Received: 6 August 2023 | Accepted: 13 August 2023

DOI: 10.1002/lio2.1144

ORIGINAL RESEARCH

Laryngoscope
Investigative Otolaryngology

End-to-end deep learning classification of vocal pathology using stacked vowels

George S. Liu MD[1,2] | Jordan M. Hodges BS[3] | Jingzhi Yu BA[4] | C. Kwang Sung MD, MS[1,2] | Elizabeth Erickson-DiRenzo PhD[1,2] | Philip C. Doyle PhD[1,2]

[1]Department of Otolaryngology Head and Neck Surgery, Stanford University School of Medicine, Stanford University, Stanford, California, USA

[2]Division of Laryngology, Stanford University School of Medicine, Stanford University, Stanford, California, USA

[3]Computer Science Department, School of Engineering, Stanford University, Stanford, California, USA

[4]Biomedical Informatics, Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, California, USA

Correspondence
Philip C. Doyle, Otolaryngology Head and Neck Surgery, Division of Laryngology, Stanford University School of Medicine, Stanford University, 801 Welch Road, Stanford, CA 94035, USA.
Email: pdoyle2@stanford.edu

Abstract

Objectives: Advances in artificial intelligence (AI) technology have increased the feasibility of classifying voice disorders using voice recordings as a screening tool. This work develops upon previous models that take in single vowel recordings by analyzing multiple vowel recordings simultaneously to enhance prediction of vocal pathology.

Methods: Voice samples from the Saarbruecken Voice Database, including three sustained vowels (/a/, /i/, /u/) from 687 healthy human participants and 334 dysphonic patients, were used to train 1-dimensional convolutional neural network models for multiclass classification of healthy, hyperfunctional dysphonia, and laryngitis voice recordings. Three models were trained: (1) a baseline model that analyzed individual vowels in isolation, (2) a stacked vowel model that analyzed three vowels (/a/, /i/, /u/) in the neutral pitch simultaneously, and (3) a stacked pitch model that analyzed the /a/ vowel in three pitches (low, neutral, and high) simultaneously.

Results: For multiclass classification of healthy, hyperfunctional dysphonia, and laryngitis voice recordings, the stacked vowel model demonstrated higher performance compared with the baseline and stacked pitch models (F1 score 0.81 vs. 0.77 and 0.78, respectively). Specifically, the stacked vowel model achieved higher performance for class-specific classification of hyperfunctional dysphonia voice samples compared with the baseline and stacked pitch models (F1 score 0.56 vs. 0.49 and 0.50, respectively).

Conclusions: This study demonstrates the feasibility and potential of analyzing multiple sustained vowel recordings simultaneously to improve AI-driven screening and classification of vocal pathology. The stacked vowel model architecture in particular offers promise to enhance such an approach.

George Siyuan Liu, Jordan M. Hodges, and Jingzhi Yu contributed equally to this study.

This was presented as a podium presentation at the Triological Society Spring Meeting, Boston, Massachusetts, USA (COSM 2023, May 4).

Lay Summary: AI analysis of multiple vowel recordings can improve classification of voice pathologies compared with models using a single sustained vowel and offer a strategy to enhance AI-driven screening of voice disorders.

Level of Evidence: 3

KEYWORDS

artificial intelligence, deep learning, voice classification, voice disorders, voice pathology

1 | INTRODUCTION

Interest in the evaluation and classification of voice disorders is longstanding in both the basic and clinical literatures. Vocal function can be quantified using a range of objective instrumental measures including acoustic and aerodynamic analyses, as well as examination and visual description using laryngoscopy and stroboscopy.[1] However, because alterations in voice quality are complex and multidimensional, crossing the domains of frequency (pitch), intensity (loudness), and duration (temporal) features, classification of vocal pathology is challenging. While changes in voice quality may be identified using auditory-perceptual methods, subjective judgments by listeners may be influenced by multiple factors including listener experience, perceptual bias, the type of stimuli evaluated, measurement methods, etc.[2–5] Thus, contemporary objective applications that seek to exploit computerized prediction models to identify vocal fold pathologies have become of substantial interest in recent years.[6,7]

It is well documented that changes in voice quality can serve as a critical element in the diagnostic process. In fact, in some instances the underlying laryngeal abnormality may serve to characterize specific classes of voice disorders. Because voice disorders may evolve over time with subsequent changes in voice quality or intermittent loss of voice, compensations also may occur. In some instances, a volitional increase in laryngeal adductory force may be required to meet voicing requirements when the speaker experiences a reduction in voice efficiency or changes in quality.[8,9] These types of changes will have a direct influence on vocal fold vibration and subsequently on voice quality. Similarly, disorders that involve incomplete adduction of the vocal folds will result in varied degrees of air leakage with the perceptual identification of breathiness.

