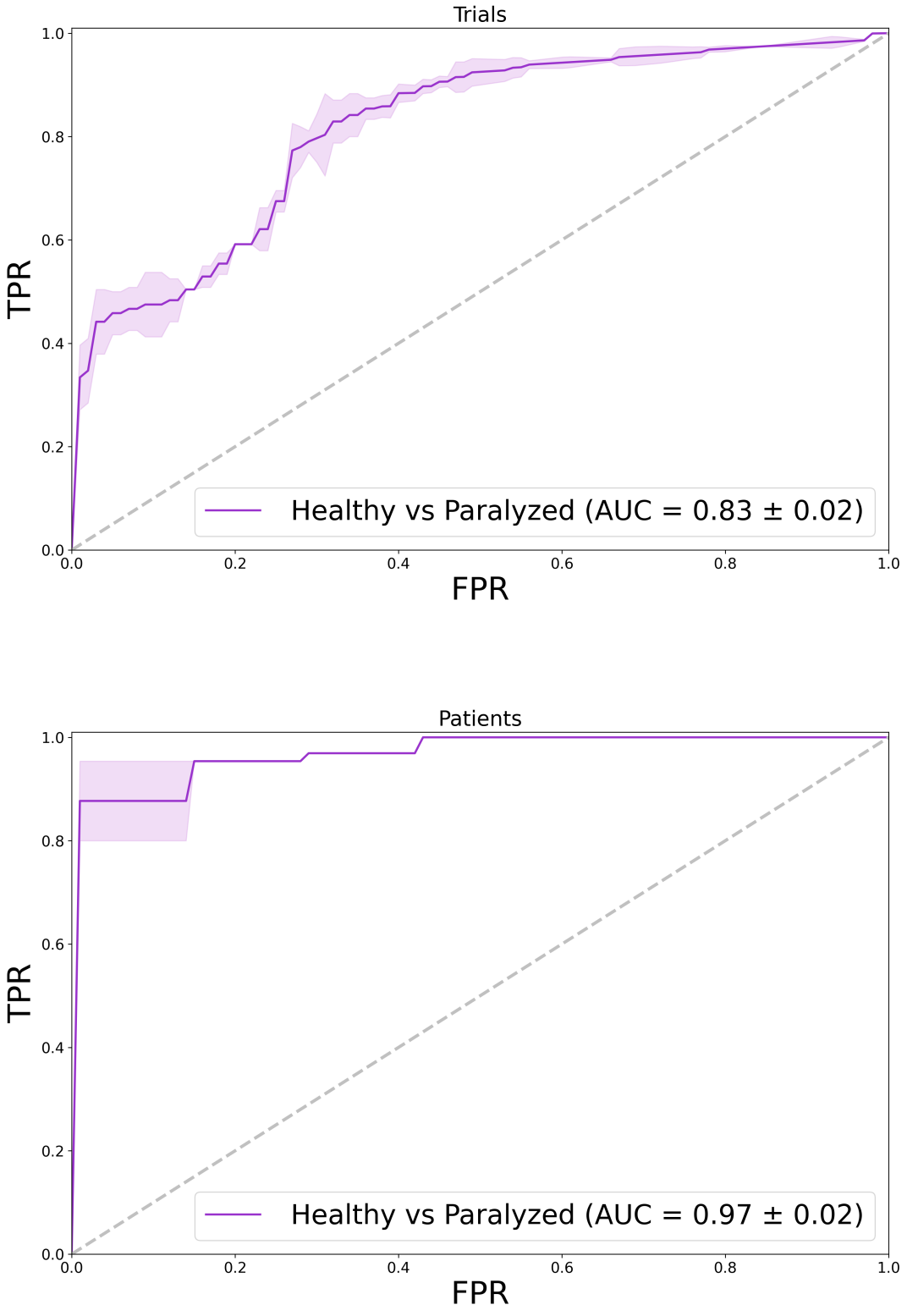

Figure B.3: Comparison of ROC-AUC curves for patients and trials in binary classification. FPR/TPR denote False/True Positive Rate. The shaded area represents the Median Absolute Deviation (MAD) of the AUC scores of the 5 folds.

