



Published in final edited form as:

Laryngoscope. 2025 July ; 135(7): 2428–2436. doi:10.1002/lary.32036.

A Deep-Learning Model for Multi-class Audio Classification of Vocal Fold Pathologies in Office Stroboscopy

Yeo Eun Kim, BA^{*1}, Maria Dobko, BS^{*1}, Haomiao Li, MS^{*2}, Tianlan Shao, BS¹, Preethi Periyakoil, BS¹, Courtney Tipton, MD¹, Christine Colasacco, BS³, Aisha Serpedin, BA⁴, Olivier Elemento, PhD⁵, Mert Sabuncu, PhD⁶, Michael Pitman, MD³, Lucian Sulica, MD¹, Anaïs Rameau, MD MS MPhil FACS¹

¹Sean Parker Institute for the Voice, Department of Otolaryngology-Head and Neck Surgery, Weill Cornell Medicine, New York, New York, U.S.A.

²Cornell Tech, New York, NY, USA

³The Center for Voice and Swallowing, Department of Otolaryngology-Head and Neck Surgery, New York-Presbyterian Hospital, Columbia University Irvine Medical Center, New York City, New York, U.S.A.

⁴Weill Cornell Medicine-Qatar, Doha, Qatar.

⁵Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, New York, U.S.A.

⁶Department of Radiology, Weill Cornell Medicine, New York, New York, U.S.A.

Abstract

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

Methods: Patients with UVFP (n=105), VF lesions (n=63), and HVF (n=41) were retrospectively identified. Voice samples were extracted from stroboscopic videos (Pentax Laryngeal Strobe Model 9400), including sustained /i/ phonation, pitch glide, and /i/ sniff task. Extracted audio files were converted into Mel-spectrograms. Voice samples were independently divided into training (80%), validation (10%), and test (10%) by patient. Pre-trained ResNet18 models were trained to classify 1) HVF and pathological VF (lesions and UVFP), and 2) HVF, UVFP, and VF lesions. Both classifiers were further validated on an external dataset consisting of 12 UVFP, 13 VF lesions, and 15 HVF patients. Model performances were evaluated by accuracy and F1-score.

Results: When evaluated on a hold-out test set, the binary classifier demonstrated stronger performance compared to the multi-class classifier (accuracy 83% vs 40%; F1-score 0.90 vs 0.36). When evaluated on an external dataset, the binary classifier achieved an accuracy of 63% and F1-score of 0.48, compared to 35% and 0.25 for the multi-class classifier.

Corresponding Author: Anaïs Rameau, MD, MPhil, Assistant Professor of Otolaryngology, Department of Otolaryngology – Head and Neck Surgery, Sean Parker Institute for the Voice, Weill Cornell Medicine, 240 East 59th St, New York, NY 10022, anr2783@med.cornell.edu, Fax: (646) 962-0393.

*Co-first author.

Conclusions: Deep-learning classifiers differentiating HVE, UVFP, and VF lesions were developed using voice data from stroboscopic videos. Although healthy and pathological voice were differentiated with moderate accuracy, multi-class classification lowered model performance. The model performed poorly on an external dataset. Voice captured in stroboscopic videos may have limited diagnostic value, though further studies are needed.

Level of Evidence: 4

Graphical Abstract

Figure: 1

Caption: A deep-learning model was developed and validated for distinguishing healthy vocal folds (HVF), unilateral vocal fold paralysis (UVFP), and vocal fold lesions from voice data derived from stroboscopy videos. When tested on an external dataset, the model achieved lower performance. Overall, the utility for voice data captured with stroboscopic equipment in augmenting disease classification seems to be equivocal at best.

Keywords

Deep learning; artificial intelligence; convolutional neural network; computer-aided diagnosis; laryngology

Introduction

Videolaryngostroboscopy (VLS) is the mainstay of objective laryngeal evaluation in patients with dysphonia and is currently almost exclusively performed in otolaryngology – head and neck surgery practices. Unlike traditional office laryngoscopy with a constant light source, VLS employs a stroboscopic light that flashes faster than the patient's fundamental frequency. This allows physicians to capture detailed images of vocal folds (VF) at various phases of the glottal cycle, enabling slow-motion visualization of VF oscillations¹. Additionally, stroboscopic videos provide a unique opportunity to integrate high-quality dynamic imaging with synchronized voice recordings.

Compared to state-of-the-art assessment by otolaryngologists with access to VLS, voice evaluation in the primary care setting remains rudimentary, largely relying on patient-reported symptoms and auditory-perceptual assessment by untrained clinicians. This unfortunately often leads to delays in diagnosis and management of laryngeal disorders². Even when patients are referred to specialists, otolaryngologists often face difficulties triaging referrals based on primary care clinician assessment³. Furthermore, access to VLS varies significantly by region⁴, with relative scarcity in the contiguous Mountain, West South Central and East South Central divisions of the United States, where referral for dysphonia evaluation is particularly challenging.

Applications of deep learning (DL), in upper aerodigestive tract endoscopy, dubbed “videomics” by Paderno et al.⁵, have gained interest in the past decade⁶, partly motivated by the desire to improve access to diagnostic exams in surgical deserts. DL techniques using convolutional neural networks (CNN) have been applied for analysis of laryngoscopic

images, such as the classification of various VF pathologies, with performance comparable to that of human experts^{2,7,8}. The overall vision for implementing these algorithms in low-resource settings is to provide clinical decision support (CDS) during endoscopy, potentially to non-experts who could screen for pathologies requiring urgent evaluation by specialists⁹.