Although classification of voice disorders based on objective and subjective characteristics has been longstanding,[10,11] the application of deep learning approaches may provide valuable information. However, to date, limited work has been performed with the objective of classifying the underlying categories of voice disorders. Empirical efforts may provide additional insights for both identifying and classifying a range of vocal fold abnormalities. Ultimately, assessment methods that move beyond the binary distinction between healthy and pathologic voices can be of substantial importance. Because of this need, machine learning methods may offer valuable insights. Such approaches generally rely on the analysis of either a sustained vowel sample, or in some cases standardized sentences.[12] Because vowels are relatively steady state, quasiperiodic entities, they offer a distinct advantage over the use of sentences or other types of running speech. Additionally, use of vowel stimuli for vocal analysis is increasingly being used in situations of nonoptimized recording conditions.[13]

Based on existing data, deep learning algorithms may be used to identify latent relationships between vowel samples which can then assist in the classification of different vocal abnormalities. Classification models built from voice samples have been developed utilizing various strategies, including both deep learning and more traditional machine learning approaches.[14] Briefly, many current models utilize support vector machines, multilayer perceptrons, and random forests, with deep learning models gaining in popularity.[7] Feature extraction using deep learning almost always occurs as a distinct step before training and prediction. Presently, Mel-frequency cepstral coefficients (MFCCs) remain the most common features that are relied on for voice prediction.[6,7,15–18] Alternatively, the analysis of spectrograms has also been explored in analyzing auditory data for similar tasks.[19]

The current popularity of the “2 step approach” (i.e., expert derived feature extraction followed by training) leaves open the question of the efficacy of a fully end-to-end approach (i.e., training with minimally processed, raw data) to classification. Recent research has reported accuracy as high as 92% in identifying vocal patterns associated with Parkinson's disease[17] and 89% for classification of environmental sounds with an end-to-end approach.[20] Yet, potential features that may be useful for classifying other types of voice pathology remain overlooked solely by extracting MFCCs. In addition, the use of MFCCs is also dependent on the temporal window size applied in the transformation process.[6] Consequently, the use of raw audio samples instead of MFCCs to train a deep neural network may better predict categories of vocal fold pathology when employing a fully end-to-end approach.

Therefore, the present project sought to develop an end-to-end deep learning framework, named *VocalPathNet*, for detecting and classifying vocal fold pathology using recordings of multiple sustained vowels. To do this, we trained a baseline model that took individual vowel recordings as inputs and compared its results for vocal fold pathology classification with the proposed framework; this involved a process that analyzed stacked vowels recordings to classify healthy and pathological voice recordings and distinguish vocal pathologies of hyperfunctional dysphonia and laryngitis.[21]

2 | METHODS

Voice data used in this project was accessed from the Saarbruecken voice database (SVD) hosted by Saarland University.[21] This database contains labeled healthy and pathological audio samples consisting of both sustained vowels (/a/, /i/, and /u/) and sentences.[22,23] The vowel samples were recorded in three different pitches: low, neutral, and high. The database includes recordings from over 2000 German-speaking individuals with over 70 class labels.[21] We used a subset of the voice recordings, including recordings from healthy patients and patients with two vocal fold pathologies: hyperfunctional dysphonia and laryngitis. These pathologies were chosen based on their presence in related works.[15] Only pathologic voice recordings with tags of either hyperfunctional dysphonia or laryngitis were included (i.e., voice recordings with tags of both laryngitis and hyperfunctional dysphonia were excluded). Voice recordings of the vowels /a/, /i/, and /u/ produced at a low, neutral, and high pitch were used. These vowels and pitches were chosen because of their availability in the SVD dataset. Altogether, our dataset consisted of 687 healthy participants, 207 hyperfunctional dysphonia patients, and 127 laryngitis patients, with up to 9 vowel recordings per participant, depending on the model (Table 1). Age and sex demographic characteristics of participants are shown in Table 2.

The raw audio data were extracted at a sampling rate of 44,100 Hz using the *librosa* package in Python.[24] Due to the variability in the recording lengths, this sampling rate was chosen to provide a relatively large number of data points that could be included even after trimming. We also examined the presence of silence in the raw audio data and did not find substantial periods of silence. All recordings were trimmed to clips of a duration of 0.50 s.