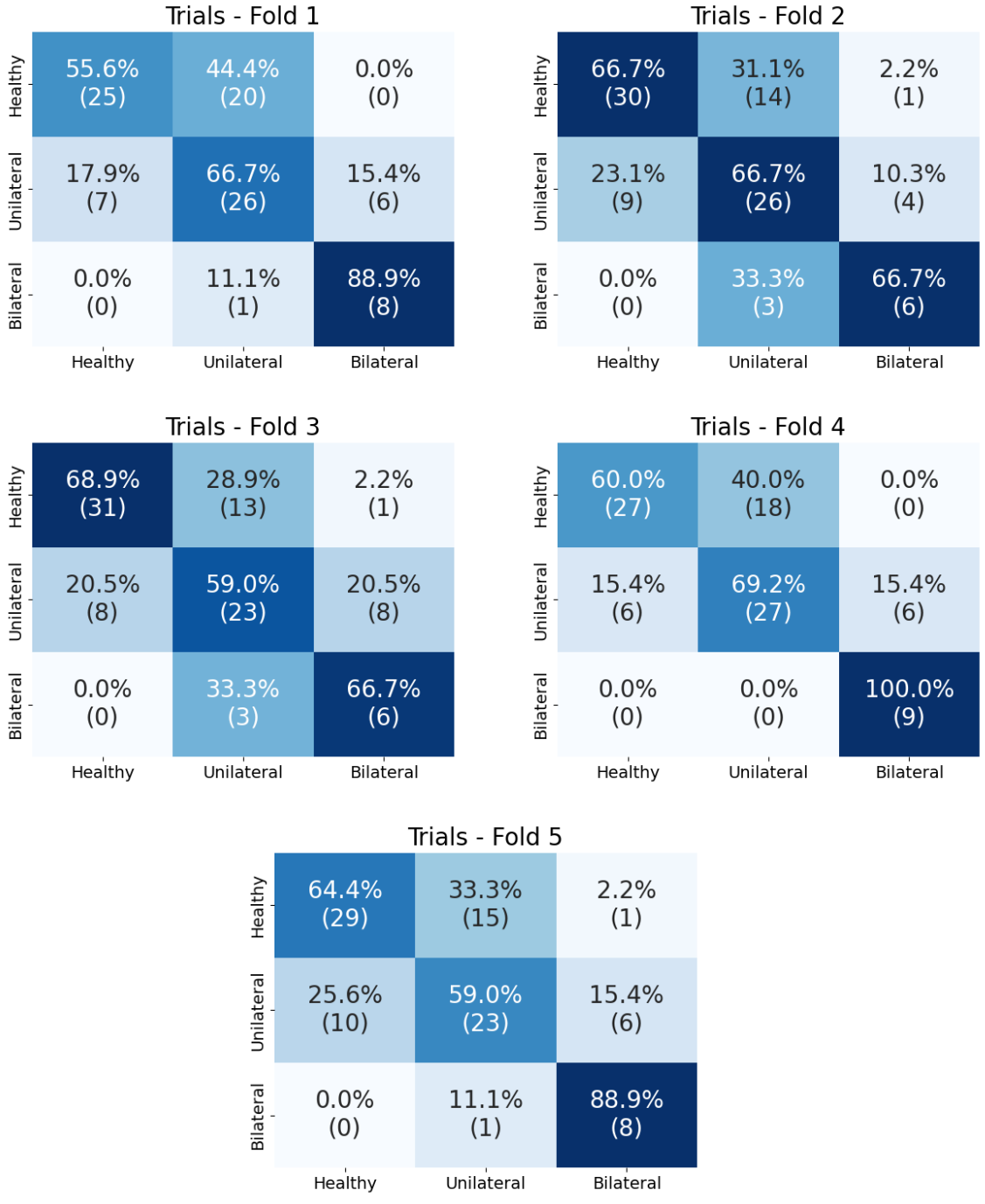

Figure B.4: Confusion matrices for trials with a focus on the 5 different folds for multi-class classification.