DL applications in healthcare have expanded into voice, speech, and respiratory sounds, with rapid expansion during the COVID-19 pandemic. This growing field of data science focused on identification of acoustic biomarkers of health and disease has been recently named “audiomics.”¹⁰ Combined with video data, VLS audio recordings facilitate the integration of videomics and audiomics through multimodal machine learning. Currently, research on DL capabilities for voice-based diagnosis of VF pathologies is inconclusive, primarily due to a lack of large, clinically validated voice datasets that include sociodemographic information. Further, voice prints are considered biometric identifiers under the Health Insurance Portability and Accountability Act (HIPAA), complicating data sharing¹¹.

While CNN-based voice classifiers differentiate between binary states (i.e. healthy and pathological voices) with sensitivity^{3,12–14,15}, their performance declines when classifying multiple vocal states¹⁶ – suggesting their potential as screening tools and not as diagnostic tools. Their performance in pathology discrimination based on voice is akin to that of auditory-perceptual judgement by experts. In a Taiwanese study, expert laryngologists with more than 10 years of experience demonstrated limited accuracy in distinguishing different VF pathologies based on voice samples of sustained vowel sound /a:/ along with the recording of a Mandarin passage¹⁶, with a maximum overall accuracy of 60%, though superior to that of general otolaryngologists, with a maximum overall accuracy of 40%.

Here, we aimed to investigate the screening and diagnostic capabilities of CNNs trained on voice recordings from VLS. Acoustic classifiers were developed to 1) differentiate voices produced by patients with healthy vocal folds (HVF) and pathological VFs; and 2) discriminate voice produced by patients with HVF, unilateral vocal fold paralysis (UVFP) and vocal folds with lesion (VF lesion). These classifiers were also tested on an external dataset to evaluate generalizability (Figure 1). We hypothesized that these classifiers would 1) distinguish normal from pathological voice with good accuracy, but not perform as well in multi-class discrimination and 2) perform with reduced accuracy on the external dataset.

We chose to focus on HVF, UVFP and VF lesions, as they are clinically associated with significantly different voice outputs. UVFP results in the complete immobility of a vocal fold, imparting a breathy quality to the voice due to inadequate closure of the VFs during phonation¹⁷. VF lesions often limit proper vocal fold closure, causing breathiness, and lead to phase asymmetry, resulting in a grainier voice¹⁸. To our knowledge, this is the first study evaluating the screening and diagnostic capabilities of voice-based CNN models based on acoustic samples extracted from VLS recordings.

Methods

Data

This project was approved by the Weill Cornell Medicine and the Columbia University Irvine Medical Center Institutional Review Board. A retrospective dataset of VLS recordings was collected at Sean Parker Institute for Voice at Weill Cornell Medicine. All adult patients who were seen by our department with HVE, UVFP, and VF lesions with available recordings were included. Those with multiple diagnoses (e.g. both UVFP and VF lesion) and those with inaudible audio recordings were excluded. A total of 209 videos were included in the dataset, with 41 HVE, 105 UVFP, and 63 VF lesions (33 unilateral VF polyps, 16 cysts, 14 squamous cell carcinoma (SCC)). Diagnoses were made by the senior laryngologist in the team with over 15 years of experience. A second retrospective dataset of 40 VLS recordings was collected at Columbia University Irvine Medical Center. This dataset included 15 videos of HVE, 12 with UVFP, and 13 with VF lesions (7 unilateral VF polyps, 1 cyst, 5 SCC).

Of note, the diagnosis of SCC was confirmed on biopsy at both institutions. There was no patient overlap between the HVE, UVFP, and VF lesion cohorts. Each patient only contributed one stroboscopy exam that was performed prior to any intervention.

Audio Data Collection

All voice samples were recorded with the Kay-Pentax Stroboscopy System (Laryngeal Strobe: Model 9400; New Jersey, USA) during the stroboscopic examination in standard clinical rooms without sound treatment. A lapel microphone, placed on the patient's collar, and a laryngeal microphone, held at the neck (Model 7175–6000, Model MU0P00000718, Pentax Medical; New Jersey, USA), were used at Weill Cornell Medicine, without controlling for mouth-to-microphone distance. A microphone attached to the camera head (Audio-Technica; Tokyo, Japan) was used at Columbia University Irvine Medical Center. Recordings were performed at a sampling rate of 44kHz with a 16-bit resolution and were embedded in video files (MP4). Audio data was extracted in the wav format using the Adapter tool (Macroplant; Massachusetts, USA). Background noise included that produced by the Pentax Endoscopy System equipment, especially by the Laryngeal Strobe flickering light mechanism, at both sites.

Audio File Extraction

Specific audio segments were extracted from the audio recordings by a trained data extractor and researcher (YK). Three voice tasks, sustained /i/, pitch glides, and alternating /i/ sniff, elicited during VLS were isolated from each patient's recording. These tasks were chosen because they were performed consistently through all stroboscopic examinations. In particular, alternating /i/ sniff was included as vocal fold movement changes may be observed on stroboscopy in neurogenic dysphonia¹⁹ leading us to think that this task may carry acoustic diagnostic information. Only segments containing the patient's voice were extracted. Recordings that were inaudible or contained the examiner's voice were excluded. Ten percent of cases from each class was verified by two expert laryngologists (AR, CT) for correct extraction technique.

Three datasets were created: 1) longest sustained /i/ from each patient (n=209), 2) longest sustained /i/, alternating /i/ sniff with the most alterations, and pitch glides covering the most range from each patient (n=423), and 3) all voice tasks from all patients (n=681). In Dataset 2, the longest voice segment was extracted for each task in case of multiple performance of a task. If a patient did not perform all voice tasks besides sustained /i/, only the available voice tasks were included in the dataset.

