ConVox: A Robust Deep Learning Approach for Accurate Voice Disorder Detection with Multilingual Capabilities

Jason Hao*
Computer Systems Lab
Thomas Jefferson High School
for Science and Technology
Alexandria, United States
2025jhao@tjhsst.edu

Shaurya Jain*
Computer Systems Lab
Thomas Jefferson High School
for Science and Technology
Alexandria, United States
2025sjain@tjhsst.edu

Soham Jain*
Computer Systems Lab
Thomas Jefferson High School
for Science and Technology
Alexandria, United States
2025sjain1@tjhsst.edu

Anmol Karan*
Computer Systems Lab
Thomas Jefferson High School
for Science and Technology
Alexandria, United States
2025akaran@tjhsst.edu

*Denotes equal contribution from all authors, with names listed alphabetically

***Abstract*—Voice disorders significantly impact an individual's ability to communicate verbally, particularly affecting the elderly community. Diagnosing these disorders is complex, often hindered by the limitations of traditional imaging techniques. This study presents a novel deep learning framework for voice disorder detection through audio classification, addressing the challenge of diagnosing these disorders that affect a large proportion of elderly adults in America. Our model, ConVox, utilizes a sequential stack of one-dimensional convolutional neural networks to conduct binary classification of voice disorders. We leverage four large datasets: Advanced Voice Function Assessment Databases, Saarbrücken Voice Database, TORGO Database, and UA Speech Database, which together comprise 22,883 audio samples in Waveform Audio File Format. The model achieved notable accuracies of 99.89% in training, 99.91% in validation, and 99.74% in testing, outperforming existing models. With an area-under-curve of 0.999995, precision of 0.9972, and recall of 0.9994, our model demonstrates exceptional performance in accurately identifying voice disorders with a very low rate of false positives and false negatives. Additionally, this model demonstrates promising performance across multiple languages and voice pathologies. ConVox's comprehensiveness and high accuracy demonstrate that it is a promising tool for audio classification, potentially enhancing healthcare outcomes for individuals with voice disorders.**

***Index Terms*—voice disorder detection, deep learning, CNN**

I. Introduction

Voice disorders are an extensive issue worldwide, with a notable prevalence among the elderly. In fact, Cleveland Clinic [1] quantifies that 3-9% of the U.S. population experiences a voice disorder at some point in their lives. These disorders can severely damage one's ability to speak, manifesting in various forms of muscle impairment, fatigue, and tissue damage. The implications of these conditions are profound, affecting communication, social interaction, and overall quality of life.

According to the American Speech-Language-Hearing Association [2], a disturbance in the respiratory system, nervous system, laryngeal muscles, pharynx, or oral cavity are the most common causes of voice disorders. Structural causes such as vocal folds—which can develop nodules, polyps, or cysts—can lead to voice disorders. These structural abnormalities can result from repeated trauma or misuse of the voice, causing changes to the vocal fold tissue that damage normal vocal function. On the other hand, neurogenic causes originate from abnormalities in the central or peripheral nervous system. Spasmodic dysphonia (SD) is one such disorder which originates from the basal ganglia of the brain. This disorder affects the neural control of the larynx, causing spasms in speech [3]. These conditions are just a few out of the various causes and classes of voice disorders.

Diagnosing voice disorders presents a significant challenge for healthcare professionals due to the complexity and variety of potential causes. An effective, automated voice disorder detection system would significantly advance efforts to ensure the wellbeing of individuals at risk for voice disorders. Currently, existing systems, primarily for neurogenic voice disorders, are diagnosed with ineffective imaging techniques. For example, Ludlow et al. [4] show that SD is challenging to diagnose because the larynx often appears normal on standard imaging tests. Hence, in this paper, we introduce ConVox, a deep learning model that uses a sequential stack of one-dimensional convolutional neural networks (CNNs) to detect voice disorders through analyzing speech samples.

II. Literature Review

A. Mel-Frequency Cepstral Coefficient

Machine learning approaches for voice disorder classification have been studied for quite some time. Most models rely heavily on the robustness of their data processing, requiring discrete values from an originally continuous signal. Derived from the Short-Time Fourier Transform, the Mel-frequency cepstral coefficient (MFCC) is one such method [5]. The MFCC is mapped to the Mel scale, an approximation of how loud humans perceive pitch, and is represented as a magnitude. Peng et al. [6] followed a similar process, but utilized a logarithmic scale rather than a pure Mel scale. Verde et al. [7] mentioned several other features often used for classification: Fundamental Frequency (F_0), Jitter, Shimmer, Harmonic to Noise Ratio, and the first and second derivatives of the cepstral coefficients. As the researchers point out, F_0 has no exact formula but is only determined through approximations such as the Hilbert-Huang Transformation. Thus, the standard feature is the MFCC.