To explore the efficacy of our proposed deep learning framework, we trained three different model architectures: (1) a baseline model architecture that takes in individual neutral-pitch vowel recordings as inputs; (2) a *VocalPathNet* stacked-vowel model architecture that takes in stacked recordings of three vowels (/a/, /i/, and /u/) in the neutral pitch as inputs; (3) a *VocalPathNet* stacked-pitch model architecture that takes in stacked recordings of the three pitches (low, neutral, and high) for each vowel (/a/, /i/, and /u/). Conceptually, stacked recordings were constructed in a way that would be akin to playing the recordings simultaneously albeit via different input streams so that the recordings were separable by the listener. All model architectures were used to train candidate models for the task of 3-class multiclass classification of healthy, hyperfunctional dysphonia, and laryngitis voice recordings.

TABLE 1 Counts of participants by condition and sex.

Class	Male	Female	Total
Healthy	259	428	687
Hyperfunctional dysphonia	42	165	207
Laryngitis	75	52	127
Total pathological (hyperfunctional dsyphonia and laryngitis)	117	217	334

For the baseline model architecture, a total of 3063 neutral pitch vowel audio samples (/a/, /i/, and /u/) from all included participants were split into training, validation, and testing datasets in a 60:20:20 ratio, stratified by class labels. For the *VocalPathNet* stacked-vowel model architecture, neutral pitch vowel recordings of the vowels were stacked in the order of /a/, /i/, and /u/. The final sample count after stacking was 1021. For the *VocalPathNet* stacked-pitch model architecture, the three pitch vowel recordings of each vowel were stacked in the order of low, neutral, and high. The final sample count after stacking was 3063.

For our deep learning frameworks, we utilized a 1-dimensional convolutional neural network (1-D CNN) model architecture. This model architecture uses convolutional filters and weight sharing to analyze the one-dimensional, time domain audio data. The sequential nature of the audio data and presence of semi-repetitive qualities, such as jitter and shimmer, makes the 1-D CNN architecture apropos for identifying relevant features within the recordings. Following the convolutional and pooling layers, the sex and age demographic data of each speaker were combined with the audio data. These were then fed into fully connected layers utilizing the rectified linear unit activation function. The prediction layer utilized three outputs, one for each class. The architectures of the baseline and *VocalPathNet* CNN models are shown in Figure 1. The key difference between the baseline and *VocalPathNet* model architectures is that the latter had input layers that accepted a quasi-one-dimensional input of the three-channel stacked vowel and stacked pitch input data. Models were implemented using the *Keras* deep learning library.[25] Models used the categorical cross entropy loss function for three-class multiclass classification of healthy, hyperfunctional dysphonia, and laryngitis class labels.

For model training, the *Adam* optimization algorithm was used to update the initial learning rate using exponentially weighted moving averages of the gradient and squared gradient of each learnable parameter.[26] Hyperparameter tuning was done via a combination of the random search feature in the *keras_tuner* package, as well as manual training and error analysis to optimize the validation loss. The hyperparameters that were tuned were the number of filters, filter size, size of the pooling layers, batch size, and the initial learning rate.

For model evaluation, the best performing candidate models of each model architecture were compared for each voice classification task using two performance metrics: area under the receiver operating

TABLE 2 Counts of the age bin distribution across the participant dataset by sex.

	Age brackets									
Sex	1–10	11–20	21–30	31–40	41–50	51–60	61–70	71–80	81–90	Total
M	75	29	122	74	43	16	17	0	0	376
F	1	197	217	67	70	54	28	10	1	645

Abbreviations: M, male; F, female.