Data Pre-processing

Mel-spectrograms are useful for visualizing the amplitude of different frequency bands over time and are often used in voice-based classifiers¹². Our audio samples were converted to Mel-spectrograms using the Librosa package,²⁰ and resampled to a sampling rate of 48kHz to ensure high fidelity. We used a Fast Fourier Transform window size of 2048 and a hop length of 256, balancing time resolution and computational efficiency.

To enrich the dataset and enhance the model's robustness, pitch augmentation was applied randomly to half of the training dataset. This technique modifies the pitch of the audio samples without altering the playback speed. Mel-spectrograms are the representation of the raw data and convolutional neural network (CNN) learn the features automatically. CNN retrieve different features based on patterns in data through a learning process on labelled data. It is the nature of the neural network that mitigates the need for manual feature selection. Though CNNs can handle raw Mel-spectrograms without feature selection, they still benefit from data processing and careful architecture design, as described below. We also used data augmentation to improve the model's generalizability. When a CNN is trained on Mel-spectrograms, it typically takes the entire spectrogram as input, i.e. all the coefficients that represent the energy distribution across different frequency bands on the Mel scale over time.

Datasets 1, 2, and 3 were independently divided into training (80%), validation (10%), and test (10%), using random sampling stratified by class (HVF, VF lesions, UVFP). Stratifying by class ensures a balanced representation of each category in both training and validation. Further, splitting by patients prevents data leakage between the training and validation sets, thereby preserving the integrity of our model's evaluation and preventing overestimation of its performance. We do acknowledge the small size of the external dataset, but the inclusion of even a small external validation dataset strengthens our analysis compared to most other clinical ML studies that only rely on internal test set.

For explainable ML and feature importance analyses, the audio samples were processed using OpenSmile with the eGeMAPS configuration file²¹. This automatically extracts 88 acoustic features per sample, providing information on^{22,23}:

- *Vocal folds*, including pitch (logarithmic F0) and jitter (period-to-period deviations)
- *Intensity*, including loudness (perceived signal intensity), harmonics-to-noise ratio (HNR), and shimmer (variations in peak amplitudes)
- *Vocal tract*, including Formant, or F1–3 frequency, bandwidth, and amplitude

- *Spectral balance*, including alpha ratio, Hammarberg index, spectral slope, and Mel-frequency cepstral coefficient, or MFCC 1–4

Audio Classifier

Binary and multi-class classification tasks were performed on each of the three datasets described above. Binary classification was performed to distinguish between healthy and pathological voice, and multi-class classification was performed to distinguish between HVE, VF lesions, and UVFP. The same data processing techniques were used across the different models. Models were trained using the Weill Cornell dataset and externally validated on the Columbia dataset.

A pre-trained ResNet18 was chosen to preserve the CNN architecture while limiting the depth of the network²⁴. Although more commonly used for image-based tasks, ResNet models have been used to learn and identify patterns in audio data that have been pre-processed into spectrogram images²⁵. ResNet-18 is a deep convolutional neural network with 18 layers designed to improve learning by incorporating residual connections, originally trained on the ImageNet dataset²⁶. Transfer learning leverages this pre-trained model by fine-tuning it on a new task, enabling efficient learning with limited data by utilizing the previously learned features.

AdamW was used as the model optimizer, with a starting learning rate of $3e-4$ and linear scheduling with step size 2. AdamW decouples weight decay from optimization steps, aiming to improve training performance and generalization²⁷. Weighted Cross Entropy loss function was applied, as it has been shown to mitigate the effects of class imbalance and improve the accuracy of binary classifiers, for instance in detecting diabetic retinopathy²⁸. Training was stopped after 15 epochs without validation loss improvement, using a batch size of 16.

PyTorch Lightning was employed for model development and training processes²⁹. The code is available at https://github.com/marichka-dobko/vocal_fold_pathology_classifier.

Explainable Machine Learning Analysis and Assessing Feature Importance

To assess how the ResNet image-based neural network performs compared to other machine learning models, and to assess the effects of acoustic features on classification performance, three ML algorithms of increasing complexity (i.e. harder to interpret) were used. These included Logistic Regression (LR; most simple linear model), Random Forest (RF; more complex decision tree-based model), and Multi-Layer Perceptron (MLP; most “black box” neural network)²².

To generate independent test and train data splits, a stratified shuffle split sampling scheme was used. This method ensures that the training and validation sets maintain the same class distribution as the original dataset, which is especially important when dealing with imbalanced data. Compared to other cross-validation methods, such as k-fold cross validation, stratified shuffle split ensures consistent representation of minority classes and can contribute to improved model performance³⁰. Additionally, stratified shuffle split introduces randomness by generating new splits in each iteration, enabling multiple rounds

of evaluation for robust and reliable performance estimation. For each iteration, a random selection of 20% of the dataset was used to create a held-out test set, and the remaining 80% was used for training. This process was repeated 50 times, and the three classifiers were fitted and tested for each test/train split. The best models with highest Area Under the Receiver Operating Characteristic Curve (AUROC) on validation set were saved to test on the external dataset separately.

Kernel SHAP (SHapley Additive exPlanations) was used to determine the top acoustic features for each model, as it is capable of computing feature importance on any trained model regardless of complexity³¹. The absolute SHAP values were computed on the validation set for each of the 50 splits and the average was recorded across all classes. These per-split averages were then stacked and averaged again across the 50 splits to obtain an overall mean SHAP value per feature. We further multiplied each feature's mean SHAP value by the model's median AUROC (as a weight). Finally, the top 5 features with the highest weighted mean SHAP values were reported for each model.