B. Deep Learning Approaches

Deep Learning has recently received notable attention for its versatility in image classification, text generation, speech recognition, and environmental modeling. Consequently, it has been applied to the voice disorder diagnosis problem as well. Chainani et al. [8] used deep learning as a supplement to their convolutional neural network and Long-Short Term Memory architecture. Combining a 1-D CNN, an LSTM layer, and SinRU activation function, they achieved an accuracy of 70.62% for classifying four voice disorders taken from the Saarbrücken Voice Database (SVD). Similarly, [6] used a deep transfer learning approach on the VOICED set. Their model consisted of a CNN and a Support Vector Machine Classifier. To decrease feature dimensionality, they employed linear local tangent space alignment. Their method achieved a sensitivity of 99.6%, specificity of 98.9%, accuracy of 98.5%, and F1 score of 99.6%. Finally, Fang et al. [9] pitted a deep neural network (DNN) against a support vector machine and a Gaussian mixture model. On the Massachusetts Eye and Ear Infirmary dataset, their DNN scored a higher accuracy, with a 99.32%, 94.26%, and 90.52%, demonstrating varying results among different datasets.

Joshy et al. [10] also made a DNN and CNN, but their goal was to instead classify the severity of dysarthria as opposed to identifying it. On the UA Speech dataset, their accuracy was 96.18%; on TORGO, their accuracy was 93.24%. The highest accuracy reached for severity classification was 99% in the study conducted by Joshy et al. [10], where the authors proposed a novel cross-modal network structure, combining audio and video recordings from UA Speech, extracting the MFCC, and inputting into their 2-D CNN. Despite the success in severity classification, the same cannot be said for dysarthria and voice disorder detection. The greatest was achieved by Verde et al. [7], who had an accuracy of 90% on the MEEI dataset. Even then, they could not replicate it on SVD and VOICED. Furthermore, attempts made by [12], [13], [14], [15], [16], [17], [18], and [19] achieved accuracies of only 57%, 89%, 86%, 79%, 62.87%, 71%, 82%, and 57.5%, respectively.

It is clear that there exists few models that can break the 90% mark, with a few exceptions ([20], [21], [22], [23]). As a result, we propose a deep learning framework that conducts binary classification to identify voice disorders. Rather than using only one or two datasets, we will use four large datasets containing audio recordings.

III. Methodology

Source code for the project can be found at:

https://github.com/sjain2025/Voice-Disorder-Classification

A. Data Acquisition

We utilized four large datasets to train ConVox: the Advanced Voice Function Assessment Databases (AVFAD), Saarbrücken Voice Database, TORGO Database, and the UA Speech dataset. Each of these databases consisted of voice recordings in Waveform Audio File Format (WAV) from individuals who were either healthy or diagnosed with a vocal pathology. The TORGO and UA Speech datasets compose of English speakers, whereas the SVD and AVFAD datasets are in German and Portuguese respectively. The sampling rates for the AVFAD, SVD, TORGO, and UA Speech datasets are 48, 50, 16, and 16 kHz respectively.

The AVFAD dataset consists of 709 individuals: 363 who are healthy, and 346 who have a voice disorder. Of the 8,648 uncompressed files, we used the 1,553 files which were in WAV format and in the range between 2 to 8 seconds. We also used all 6,602 audio files from the TORGO Database that fit these criteria. In addition, we initially extracted 1,494 total WAV files of sentences from SVD: 632 were recordings from healthy individuals and 862 were recordings from people with voice disorders. From this database, 732 fit the criteria and were used in the final concatenated dataset. Finally, the UA Speech Database consists of 75,364 voice recordings, from which we randomly selected 14,000 WAV files to train the model. Nonetheless, the model was tested on all 75,364 voice recordings from the UA Speech Database to ensure generalizability and prevent overfitting. By doing so, we maintained balance in the sizes of all four datasets, ensuring that ConVox was both robust and feasible to process. Altogether, the concatenated dataset consisted of 22,883 WAV files.