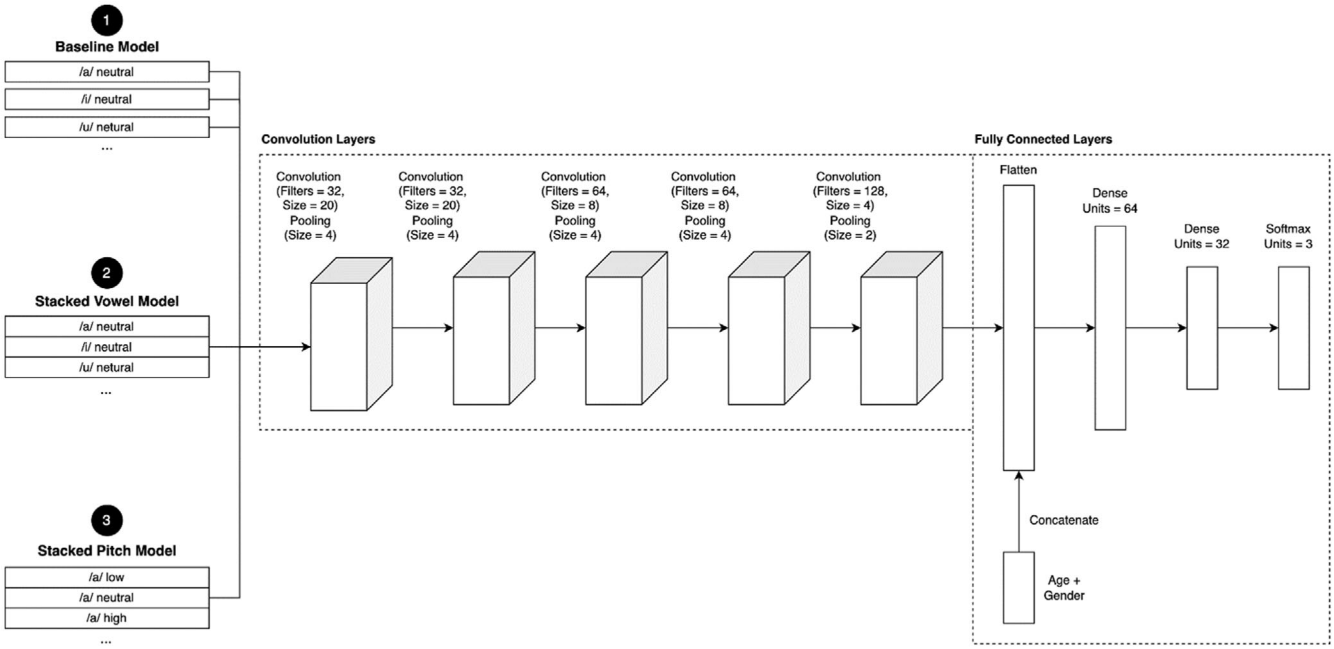
1
Baseline Model
/a/ neutral
/i/ neutral
/u/ netural
...
2
Stacked Vowel Model
/a/ neutral
/i/ neutral
/u/ netural
...
3
Stacked Pitch Model
/a/ low
/a/ neutral
/a/ high
...
Convolution Layers
Convolution (Filters = 32, Size = 20) Pooling (Size = 4)
Convolution (Filters = 32, Size = 20) Pooling (Size = 4)
Convolution (Filters = 64, Size = 8) Pooling (Size = 4)
Convolution (Filters = 64, Size = 8) Pooling (Size = 4)
Convolution (Filters = 128, Size = 4) Pooling (Size = 2)
Fully Connected Layers
Flatten
Dense Units = 64
Dense Units = 32
Softmax Units = 3
Concatenate
Age + Gender

FIGURE 1 Schematic of convolutional neural network architectures for baseline, *VocalPathNet* stacked-vowel, and *VocalPathNet* stacked-pitch models. The convolution layers are 1-dimensional for the baseline model and 2-dimensional for the *VocalPathNet* models.

curve (AUROC) and the F1 score. To address class imbalance, the AUROC was weighted by the number of samples representing each class label. The F1 score was calculated as the harmonic mean of the precision and recall. The F1 score metric was chosen because it works well for imbalanced datasets. Overall, F1 score metrics for models were calculated using the micro-average weighting scheme, to further account for the class imbalance. The precision and recall scores used to calculate F1 scores were also evaluated.

3 | RESULTS

For the task of multiclass classification of healthy and dysphonic voices, the baseline model and *VocalPathNet* stacked-pitch model achieved similar F1 scores (F1 score 0.77 vs. 0.78, respectively), while the *VocalPathNet* stacked-vowel model's F1 score was higher (0.80). The baseline model, stacked vowel and stacked pitch models had similar AUROC scores (0.90, 0.90, and 0.89, respectively; Table 3). These AUROC values should be interpreted with caution because our dataset was imbalanced.[27] As a reference, human expert classification of related vocal pathologies achieved an accuracy around 0.6.[18]

We further examined class-specific classification performance metrics for each model (Table 4). For the classification of hyperfunctional dysphonia, the stacked vowel model achieved a higher F1 score than the baseline and stacked pitch models (F1 score 0.56 vs. 0.49 and 0.50, respectively). The baseline, stacked vowel, and stacked pitch models achieved similar F1 scores for the classification of healthy and laryngitis voice recordings (Table 4). Precision and recall values for the models' test performance results are shown in Tables 3 and 4. Confusion matrices showing the test performance of the baseline, stacked vowel, and stacked pitch models are shown in Figure 2.

TABLE 3 Comparison of test performance of multiclass classification by the three models.