Statistical Analysis

The performance of DL neural network was evaluated by the following metrics: accuracy, precision, recall (sensitivity; referring to pathology), and F1-score, calculated using Python. F1-score is a harmonic mean of precision (positive predictive value) and recall (sensitivity). This metric is useful in datasets with class imbalance, as it combines how well the model identifies true positives (recall) and avoids false positives (precision). The mean AUROC, along with accuracy, precision, recall, and F1-score, was computed for the three ML models to evaluate their average performance on across 50 bootstrapping iterations. Demographics and audio sample duration were compared using Student's t-test and ANOVA for quantitative variables and Fischer's exact test for qualitative variables. Statistical analysis was performed using R Version 3.6.2 (R Foundation for Statistical Computing, Vienna, Austria).

Results

The mean age of patients included in the Weill Cornell dataset was 57±19 years, with 122 (58%) females. The mean age of patients included in the Columbia dataset was 51±21 years, with 19 (48%) females. There were no significant differences in age ($p=0.0511$) or sex ($p=0.2255$) between the two cohorts. There were no significant differences in the duration of audio files between the HVF, VF lesion, and UVFP cohorts in the Weill Cornell dataset ($p=0.4120$).

Performance of Deep Learning Neural Networks

When evaluated on a hold-out test set, the binary audio classifier trained on Dataset 1 achieved an accuracy of 83% and F1-score of 0.90 (Table 1). The binary classifier trained on Dataset 2 achieved an accuracy of 75% and F1-score of 0.69, whereas the classifier trained on Dataset 3 achieved an accuracy of 75% and F1-score of 0.72 (Table 1). In comparison, the multi-class audio classifier trained on Dataset 1 achieved an accuracy of 40% and F1-score of 0.36 (Table 1). The multi-class classifier trained on Dataset 2 achieved an

accuracy of 45% and F1-score of 0.43, whereas the classifier trained on Dataset 3 achieved an accuracy of 40% and F1-score of 0.39 (Table 1).

When evaluated on the external dataset, the binary audio classifier trained on Dataset 1 achieved an accuracy of 63% and F1-score of 0.48, whereas the multi-class audio classifier trained on Dataset 1 achieved an accuracy of 35% and F1-score of 0.25 (Table 2). The binary classifiers trained on Datasets 2 and 3 both achieved an accuracy of 60%, whereas the multi-class classifiers trained on Datasets 2 and 3 achieved accuracies of 43% and 45%, respectively (Table 2). Other performance metrics of both binary and multi-class classifiers are detailed in Tables 1–2.

Performance of Machine Learning Models and Assessing Feature Importance

The top 5 features for Logistic Regression and Random Forest included Formant 1–3 parameters and MFCC; top features for MLP included temporal (unvoiced segment length) and spectral (Hammarberg Index) parameters in addition to the aforementioned features (Supplemental A).

Of the three ML models, MLP achieved the highest performance (accuracy=89%, F1-score=0.89, AUROC=0.89) (Table 3). However, the performance of all three models globally degraded on the multi-class classification task, with the Random Forest demonstrating highest performance metrics (accuracy=65%, F1-score=0.63, AUROC=0.79) (Table 3). Similar trends were noticed on the external dataset. All three models underperformed across every performance metric for both binary and multi-class classification tasks, with multi-class classifiers demonstrating globally inferior performance (Table 4).

Discussion

Here, we investigated the screening and diagnostic capabilities of DL-based classifiers in differentiating between voices produced with HVF and pathological VFs using acoustic data collected from VLS recordings. The binary classifier trained on single-vowel phonation distinguished normal from pathological voice with 80% accuracy. However, the accuracy dropped to 40% when distinguishing voices produced with HVF and voices produced with VF lesions or UVFP. In other words, while there is moderate success in binary classification of healthy versus pathological voices, the performance significantly diminishes in the multi-class classification of HVF, VF lesions, and UVFP. This aligns with the results described by Hu et al., where the model accuracy was 95% for the binary distinction of healthy from pathological voices, but dropped to 67% when discriminating between five classes, including VF lesions and UVFP¹⁶. The performance of our DL classifiers appears similar to human expert aptitude in discriminating different etiologies of dysphonia¹⁶. Consistent with prior findings^{16,32,33}, the present study suggests that classifiers trained on voice data alone have limited diagnostic utility but may serve as screening tool for general dysphonia.

Though voice alone may have limited diagnostic value, it may bolster the precision of DL-based CDS tools if combined with other modalities, such as imaging or clinicodemographic information. Wang et al. demonstrated that a multimodal learning system combining acoustic signals and medical records outperformed unimodal systems in pathological voice

classification³³. Kwon et al. showed that the classification accuracy of glottic cancer improved when image and voice classifiers were combined with ensemble learning³². Such multimodal approaches are more representative of the clinical decision-making process by human experts, and are expected to provide more reliable results in low validity settings³⁴. With DL models applied to VLS, it will be important to explore whether combining audio with video data enhances performance compared to using video input alone.

The voice dataset used for this study has several advantages. First, unlike previous studies that recorded voice data in controlled environments^{3,12–16}, the present study used voice data collected during office laryngoscopy, i.e. a real-world clinical setting. Hasebe et al. evaluated the impact of Gaussian noise on the performance of voice-based CNN classifiers, finding a substantial decrease in performance with increasing intensities of added noise³⁵. This underscores the challenges of using noisy acoustic data and highlights the value of such data for real-world implementation of AI. Thus, training classifiers on a noisy acoustic dataset representative of the realistic clinical environment is a key strength of our study. Second, each patient only contributed one recording to the dataset. Although this restricted the dataset size, it helped maintain dataset heterogeneity and avoid overrepresenting any individual patient. Finally, only voice samples collected during initial visits or pre-intervention visits were included to accurately represent the pathological conditions.