B. Preprocessing

We resampled all of the WAV files in each dataset to 16 kHz, and then filtered out all files that were less than 2 seconds and more than 8 seconds. We chose this range because training on excessively long files was unnecessary, as disorders could be accurately detected within the optimal 2 to 8 seconds of speech. We also omitted files shorter than 2 seconds because these audio files predominantly consisted of speakers making vowel sounds, without capturing complex sounds or sentences. The filtered and resampled files were then padded with zeros to make all sequences a uniform 8 seconds long at a sampling rate of 16 kHz. Using the librosa library, we extracted MFCCs from each of these files and used them for training the model. The MFCCs function on a perceptual scale, which also better captures the timbre of the sounds.

$$\text{Mel}(f) = 2595 \log \left(1 + \frac{f}{700}\right) \tag{1}$$

Using the formula in (1), a frequency that is in Hertz (f) can be transformed to the Mel scale. Files of healthy speakers were assigned a ‘0’, and files of pathological speakers were assigned a ‘1’. Then, we combined all of the eight second long sequences into a `tf.data.Dataset` object using the `from_tensor_slices` method. Furthermore, we implemented a batch size of 32.

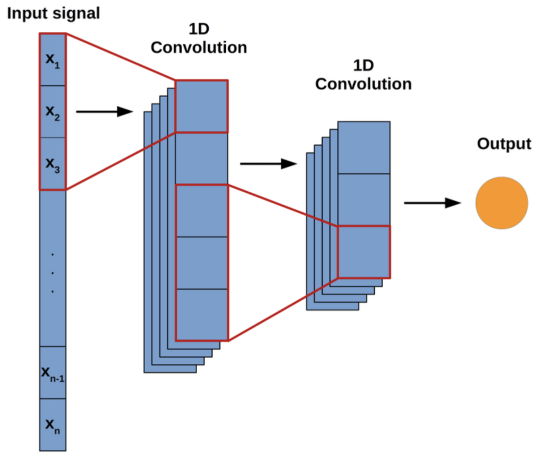


Fig. 1. 1-D CNN Structure [25]

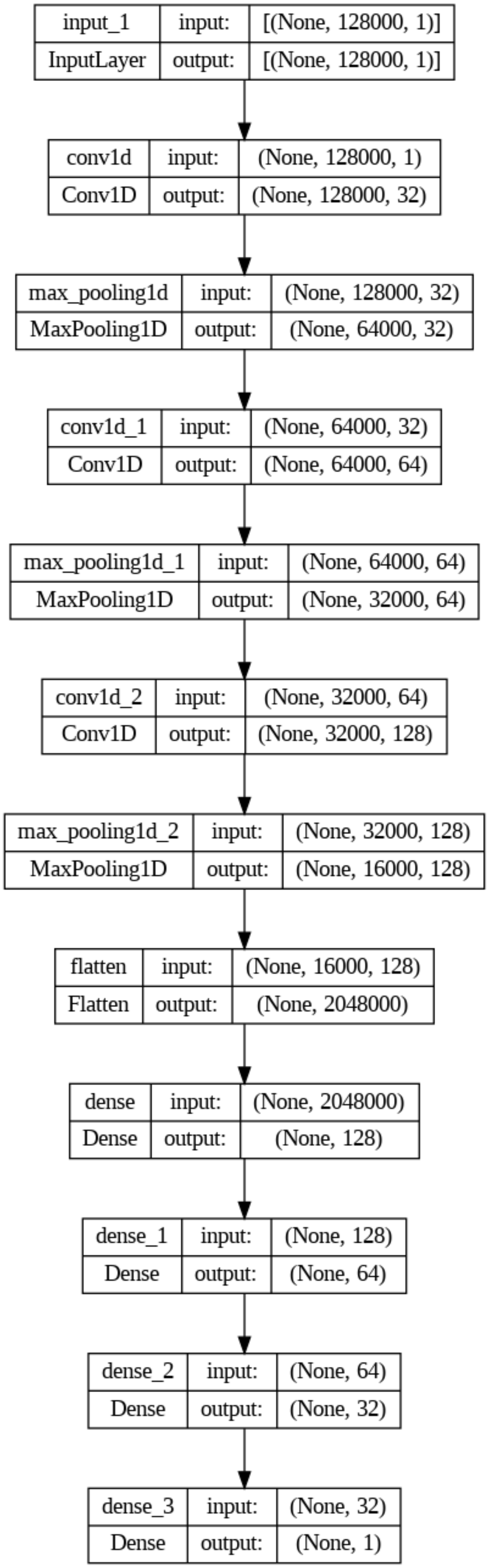
C. Model

The model that we constructed is a sequential neural network designed for binary classification, utilizing a combination of convolutional and dense layers to process one-dimensional input data. As illustrated in Fig. 1, the 1-D CNN structure begins with an input signal represented as a one-dimensional array of data points. This input signal undergoes multiple stages of convolution using 1-D convolutional layers. Each convolutional layer applies a set of filters to the input signal, extracting local features and producing a series of feature maps. The feature maps generated by the first convolutional layer are then fed into subsequent convolutional layers for further processing.