Model	Precision	Recall	F1 score	Area under the receiver operating curve
Baseline	0.78	0.78	0.77	0.90
Stacked vowel	0.81	0.79	0.80	0.90
Stacked pitch	0.78	0.78	0.78	0.89

Note: Reported precision, recall, and F1 scores are weighted by class using the microaverage weighting scheme.

4 | DISCUSSION

Advances in artificial intelligence have aided the objective analysis of voice samples to predict vocal pathology. Most prior machine learning models for detecting vocal pathology have two characteristics. First, a two-stage approach is used in which expert-derived voice features, most commonly the MFCC,[6,12,15,16] are calculated from the raw voice data and used to predict vocal pathology.[28] Second, a single sustained vowel recording (e.g., selected vowel /a/ samples) is used as the initial raw data input.[15,16,29] Even when the model is trained with different sustained vowel recordings (e.g., selected vowel /a/, /i/, or /u/ samples), only an individual vowel recording is input to generate a prediction of vocal pathology. Our neural network model, *VocalPathNet*, is distinguished in both aspects as it directly learns from voice data via an end-to-end deep learning approach and uses stacked vowel

TABLE 4 Comparison of class-specific metrics in test classification performance by the three models.

	Healthy			Hyperfunctional dysphonia			Laryngitis		
Model	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Baseline	0.83	0.90	0.86	0.51	0.46	0.49	0.94	0.65	0.77
Stacked vowel	0.88	0.86	0.87	0.50	0.63	0.56	0.94	0.65	0.77
Stacked pitch	0.83	0.89	0.86	0.51	0.49	0.50	0.94	0.65	0.77

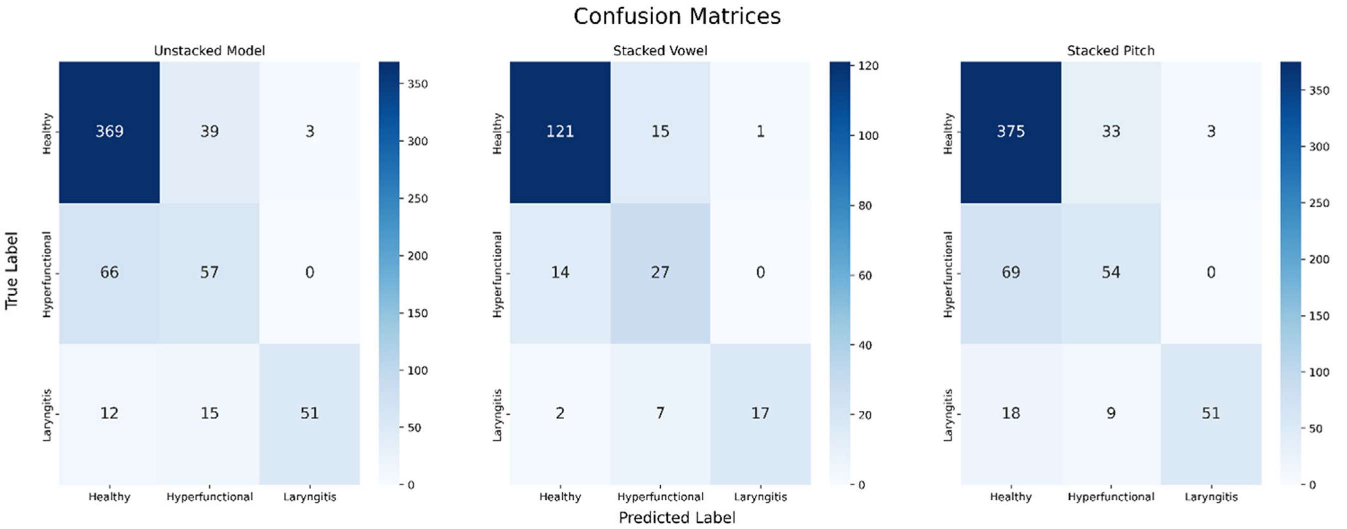

FIGURE 2 Confusion matrices of test classification results from the unstacked (baseline), stacked vowel, and stacked pitch models on reserved test subsets.

recordings (/a/, /i/, and /u/ in neutral pitch) or stacked pitch recordings as inputs to simultaneously analyze multiple sustained vowels to generate vocal pathology predictions. To our knowledge, this is the first investigation of a deep learning strategy that combines end-to-end learning and stacked vowel inputs for vocal pathology classification. The primary contribution of this work is the proof-of-concept demonstration of a novel CNN model architecture that can input stacked vowels and potentially improve classification of vocal pathology compared with the conventional use of single vowel inputs. We also believe that the product of this work sets the stage for future investigations.