Analysis of feature importance demonstrated that, across three ML models of increasing complexity, features associated with first to third formants and MFCCs were key contributors to model prediction. In a healthy voice, the fundamental frequency energy is localized at the first two formants of vowel /AH/; this is shown to be affected in dysphonic voices with polyp and adductor spasmodic dysphonia³⁶. Glottal insufficiency is correlated with changes in the intensity region of the first and third formants (F1, F3)³⁷, and previous studies have shown an increase in this frequency region in patients with UVFP^{38,39}. MFCC-related features, particularly MFCC 0–4, have been shown to be key contributors in ML-based discrimination in voices with vocal fold polyps and UVFP⁴⁰ and distinguishing healthy from pathological voice⁴¹. These coefficients are thought to assess the vocal tract independently of the vocal folds.

Using a large standard acoustic parameter set of 88 features, we demonstrate that machine learning algorithms of increasing complexity can differentiate between pathological from healthy voices with high accuracy, but underperform when discriminating between three classes. Further, both the binary and multi-class classifiers performed poorly on the external validation set. Similarly, the DL binary and multiclass classifiers achieved low performance on the external dataset, suggesting limited generalizability across varied datasets. This also shows that voice DL classifiers, sensitive to external variables such as microphone type, recording specifications, and background noise, may not perform reliably in diverse, real-world clinical settings. The classifiers also showed poor generatability to different voice tasks. The binary model trained on Dataset 1 (longest sustained /i/ from each patient) outperformed the binary models trained on Dataset 2 (sustained /i/, alternating /i/ sniffs, and pitch glides) and Dataset 3 (all voice tasks) on all evaluation metrics. The observed trends may be due to overfitting, a phenomenon in which the model learns the training data too closely and cannot generalize to novel datasets. This effect is exacerbated when the size

of the training dataset is small, as in our study. Using a multi-institutional training dataset could have diversified background noise and collection methods, potentially enhancing our classifiers' generalizability.

Recognizing that a small training dataset and class imbalance can severely impair model performance, we applied augmentation techniques to artificially increase dataset size and introduce heterogeneity⁴². However, given the short duration of our audio samples, methods like masking or time shifting did not significantly enhance model performance; pitch augmentation was used instead. Weighted Cross Entropy loss function was employed to address class imbalance in ResNet models, and stratified shuffle split was used for more robust evaluation of ML models. Our models continued to perform poorly on validation despite these techniques, highlighting that dataset size significantly impacts machine learning training and that regularization cannot fully overcome this limitation. Unlike the binary classifiers, which achieved highest accuracy when trained on the smallest dataset (Dataset 1, n=209), the multi-class classifiers performed better with larger training datasets (Dataset 2, n=423; Dataset 3, n=681). This suggests that while the current dataset sufficed for simple binary classification, larger datasets are necessary for more complex multi-class tasks. It is important to note, however, that the multi-class classifier trained on the largest dataset (Dataset 3) still underperformed compared to the binary classifier trained on the smallest dataset (Dataset 1). This underscores the importance of prioritizing dataset quality and diversity over sheer size. Balancing the quality, quantity, and diversity of the training datasets will be the key for improving model performance in audiomics moving forward.

In addition to the small dataset size and lack of heterogeneity, our study had several limitations. The audio samples were derived from a retrospective dataset containing diverse, and sometimes limited, voice tasks during VLS. The examiner's voice occasionally overlapped with the patient's voice, which required the audio samples to be trimmed. Truncating voice tasks can misrepresent the phonation time and quality, which are important clinical features in UVFP and VF lesions. To address this, the longest trimmed sample was selected for each patient in Dataset 1 and 2. We also demonstrated that there were no significant differences in the duration of the audio samples between the three classes. Additionally, we excluded patients with multiple VF pathologies. This may not reflect real-world scenarios, where UVFP is often accompanied by lesions and other pathologies.

Further, all audio samples were isolated by a trained non-expert researcher, though 10% of the dataset was verified by two expert laryngologists to ensure data quality. There is also a lack of explainability in the model's predictions. Although not explicitly shown in the results, there were a few cases that were misclassified with high confidence by the DL algorithms. This emphasizes a need for understanding the model's decision-making process to shed light on the cause of misclassification. Finally, the poor algorithmic performance on the external dataset may be the result of different microphone set up and background noise. Such real-world limitations are important to incorporate, however, to evaluate the implementation potential of AI algorithms in different clinical settings. Our findings are reinforced by a recent scoping review, demonstrating that 88% of 82 studies using voice data for detection of vocal fold pathology since 2000 were aimed at this binary task⁴³. None of the studies included in the scoping review conducted external

validation of their AI models. This lack of external validation is a significant gap in the current literature, and we believe that our study addresses this issue by validating our models on an external dataset. Our study prioritizes bias mitigation, interpretability, and robustness over pure performance optimization. Explainability of models is crucial, as otolaryngologists are more likely to trust ML-CDST output when provided with a specific explanation of its logic⁴⁴. Model agnostic feature importance analysis was used to assess the key acoustic features contributing to model prediction and whether these align with known pathophysiological changes. Further, only the default hyperparameters were used and optimizing for performance gains was avoided to focus on robustness, interpretability, and trustworthiness of the models. Future analyses using feature selection for CNN training and visualization techniques, such as t-distributed stochastic neighbor embedding⁴⁵ and gradient-weighted class activation mapping⁴⁶, could further enhance our efforts by elucidating features in the Mel-spectrogram contributing to model predictions in “black box” neural networks⁴⁷, enhancing interpretability and trust in AI-based diagnostic tools.