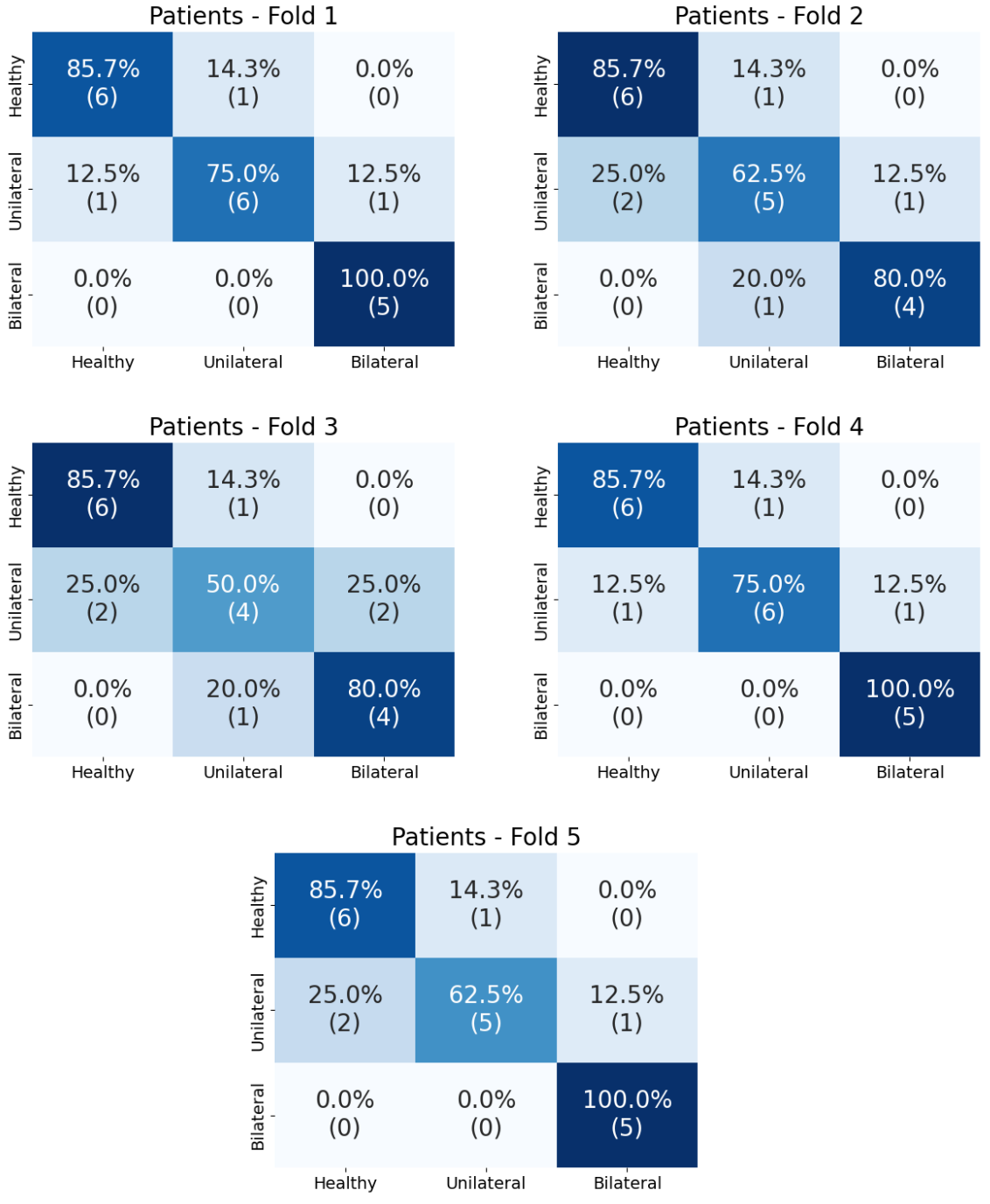

Figure B.5: Confusion matrices for patients with a focus on the 5 different folds for multi-class classification.