The proposed framework offers several advantages. The rationale for using stacked vowel inputs is that the different vowels, analyzed together, provide a richer set of information than single vowel recordings for analyzing the effects of vocal pathology on different expression patterns. For example, muscle tension dysphonia is associated with problems with tonal pitch variation, particularly with targeting higher pitches.[30] The use of an end-to-end deep learning approach is apropos for analyzing the stacked vowel inputs, as voice features for stacked vowel inputs have not yet been derived. Further, an end-to-end learning approach also can assess all potentially relevant voice features which overcomes limitations of a two-step approach which requires pre-selection of expert derived voice features that can substantially affect the performance of the machine learning classifier, such as the choice of the temporal window size for the use of MFCCs.[6] The *VocalPathNet* framework requires minimal additional preprocessing to concatenate single vowel recordings into a stacked vowel input. Sustained vowel recordings are relatively steady state and quasiperiodic, so recordings can be stacked without the need to match phases and temporally align them.

This study is preliminary and has limitations. As in many proof-of-concept studies, the size of the training dataset is small which may limit the performance of models. This may explain in part differences in the performance between the *VocalPathNet* stacked vowel and stacked pitch models, as the baseline, stacked vowel, and stacked pitch models have successively higher numbers of model parameters and higher risks of overfitting to a small training dataset size. Prior research suggests that a training dataset size of 20,000 chest x-rays may be sufficient to enable CNN models to perform binary classification triaging of normal versus abnormal chest x-rays.[31] Our processed voice recordings had 10,000 samples which is two orders of magnitude smaller than the number of pixels in a 2000 × 2500 pixels x-ray. A back of the envelope calculation suggests, therefore that a training dataset size of around 1000 voice recording samples may be sufficient to train a voice classification model; this is approximately the size of our entire dataset. It would be reasonable to expect that *VocalPathNet* model classification results would improve with expansion of our training dataset size by an order of magnitude.

The present work used the SVD dataset, one which contained a class imbalance with more healthy than dysphonic voice recordings and among vocal pathology recordings, more hyperfunctional dysphonia than laryngitis recordings. We explored the use of class weighted loss functions and over- and under-sampling of classes to address imbalance in the dataset, however these strategies did not improve our classification results likely because of the small dataset size (data not shown). Other methods to address class imbalance, such as future development of algorithms for data augmentation of voice recordings,[32] could potentially help address this issue in future work. The generalizability of the current model was also only evaluated on a test hold-out set from a single open dataset (SVD[21]); future iterations of the models could be validated using additional open labeled datasets of voice recordings, such as the MEEI Voice Disorders Database[33] and/or the VOICED database.[34] The training dataset could further be broadened to expand the application and generalizability of the *VocalPathNet* framework. This could be done by including recordings of sustained vowels in different pitches, additional vocal pathologies, and recordings in other languages.

A longstanding challenge in the practical application of deep learning models is interpretability. Visualization techniques, such as saliency maps and occlusion maps,[35] are well established in computer vision for providing insight into the specific regions of images that are used for image classification and detection.[36–38] Approaches using spectrograms[19] could allow application of visualization techniques to delineate areas of spectrograms that are influential on vocal pathology predictions. Visualization methods are also being developed for audio deep neural networks and may be more available in the future.[39] Lastly, the issue of voice classification and associated terminology poses a challenge.[39] As an example, the term "hyperfunctional" voice disorders[8,9] represents a pathophysiologic process that may result in a range of glottal pathology. That is, such disorders may represent changes to the vocal fold(s) that could result in varied levels of edema, lesions on the membranous glottis (i.e., nodules or polyps), or pathology at the posterior glottis (contact ulcers). Consequently, as deep learning models progress, more specific information on the presence and location of pathology in ground truth class labels would be beneficial.

5 | CONCLUSION

A deep neural network framework, *VocalPathNet*, can input multiple sustained vowel recordings simultaneously to potentially improve classification of vocal pathologies using raw audio recordings. This preliminary work sets the groundwork to improve voice pathology classification using more of the available voice recording data in future deep learning applications.