Conclusion

DL-based binary and multiclass classifiers were developed to discriminate normal voices from voices with VF lesions and UVFP. Although the algorithm demonstrated high discrimination between normal and pathological voice, its ability to discriminate between HVE, VF lesion, and UVFP was mediocre. Both models generalized poorly to the external validation dataset, suggesting their limited utility in heterogeneous healthcare settings. Acoustic features associated with prediction were elucidated to enhance the interpretability and reliability of the models. Our findings highlight that the voice data collected during office stroboscopy can serve as an accurate screening tool but lacks the reliability for definitive diagnosis, and that training data from diverse real-world setting is critical to ensure generalizability of voice-based DL classifiers. Future work should explore multimodal approaches combining voice and VLS imaging, and employ methods to enhance model explainability.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Fundings and Conflict of Interest:

Anaïs Rameau was supported by Paul B. Beeson Emerging Leaders Career Development Award in Aging (K76 AG079040) from the National Institute on Aging and by the Bridge2AI award (OT2 OD032720) from the NIH Common Fund. Anaïs Rameau is a medical advisor for Perceptron Health, Inc. and of Sound Health Systems, Inc.

References

1. Stroboscopy Woo P.. Plural Publishing, Incorporated, 2009.
2. Ren J, Jing X, Wang Jet al. Automatic Recognition of Laryngoscopic Images Using a Deep-Learning Technique. *Laryngoscope* 2020; 130:E686–e693. [PubMed: 32068890]
3. Reid J, Parmar P, Lund T, Aalto DK, Jeffery CC. Development of a machine-learning based voice disorder screening tool. *Am J Otolaryngol* 2022; 43:103327.

4. Davis RJ, Exilus S, Best S, Willink A, Akst LM. The Geographic Distribution of Videolaryngostroboscopy in the United States. *Journal of Voice* 2023; 37:798.e797–798.e714.
5. Paderno A, Holsinger FC, Piazza C. Videomics: bringing deep learning to diagnostic endoscopy. *Curr Opin Otolaryngol Head Neck Surg* 2021; 29:143–148. [PubMed: 33595977]
6. Yao P, Usman M, Chen YHet al. Applications of Artificial Intelligence to Office Laryngoscopy: A Scoping Review. *Laryngoscope* 2022; 132:1993–2016. [PubMed: 34582043]
7. Cho WK, Choi SH. Comparison of Convolutional Neural Network Models for Determination of Vocal Fold Normality in Laryngoscopic Images. *J Voice* 2022; 36:590–598. [PubMed: 32873430]
8. Tran BA, Dao TTP, Dung HDQet al. Support of deep learning to classify vocal fold images in flexible laryngoscopy. *Am J Otolaryngol* 2023; 44:103800.
9. Yao P, Witte D, Gimonet Het al. Automatic classification of informative laryngoscopic images using deep learning. *Laryngoscope Investig Otolaryngol* 2022; 7:460–466.
10. Bensoussan Y, Elemento O, Rameau A. Voice as an AI Biomarker of Health-Introducing Audiomics. *JAMA Otolaryngol Head Neck Surg* 2024; 150:283–284. [PubMed: 38386315]
11. Han Y, Li S, Cao Y, Ma Q, Yoshikawa M. Voice-Indistinguishability: Protecting Voiceprint In Privacy-Preserving Speech Data Release 2020 IEEE International Conference on Multimedia and Expo (ICME), 2020:1–6.
12. Chen Z, Zhu P, Qiu W, Guo J, Li Y. Deep learning in automatic detection of dysphonia: Comparing acoustic features and developing a generalizable framework. *Int J Lang Commun Disord* 2023; 58:279–294. [PubMed: 36117378]
13. Wu H, Soraghan J, Lowit A, Di Caterina G. Convolutional neural networks for pathological voice detection: IEEE:1–4.
14. Hung C-H, Wang S-S, Wang C-T, Fang S-H. Using SincNet for learning pathological voice disorders. *Sensors* 2022; 22:6634. [PubMed: 36081092]
15. Powell ME, Rodriguez Cancio M, Young Det al. Decoding phonation with artificial intelligence (DeP AI): Proof of concept. *Laryngoscope Investig Otolaryngol* 2019; 4:328–334.
16. Hu HC, Chang SY, Wang CHet al. Deep Learning Application for Vocal Fold Disease Prediction Through Voice Recognition: Preliminary Development Study. *J Med Internet Res* 2021; 23:e25247.
17. Hartl DM, Hans S, Vaissière J, Riquet M, Brasnu DF. Objective voice quality analysis before and after onset of unilateral vocal fold paralysis. *J Voice* 2001; 15:351–361. [PubMed: 11575632]
18. Altman KW. Vocal fold masses. *Otolaryngol Clin North Am* 2007; 40:1091–1108, viii. [PubMed: 17765697]
19. Welch B, Gold K, Gartner-Schmidt J, Petrov A, Law A, Helou LB. Variability of Maximum Glottal Angle on Clinical Sniff Task Differs in Patients With Functional and Organic Laryngeal Pathologies Compared to Healthy Controls. *J Voice* 2024.
20. McFee B, Raffel C, Liang Det al. *librosa: Audio and Music Signal Analysis in Python*. 2015.
21. audEERING GmbH. *openSMILE (Version 2.3)*, 2017.
22. Low DM, Rao V, Randolph G, Song PC, Ghosh SS. Identifying bias in models that detect vocal fold paralysis from audio recordings using explainable machine learning and clinician ratings. *medRxiv* 2024.
23. Eyben F, Scherer KR, Schuller BWet al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing* 2015; 7:190–202.
24. Ramzan F, Khan MU, Rehmat Aet al. A Deep Learning Approach for Automated Diagnosis and Multi-Class Classification of Alzheimer's Disease Stages Using Resting-State fMRI and Residual Neural Networks. *Journal of Medical Systems* 2019; 44.
25. Malekroodi HS, Madusanka N, Lee BI, Yi M. Leveraging Deep Learning for Fine-Grained Categorization of Parkinson's Disease Progression Levels through Analysis of Vocal Acoustic Patterns. *Bioengineering (Basel)* 2024; 11.
26. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition, 2016:770–778.
27. Loshchilov I, Hutter F. Decoupled weight decay regularization. *arXiv preprint arXiv:171105101* 2017.