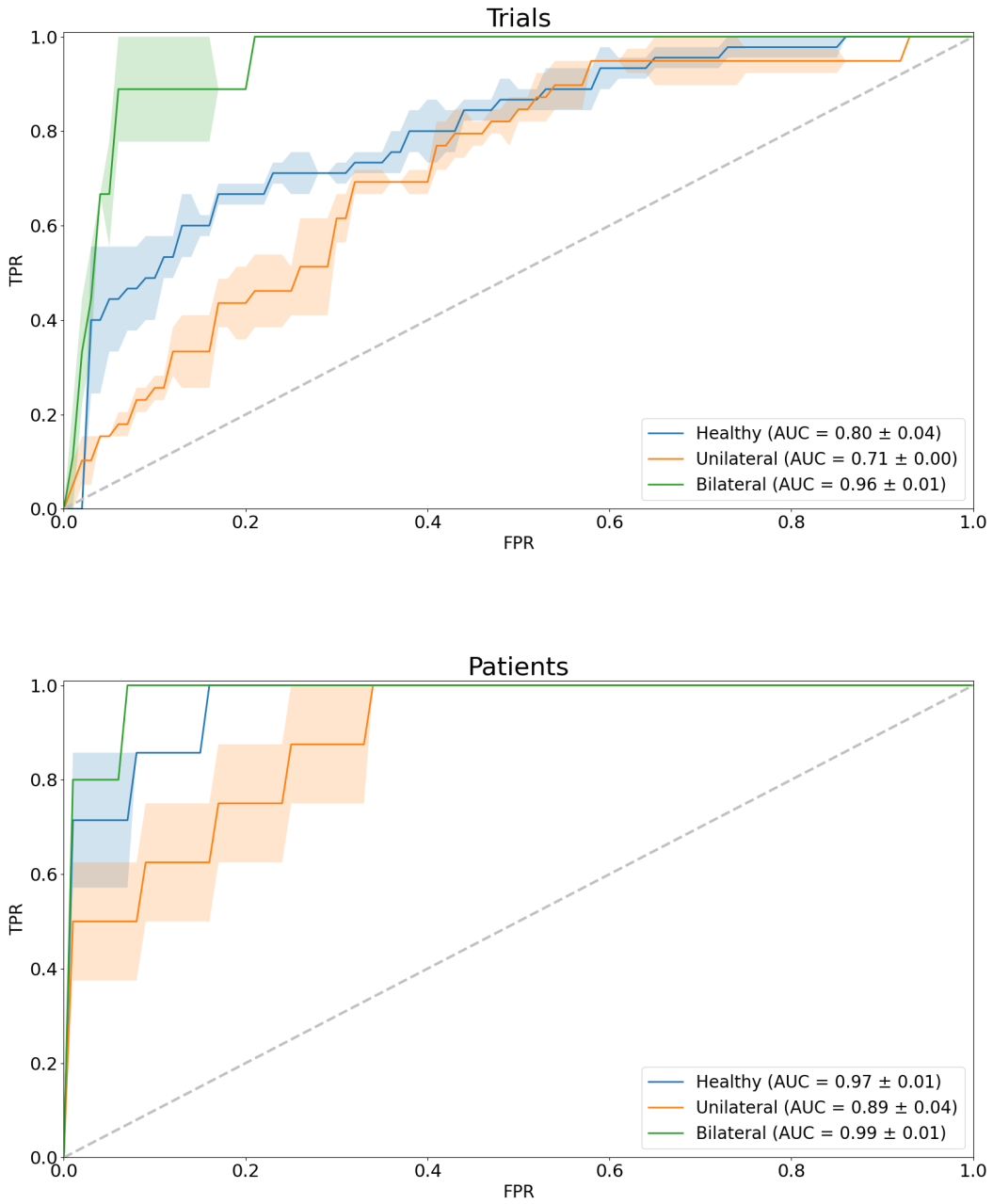

Figure B.6: Comparison of ROC-AUC curves for patients and trials in binary classification. FPR/TPR denote False/True Positive Rate. The shaded area represents the Median Absolute Deviation (MAD) of the AUC scores of the 5 folds.

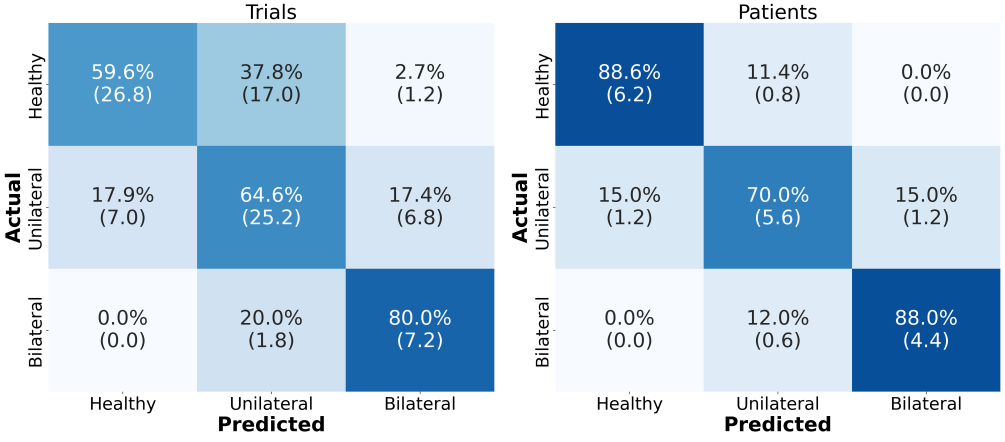

Figure B.7: Median confusion matrices obtained by the non-pretrained ResNet multiclass classification model across folds. Here, each entry shows the percentage of subjects and trials actually categorized in the class characterizing its row and predicted as the class assigned to its column. The median is computed between the different percentages obtained across folds.