ORCID

George S. Liu https://orcid.org/0000-0002-2066-5734
C. Kwang Sung https://orcid.org/0000-0001-9795-6452
Elizabeth Erickson-DiRenzo https://orcid.org/0000-0002-9841-0184
Philip C. Doyle https://orcid.org/0000-0002-2715-3619

REFERENCES

1. Mehta DD, Hillman RE. Current role of stroboscopy in laryngeal imaging. *Curr Opin Otolaryngol Head Neck Surg*. 2012;20(6):429-436. doi:10.1097/MOO.0b013e3283585f04
2. Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Hear Res*. 1993;36(1):21-40. doi:10.1044/jshr.3601.21
3. Kreiman J, Gerratt BR, Ito M. When and why listeners disagree in voice quality assessment tasks. *J Acoust Soc Am*. 2007;122(4):2354-2364. doi:10.1121/1.2770547
4. Kreiman J, Gerratt BR. The perceptual structure of pathologic voice quality. *J Acoust Soc Am*. 1996;100(3):1787-1795. doi:10.1121/1.416074
5. Kreiman J, Gerratt BR, Berke GS. The multidimensional nature of pathologic vocal quality. *J Acoust Soc Am*. 1994;96(3):1291-1302. doi:10.1121/1.410277
6. Chen L, Chen J. Deep neural network for automatic classification of pathological voice signals. *J Voice*. 2022;36(2):288.E15-288.E24. doi:10.1016/j.jvoice.2020.05.029
7. Sáenz-Lechón N, Godino-Llorente JI, Osma-Ruiz V, Gómez-Vilda P. Methodological issues in the development of automatic systems for voice pathology detection. *Biomed Signal Process Control*. 2006;1(2):120-128. doi:10.1016/j.bspc.2006.06.003
8. Hillman RE, Holmberg EB, Perkell JS, Walsh M, Vaughan C. Objective assessment of vocal hyperfunction: an experimental framework and initial results. *J Speech Hear Res*. 1989;32(2):373-392. doi:10.1044/jshr.3202.373
9. Hillman RE, Stepp CE, Van SJH, Zañartu M, Mehta DD. An Updated Theoretical Framework for Vocal Hyperfunction. *Am J Speech Lang Pathol*. 2020;29(4):2254-2260. doi:10.1044/2020_AJSLP-20-00104
10. Murry T, Singh S, Sargent M. Multidimensional classification of abnormal voice qualities. *J Acoust Soc Am*. 1977;61(6):1630-1635. doi:10.1121/1.381439
11. Prosek RA, Montgomery AA, Walden BE, Hawkins DB. An evaluation of residue features as correlates of voice disorders. *J Commun Disord*. 1987;20(2):105-117. doi:10.1016/0021-9924(87)90002-5
12. Godino-Llorente JI, Gomez-Vilda P. Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. *IEEE Trans Biomed Eng*. 2004;51(2):380-384. doi:10.1109/TBME.2003.820386
13. van der Woerd B, Wu M, Parsa V, Doyle PC, Fung K. Evaluation of Acoustic Analyses of Voice in Nonoptimized Conditions. *J Speech Lang Hear Res*. 2020;63(12):3991-3999. doi:10.1044/2020_JSLHR-20-00212
14. Crowson MG, Ranisau J, Eskander A, et al. A contemporary review of machine learning in otolaryngology–head and neck surgery. *Laryngoscope*. 2020;130(1):45-51. doi:10.1002/lary.27850
15. Tirronen S, Kadiri SR, Alku P. The effect of the MFCC frame length in automatic voice pathology detection. *J Voice*. 2022;27:S0892. doi:10.1016/j.jvoice.2022.03.021
16. Al-Dhief FT, Baki MM, Latiff NMA, et al. Voice pathology detection and classification by adopting online sequential extreme learning machine. *IEEE Access*. 2021;9:77293-77306. doi:10.1109/ACCESS.2021.3082565
17. Syed SA, Rashid M, Hussain S, Zahid H. Comparative analysis of CNN and RNN for voice pathology detection. *Biomed Res Int*. 2021;2021:e6635964. doi:10.1155/2021/6635964
18. Hu HC, Chang SY, Wang CH, et al. Deep learning application for vocal fold disease prediction through voice recognition: preliminary