28. Huang J, Xianhui W, Qi H, Cheng J, Taoran Z. Diabetic Retinopathy Detection with Weighted Cross-Entropy Loss, 2023:8814–8819.
29. Paszke A, Gross S, Massa F et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 2019; 32.
30. Ünalın S, Günay O, Akkurt I, Gunoglu K, Tekin HO. A comparative study on breast cancer classification with stratified shuffle split and K-fold cross validation via ensembled machine learning. *Journal of Radiation Research and Applied Sciences* 2024; 17:101080.
31. Lundberg S. A unified approach to interpreting model predictions. *arXiv preprint arXiv:170507874* 2017.
32. Kwon I, Wang S-G, Shin S-C et al. Diagnosis of early glottic cancer using laryngeal image and voice based on ensemble learning of convolutional neural network classifiers. *Journal of Voice* 2022.
33. Fang S-H, Wang C-T, Chen J-Y, Tsao Y, Lin F-C. Combining acoustic signals and medical records to improve pathological voice classification. *APSIPA Transactions on Signal and Information Processing* 2019; 8:e14.
34. Kline A, Wang H, Li Y et al. Multimodal machine learning in precision health: A scoping review. *NPJ Digit Med* 2022; 5:171. [PubMed: 36344814]
35. Hasebe K, Kojima T, Fujimura S et al. The Effect of Noise on Deep Learning for Classification of Pathological Voice. *Laryngoscope* 2024.
36. Markaki M, Stylianou Y. Voice Pathology Detection and Discrimination Based on Modulation Spectral Features. *IEEE Transactions on Audio, Speech, and Language Processing* 2011; 19:1938–1948.
37. Stevens KN, Hanson HM. Classification of glottal vibration from acoustic measurements. *Vocal fold physiology: Voice quality control* 1995:147–170.
38. Kim K-M, Kakita Y, Hirano M. Sound Spectrographic Analysis of the Voice of Patients with Recurrent Laryngeal Nerve Paralysis: The Korean Society of Laryngology, Phoniatrics and Logopedics.:128–136.
39. Mendoza E, Valencia N, Muñoz J, Trujillo H. Differences in voice quality between men and women: Use of the long-term average spectrum (LTAS). *Journal of voice* 1996; 10:59–66. [PubMed: 8653179]
40. Seedat N, Aharonson V, Hamzany Y. Automated and interpretable m-health discrimination of vocal cord pathology enabled by machine learning 2020 *IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 2020:1–6.
41. Verde L, Pietro GD, Sannino G. Voice Disorder Identification by Using Machine Learning Techniques. *IEEE Access* 2018; 6:16246–16255.
42. Abayomi-Alli OO, Damaševičius R, Qazi A, Adedoyin-Olowe M, Misra S. Data Augmentation and Deep Learning Methods in Sound Classification: A Systematic Review. *Periodical [serial online]*. Date 2022;11. Available from.
43. Liu GS, Jovanovic N, Sung CK, Doyle PC. A Scoping Review of Artificial Intelligence Detection of Voice Pathology: Challenges and Opportunities. *Otolaryngol Head Neck Surg* 2024; 171:658–666. [PubMed: 38738887]
44. Chen H, Ma X, Rives H, Serpedin A, Yao P, Rameau A. Trust in Machine Learning Driven Clinical Decision Support Tools Among Otolaryngologists. *Laryngoscope* 2024; 134:2799–2804. [PubMed: 38230948]
45. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *Journal of machine learning research* 2008; 9.
46. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization 2017 *IEEE International Conference on Computer Vision (ICCV)*, 2017:618–626.
47. Peng X, Xu H, Liu J, Wang J, He C. Voice disorder classification using convolutional neural network based on deep transfer learning. *Scientific Reports* 2023; 13:7264. [PubMed: 37142759]