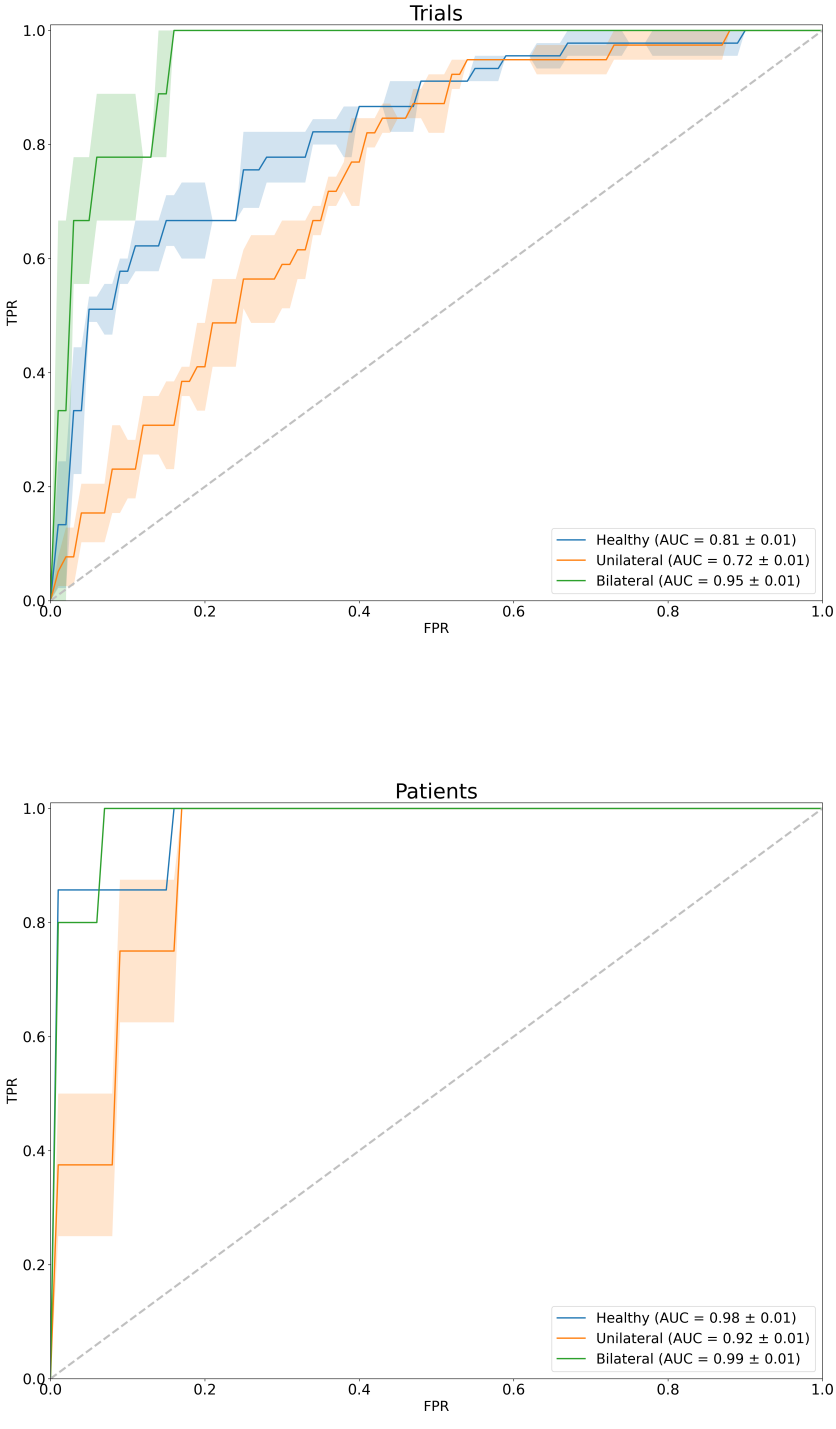

Figure B.8: Comparison of ROC-AUC curves for patients and trials in binary classification for the non-pretrained ResNet. The shaded area represents the Median Absolute Deviation (MAD) of the AUC scores of the 5 folds.

References

[1] C Blake Simpson and Clark A Rosen. *Operative Techniques in Laryngology*. Springer Nature, 2008.

[2] Albert L Merati, Yolanda D Heman-Ackah, Mona Abaza, Kenneth W Altman, Lucian Sulica, and Steven Belamowicz. Common movement disorders affecting the larynx: A report from the neurolaryngology committee of the AAO-HNS. *Otolaryngology – Head and Neck Surgery*, 133(5):654–665, 2005.

[3] Chloe Walton, Paul Carding, Erin Conway, Kieran Flanagan, and Helen Blackshaw. Voice outcome measures for adult patients with unilateral vocal fold paralysis: A systematic review. *The Laryngoscope*, 129(1):187–197, 2019.