development study. *J Med Internet Res*. 2021;23(6):e25247. doi:10.2196/25247

19. Powell ME, Rodriguez Cancio M, Young D, et al. Decoding phonation with artificial intelligence (DeP AI): Proof of concept. *Laryngoscope Invest Otolaryngol*. 2019;4(3):328-334. doi:10.1002/lio2.259
20. Abdoli S, Cardinal P, Koerich AL. End-to-end environmental sound classification using a 1D convolutional neural network. *Expert Syst Appl*. 2019;136:252-263. doi:10.48550/arXiv.1904.08990
21. Barry WJ, Putzer M. *Saarbruecken Voice Database*, Institute of Phonetics, Univ. of Saarland, http://www.stimmdatenbank.coli.uni-saarland.de/.
22. Martínez D, Lleida E, Ortega A, Miguel A, Villalba J. Voice pathology detection on the Saarbrücken voice database with calibration and fusion of scores using multifocal toolkit. In: Torre Toledano D, Ortega Giménez A, Teixeira A, et al., eds. *Advances in Speech and Language Technologies for Iberian Languages*. Springer; 2012:99-109. doi:10.1007/978-3-642-35292-8_11
23. Lee JY. Experimental evaluation of deep learning methods for an intelligent pathological voice detection system using the Saarbruecken voice database. *Appl Sci*. 2021;11(15):7149. doi:10.3390/app11157149
24. McFee B, McVicar M, Faronbi D, et al. *librosa/librosa: 0.10.0.post2*. Zenodo; 2023. doi:10.5281/zenodo.7746972
25. Chollet F. *Deep learning for humans*. Vol 12. Keras; 2023 https://github.com/keras-team/keras
26. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv14126980 Cs*. 2021.
27. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432
28. Omeroglu AN, Mohammed HMA, Oral EA. Multi-modal voice pathology detection architecture based on deep and handcrafted feature fusion. *Eng Sci Technol Int J*. 2022;36:101148. doi:10.1016/j.jestch.2022.101148
29. Areiza-Laverde HJ, Castro-Ospina AE, Peluffo-Ordóñez DH. Voice pathology detection using artificial neural networks and support vector machines powered by a multicriteria optimization algorithm. In: Figueroa-García JC, López-Santana ER, Rodriguez-Molano JI, eds. *Applied Computer Sciences in Engineering. Communications in Computer and Information Science*. Springer International Publishing; 2018:148-159. doi:10.1007/978-3-030-00350-0_13
30. Nguyen DD, Kenny DT. Impact of muscle tension dysphonia on tonal pitch target implementation in Vietnamese female teachers. *J Voice*. 2009;23(6):690-698. doi:10.1016/j.jvoice.2008.01.007
31. Dunnmon JA, Yi D, Langlotz CP, Ré C, Rubin DL, Lungren MP. Assessment of convolutional neural networks for automated classification of chest radiographs. *Radiology*. 2019;290(2):537-544. doi:10.1148/radiol.2018181422
32. Tak H, Kamble M, Patino J, Todisco M, Evans N. *RawBoost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing*. arXiv.org; 2021.
33. Eye M, Infirmary E. *Voice disorders database, version. 1.03 (CD-ROM)*. Linc Park NJ Kay Elemetrics Corp; 1994.
34. Cesari U, Pietro GD, Marciano E, Niri C, Sannino G, Verde L. A new database of healthy and pathological voices. *Comput Electr Eng*. 2018;68:310-321. doi:10.1016/j.compeleceng.2018.04.008
35. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, eds. *Computer Vision – ECCV 2014. Lecture Notes in Computer Science*. Springer International Publishing; 2014:818-833. doi:10.1007/978-3-319-10590-1_53
36. Liu GS, Zhu MH, Kim J, Raphael P, Applegate BE, Oghalai JS. ELH-net: a convolutional neural network for classifying cochlear endolymphatic hydrops imaged with optical coherence tomography. *Biomed Opt Express*. 2017;8(10):4579-4594. doi:10.1364/BOE.8.004579
37. Liu GS, Shenson JA, Farrell JE, Blevins NH. Signal to noise ratio quantifies the contribution of spectral channels to classification of human head and neck tissues ex vivo using deep learning and multispectral imaging. *J Biomed Opt*. 2023;28(1):16004. doi:10.1117/1.JBO.28.1.016004
38. Liu GS, Yang A, Kim D, et al. Deep learning classification of inverted papilloma malignant transformation using 3D convolutional neural networks and magnetic resonance imaging. *Int Forum Allergy Rhinol*. 2022;12(8):1025-1033. doi:10.1002/alr.22958
39. Krug A, Ebrahimzadeh M, Alemann J, Johannsmeier J, Stober S. Analyzing and visualizing deep neural networks for speech recognition with saliency-adjusted neuron activation profiles. *Electronics*. 2021;10(11):1350. doi:10.3390/electronics10111350

How to cite this article: Liu GS, Hodges JM, Yu J, Sung CK, Erickson-DiRenzo E, Doyle PC. End-to-end deep learning classification of vocal pathology using stacked vowels. *Laryngoscope Investigative Otolaryngology*. 2023;8(5):1312-1318. doi:10.1002/lio2.1144