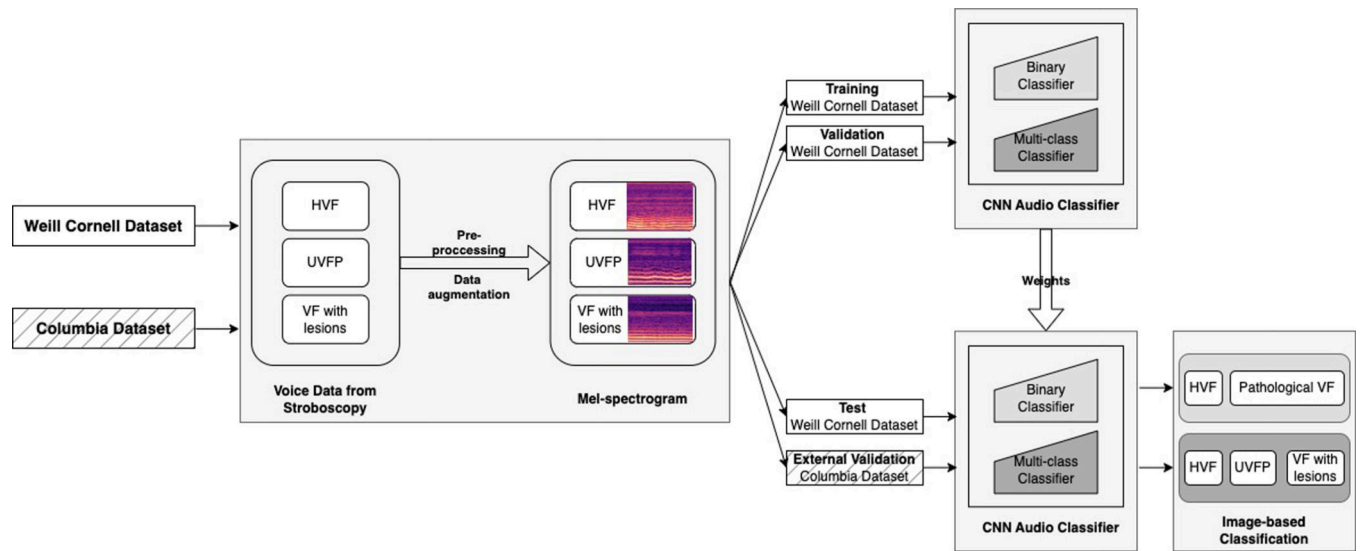


Figure 1.

Simplified schematic diagram of deep learning methods, from videostroboscopic voice data to final prediction. *CNN*: convolutional neural network; *HVF*: healthy vocal folds; *UVFP*: unilateral vocal fold paralysis *VF*: vocal folds.

Table 1:

Binary classifier (healthy vs pathologic) and multi-class classifier (healthy vs vocal fold lesion vs unilateral vocal fold paralysis) performance metrics, on Weill Cornell data

Binary Classifier						Multi-class Classifier					
	Split	Accuracy	Precision	Recall	F1-score		Split	Accuracy	Precision	Recall	F1-score
Dataset 1	Validation	0.6458	0.7187	0.6428	0.70	Dataset 1	Validation	0.25	0.2166	0.2166	0.2166
	Test	0.8333	0.9375	0.8666	0.90		Test	0.40	0.3761	0.3666	0.3643
Dataset 2	Validation	0.70	0.6458	0.7187	0.6428	Dataset 2	Validation	0.45	0.3739	0.3722	0.3603
	Test	0.75	0.6758	0.75	0.6865		Test	0.45	0.4594	0.4222	0.4311
Dataset 3	Validation	0.65	0.5659	0.5937	0.5611	Dataset 3	Validation	0.45	0.4444	0.4222	0.3984
	Test	0.75	0.7222	0.8437	0.715		Test	0.40	0.3666	0.4166	0.3888

Table 2:

Binary classifier and multi-class classifier performance metrics, on Columbia data

Binary Classifier						Multi-class Classifier					
	Split	Accuracy	Precision	Recall	F1-score		Split	Accuracy	Precision	Recall	F1-score
Dataset 1	Test	0.625	0.5694	0.5266	0.4823	Dataset 1	Test	0.35	0.3303	0.3846	0.2466
Dataset 2	Test	0.60	0.5324	0.52	0.4984	Dataset 2	Test	0.425	0.6212	0.4513	0.3673
Dataset 3	Test	0.60	0.5548	0.5466	0.5441	Dataset 3	Test	0.45	0.6706	0.4837	0.3864

Table 3:
Binary and multi-class performance metrics of three machine learning models, on Cornell data. Values are reported as (mean, 95% Confidence Interval)

Binary Classifier							Multi-class Classifier						
	Split	Accuracy	Precision	Recall	F1-score	AUROC		Split	Accuracy	Precision	Recall	F1-score	AUROC
Logistic Regression	Validation	0.89 (0.88–0.90)	0.88 (0.87–0.90)	0.89 (0.88–0.90)	0.88 (0.87–0.89)	0.88 (0.86–0.90)	Logistic Regression	Validation	0.56 (0.54, 0.58)	0.57 (0.55, 0.59)	0.56 (0.54, 0.58)	0.56 (0.54, 0.58)	0.73 (0.71, 0.74)
Random Forest	Validation	0.87 (0.86–0.88)	0.87 (0.85–0.88)	0.87 (0.86–0.88)	0.85 (0.83–0.86)	0.86 (0.84–0.88)	Random Forest	Validation	0.65 (0.63, 0.68)	0.67 (0.64, 0.69)	0.65 (0.63, 0.68)	0.63 (0.61, 0.66)	0.79 (0.78, 0.81)
Multi-Layer Perceptron	Validation	0.89 (0.88–0.91)	0.89 (0.88–0.90)	0.89 (0.88–0.91)	0.89 (0.88–0.90)	0.89 (0.87–0.90)	Multi-Layer Perceptron	Validation	0.62 (0.60, 0.63)	0.62 (0.60, 0.64)	0.62 (0.60, 0.63)	0.61 (0.59, 0.63)	0.76 (0.74, 0.77)

Table 4:
Binary and multi-class performance metrics of three machine learning models, on Columbia data.

Binary Classifier							Multi-class Classifier						
	Split	Accuracy	Precision	Recall	F1-score	AUROC		Split	Accuracy	Precision	Recall	F1-score	AUROC
Logistic Regression	Test	0.600	0.3846	0.600	0.4688	0.5680	Logistic Regression	Test	0.40	0.2483	0.40	0.3058	0.5864
Random Forest	Test	0.625	0.3906	0.625	0.4808	0.5987	Random Forest	Test	0.35	0.2142	0.35	0.2249	0.6388
Multi-Layer Perceptron	Test	0.600	0.3846	0.600	0.4688	0.5520	Multi-Layer Perceptron	Test	0.35	0.2111	0.35	0.2538	0.5343