[4] Giovanni Succo, Stefano Cirillo, Ilaria Bertotto, Elena Maldi, Davide Balmativola, Massimo Petracchini, Dario Gned, Alessandro Fornari, Gian Marco Motatto, Andrea E Sprio, et al. Arytenoid fixation in laryngeal cancer: Radiological pictures and clinical correlations with respect to conservative treatments. *Cancers*, 11(3):360, 2019.

[5] Daniel Voigt, Michael Döllinger, Anxiong Yang, Ulrich Eysholdt, and Jörg Lohscheller. Automatic diagnosis of vocal fold paresis by employing phonovibrogram features and machine learning methods. *Computer Methods and Programs in Biomedicine*, 99(3):275–288, 2010.

[6] M Ferrari, F Mularoni, S Taboni, E Crosetti, C Pessina, ALC Carobbio, N Montalto, F Marchi, A Vural, A Paderno, et al. How reliable is assessment of true vocal cord-arytenoid unit mobility in patients affected by laryngeal cancer? A multi-institutional study on 366 patients from the ARYFIX collaborative group. *Oral Oncology*, 152:106744, 2024.

[7] Hyun-Bum Kim, Jaemin Song, Seho Park, and Yong Oh Lee. Classification of laryngeal diseases including laryngeal cancer, benign mucosal disease, and vocal cord paralysis by artificial intelligence using voice analysis. *Scientific Reports*, 14(1):9297, 2024.

[8] Nat Adamian, Matthew R Naunheim, and Nate Jowett. An open-source computer vision tool for automated vocal fold tracking from videoendoscopy. *The Laryngoscope*, 131(1):E219–E225, 2021.

[9] Andreas M Kist, Pablo Gómez, Denis Dubrovskiy, Patrick Schlegel, Melda Kunduk, Matthias Echternach, Rita Patel, Marion Semmler, Christopher Bohr, Stephan Dürr, et al. A deep learning enhanced novel software tool for laryngeal dynamics analysis. *Journal of Speech, Language, and Hearing Research*, 64(6):1889–1903, 2021.

[10] A Méndez Zorrilla and B Garcia Zapirain. Vocal folds paralysis study using a pre-processing stage of Gabor filtering and Chan-Vese segmentation. In *Proceedings of the 10th IEEE International Symposium on Signal Processing and Information Technology*, pages 360–365. IEEE, 2010.

[11] Jing Yao, Fusheng Zhou, and Nan Gao. Quantitative assessment of true vocal fold movement by the lateral-approach laryngeal ultrasonography: A pilot study. *Journal of Voice*, 2024, in press.

[12] Aki Koivu, Obinna I Nwosu, Mitsuki Ota, Kristina Simonyan, and Matthew R Naunheim. Feasibility of real-time automated vocal fold motion tracking for in-office laryngoscopy. *The Laryngoscope*, 2025, in press.

[13] Dennis Gabor. Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering*, 93(26):429–441, 1946.

[14] Tony Chan and Luminita Vese. An active contour model without edges. In *Proceedings of the International Conference on Scale-Space Theories in Computer Vision*, pages 141–151. Springer, 1999.

[15] SM Nuruzzaman Nobel, SM Masfequier Rahman Swapno, Md Rajibul Islam, Mejdl Safran, Sultan Alfarhood, and MF Mridha. A machine learning approach for vocal fold segmentation and disorder classification based on ensemble method. *Scientific Reports*, 14(1), 2024. Article no. 14435.

[16] Chung-Feng Jeffrey Kuo, Po-Chun Wang, Yueng-Hsiang Chu, Hsing-Won Wang, and Chun-Yu Lai. Using image processing technology combined with decision tree algorithm in laryngeal video stroboscope automatic identification of common vocal fold diseases. *Computer Methods and Programs in Biomedicine*, 112(1):228–236, 2013.

[17] Tiffany V Wang, Nat Adamian, Phillip C Song, Ramon A Franco, Molly N Huston, Nate Jowett, and Matthew R Naunheim. Application of a computer vision tool for automated glottic tracking to vocal fold paralysis patients. *Otolaryngology – Head and Neck Surgery*, 165(4):556–562, 2021.

[18] Amaia Méndez, Begoña García, J Vicente, Ibon Ruiz, and Karen Sanchez. Objective model of vocal folds, based on glottal closure, opening angles and morphologic criteria. In *Proceedings of the 2007 9th International Symposium on Signal Processing and Its Applications*, pages 1–4. IEEE, 2007.

[19] Francesca Pia Villani, Maria Chiara Fiorentino, Lorenzo Federici, Cesare Piazza, Emanuele Frontoni, Alberto Paderno, and Sara Moccia. A deep-learning approach for vocal fold pose estimation in videoendoscopy. *Journal of Imaging Informatics in Medicine*, 2025, in press.

[20] Zihan Nie, Muhao Xu, Zhiyong Wang, Xiaoqi Lu, and Weiye Song. A review of application of deep learning in endoscopic image processing. *Journal of Imaging*, 10(11), 2024. Article no. 275.

[21] Alejandro R Marrero-Gonzalez, Tanner J Diemer, Shaun A Nguyen, Terence JM Camilon, Kirsten Meenan, and Ashli O'Rourke. Application of artificial intelligence in laryngeal lesions: A systematic review and meta-analysis. *European Archives of Oto-Rhino-Laryngology*, 282(3):1543–1555, 2025.

[22] Chiara Baldini, Kaisar Kushibar, Richard Osuala, Simone Balocco, Oliver Diaz, Karim Lekadir, and Leonardo S Mattos. Clinically-guided data synthesis for laryngeal lesion detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 54–63. Springer, 2025.

[23] Francesca Pia Villani, Alberto Paderno, Maria Chiara Fiorentino, Alessandro Casella, Cesare Piazza, and Sara Moccia. Classifying vocal folds fixation from endoscopic videos with machine learning. In *Proceedings of the 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4. IEEE, 2023.

[24] Yucong Zhang, Xin Zou, Jinshan Yang, Wenjun Chen, Juan Liu, Faya Liang, et al. Multimodal laryngoscopic video analysis for assisted diagnosis of vocal fold paralysis. *Computer Speech & Language*, 96, 2025. Article no. 101891.

[25] Peter Yao, Moon Usman, Yu H Chen, Alexander German, Katerina Andreadis, Keith Mages, and Anaïs Rameau. Applications of artificial intelligence to office laryngoscopy: A scoping review. *The Laryngoscope*, 132(10):1993–2016, 2022.

[26] Clark A Rosen, Ted Mau, Marc Remacle, Markus Hess, Hans E Eckel, VyVy N Young, Anastasios Hantzakos, Katherine C Yung, and Frederik G Dikkers. Nomenclature proposal to describe vocal fold motion impairment. *European Archives of Oto-Rhino-Laryngology*, 273:1995–1999, 2016.

[27] Douglas A Reynolds. Gaussian mixture models. In Stan Z Li and Anil K Jain, editors, *Encyclopedia of Biometrics*, pages 659–663. Berlin, Springer, 2009.

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

[29] Richard Meyes, Melanie Lu, Constantin Waubert de Puiseau, and Tobias Meisen. Ablation studies in artificial neural networks. *arXiv preprint arXiv:1901.08644*, 2019.

[30] Mukund Sundararajan and Amir Najmi. The many Shapley values for model explanation. In *Proceedings of the International Conference on Machine Learning*, pages 9269–9278. PMLR, 2020.

[31] Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21(9):1281–1289, 2018.

[32] Peter Yao, Dan Witte, Hortense Gimonet, Alexander German, Katerina Andreadis, Michael Cheng, Lucian Sulica, Olivier Elemento, Josue Barnes, and Anaïs Rameau. Automatic classification of informative laryngoscopic images using deep learning. *Laryngoscope Investigative Otolaryngology*, 7(2):460–466, 2022.

[33] Chiara Baldini, Muhammad Adeel Azam, Claudio Sampieri, Alessandro Ioppi, Laura Ruiz-Sevilla, Isabel Vilaseca, Berta Alegre, Alessandro Tirrito, Alessia Pennacchi, Giorgio Peretti, et al. An automated approach for real-time informative frames classification in laryngeal endoscopy using deep learning. *European Archives of Oto-Rhino-Laryngology*, pages 1–10, 2024.

[34] Vahid Majidnezhad and Igor Kheidorov. A hybrid of genetic algorithm and Gaussian mixture model for features reduction and detection of vocal fold pathology. *Journal of Advances in Computer Research*, 4(2):53–62, 2013.

[35] Ji Yeoun Lee. A two-stage approach using Gaussian mixture models and higher-order statistics for a classification of normal and pathological voices. *EURASIP Journal on Advances in Signal Processing*, 2012:1–8, 2012.

[36] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

[37] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[38] Timen C Ten Harkel, Guido de Jong, Henri AM Marres, Koen JAO Ingels, Caroline M Speksnijder, and Thomas JJ Maal. Automatic grading of patients with a unilateral facial paralysis based on the Sunnybrook facial grading system – a deep learning study based on a convolutional neural network. *American Journal of Otolaryngology*, 44(3):103810, 2023.

[39] Xiangyu Peng, Huoyao Xu, Jie Liu, Junlang Wang, and Chaoming He. Voice disorder classification using convolutional neural network based on deep transfer learning. *Scientific Reports*, 13(1):7264, 2023.

[40] Jianjun Ren, Xueping Jing, Jing Wang, Xue Ren, Yang Xu, Qiuyun Yang, Lanzhi Ma, Yi Sun, Wei Xu, Ning Yang, et al. Automatic recognition of laryngoscopic images using a deep-learning technique. *The Laryngoscope*, 130(11):E686–E693, 2020.

[41] E Agrimi, A Diko, D Carlotti, A Ciardiello, M Borthakur, S Giagu, S Melchionna, and C Voena. Covid-19 therapy optimization by AI-driven biomechanical simulations. *The European Physical Journal Plus*, 138(2):182, 2023.

[42] Zehui Zhao, Laith Alzubaidi, Jinglan Zhang, Ye Duan, Usman Naseem, and Yuantong Gu. Robust and explainable framework to address data scarcity in diagnostic imaging. *Computers in Biology and Medicine*, 197, 2025. Article no. 111052.

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[44] Deepshikha Bhati, Fnu Neha, and Md Amiruzzaman. A survey on explainable artificial intelligence (xai) techniques for visualizing deep learning models in medical imaging. *Journal of Imaging*, 10(10), 2024. Article no. 239.

[45] Tatsuro Ichiishi. *Game Theory for Economic Analysis*. Elsevier, 2014.

[46] Claudio Sampieri, Chiara Baldini, Muhammad Adeel Azam, Sara Moccia, Leonardo S Mattos, Isabel Vilaseca, Giorgio Peretti, and Alessandro Ioppi. Artificial intelligence for upper aerodigestive tract endoscopy and laryngoscopy: A guide for physicians and state-of-the-art review. *Otolaryngology–Head and Neck Surgery*, 169(4):811–829, 2023.

[47] Sunali Vij, Ashok K Gupta, and Dharam Vir. Voice quality following unilateral vocal fold paralysis: A randomized comparison of therapeutic modalities. *Journal of Voice*, 31(6):774–e9, 2017.

[48] Elliana Kirsh DeVore, Nat Adamian, Nate Jowett, Tiffany Wang, Phillip Song, Ramon Franco, and Matthew Roberts Naunheim. Predictive outcomes of deep learning measurement of the anterior glottic angle in bilateral vocal fold immobility. *The Laryngoscope*, 133(9):2285–2291, 2023.

[49] Ibrahim AlShourbaji, Pramod Kachare, Waleed Zogaan, Lawan Jibril Muhammad, and Laith Abualigah. Learning features using an optimized artificial neural network for breast cancer diagnosis. *SN Computer Science*, 3(3), 2022. Article no. 229.

[50] Lawan Jibril Muhammad and Alessandro Bria. Editorial: AI applications for diagnosis of breast cancer. *Frontiers in Artificial Intelligence*, 6, 2023. Article no. 1247261.

[51] Anisie Uwimana, Giorgio Gnecco, and Massimo Riccaboni. Artificial intelligence for breast cancer detection and its health technology assessment: A scoping review. *Computers in Biology and Medicine*, 184, 2025. Article no. 109391.

[52] Celine MLH Wilmes, Arsen Goril BSc, Henri AM Marres, David J Wellenstein, and Guido B van den Broek. A systematic review of the clinical impact of implementing artificial intelligence in upper aerodigestive tract endoscopy. *Head & Neck*, 2025.

[53] Yael Bensoussan, Erik B Vanstrum, Michael M Johns III, and Anaïs Rameau. Artificial intelligence and laryngeal cancer: From screening to prognosis: A state of the art review. *Otolaryngology – Head and Neck Surgery*, 168(3):319–329, 2023.

[54] Claudio Sampieri, Chiara Baldini, Muhammad Adeel Azam, Sara Moccia, Leonardo S Mattos, Isabel Vilaseca, Giorgio Peretti, and Alessandro Ioppi. Artificial intelligence for upper aerodigestive tract endoscopy and laryngoscopy: A guide for physicians and state-of-the-art review.

Otolaryngology – Head and Neck Surgery, 169(4):811–829, 2023.

[55] James D Brierley, Mary K Gospodarowicz, and Christian Wittekind. *TNM Classification of Malignant Tumours*. John Wiley & Sons, 2017.

[56] P Gorphe, P Blanchard, I Breuskin, S Temam, Y Tao, and F Janot. Vocal fold mobility as the main prognostic factor of treatment outcomes and survival in stage ii squamous cell carcinomas of the glottic larynx. *The Journal of Laryngology & Otology*, 129(9):903–909, 2015.