The summary of the model in Fig. 2 provides a detailed view of the layer configurations and parameter counts, reflecting its structure and complexity. The architecture features three convolutional layers (`Conv1D`), each followed by max-pooling (`MaxPooling1D`), which helps in extracting hierarchical features and reducing dimensionality. The convolutional layers use ReLU activation and ’same’ padding to introduce nonlinearity. Next, the model flattens the output and passes it through three dense layers with decreasing units and ReLU activation, before concluding with a final dense layer with a sigmoid activation function for binary classification. For evaluation, we tracked accuracy, area under the curve (AUC), loss, recall, and precision.

D. Training

Our dataset consists of 22,883 audio sequences, each eight seconds long, encompassing recordings from both pathological and healthy speakers. We applied a random split of 70% training data, 20% validation data, and 10% testing data. This



input_1	input:	[(None, 128000, 1)]
InputLayer	output:	[(None, 128000, 1)]
conv1d	input:	(None, 128000, 1)
Conv1D	output:	(None, 128000, 32)
max_pooling1d	input:	(None, 128000, 32)
MaxPooling1D	output:	(None, 64000, 32)
conv1d_1	input:	(None, 64000, 32)
Conv1D	output:	(None, 64000, 64)
max_pooling1d_1	input:	(None, 64000, 64)
MaxPooling1D	output:	(None, 32000, 64)
conv1d_2	input:	(None, 32000, 64)
Conv1D	output:	(None, 32000, 128)
max_pooling1d_2	input:	(None, 32000, 128)
MaxPooling1D	output:	(None, 16000, 128)
flatten	input:	(None, 16000, 128)
Flatten	output:	(None, 2048000)
dense	input:	(None, 2048000)
Dense	output:	(None, 128)
dense_1	input:	(None, 128)
Dense	output:	(None, 64)
dense_2	input:	(None, 64)
Dense	output:	(None, 32)
dense_3	input:	(None, 32)
Dense	output:	(None, 1)

Fig. 2. Model Summary

stratification ensures that ConVox was trained on a substantial portion of the data while also being rigorously validated and tested on unseen samples. To optimize the training process, we utilized an A100 Google Colaboratory GPU instance, known for its high-performance computing capabilities. The model training spanned 1048.96 seconds over the course of 20 epochs. During this period, we employed the Adam optimizer at a learning rate of 0.0001, which was optimal to balance the speed of convergence with the stability of the training process.

Binary cross-entropy loss, which the model was trained on, measures the performance of a classification model whose output is a probability value between 0 and 1. The loss for a single prediction is defined by the equation:

$$L_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}\left(Y_i \cdot \log \hat{Y}_i + (1-Y_i)\cdot \log(1-\hat{Y}_i)\right) \tag{2}$$

where:

- L_{BCE} is the Binary Cross-Entropy (BCE) loss.
- n is the number of samples.
- Y_i is the true label for the i-th sample. Y_i can take values of 0 or 1.
- $\hat{Y}_i$ is the predicted probability for the i-th sample that the label is 1. $\hat{Y}_i$ is a value between 0 and 1.

IV. Results and Discussion

We evaluated ConVox on five performance metrics: accuracy, recall, precision, loss, and AUC. Accuracy was given by the ratio of correctly predicted instances—both true positives (TP) and true negatives (TN)—to the total number of instances.

AUC is another performance metric for binary classification problems that we utilized to evaluate ConVox's performance. It represents the area under the Receiver Operating Characteristic curve, which plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The AUC value ranges from 0 to 1, where a value of 1 indicates a model with perfect accuracy and a value of 0.5 suggests a model with no discriminative ability (equivalent to random guessing). The equations for TPR, also known as recall, and FPR, also known as precision, are given as follows:

$$\text{Recall} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{Precision} = \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{4}$$

The AUC metric is particularly useful because it provides a single scalar value that summarizes the performance of the model across all possible classification thresholds. It is given by the following equation:

$$\text{AUC} = \int_0^1 \text{TPR}\ d(\text{FPR}) \tag{5}$$

The equation for AUC in discrete form is given as follows:

$$\text{AUC} = \sum_{i=1}^{n-1}(\text{FPR}_{i+1} - \text{FPR}i) \cdot \frac{(\text{TPR}_{i+1} + \text{TPR}_i)}{2} \tag{6}$$

TABLE I
Comparison of Dataset Performance

	Dataset				
	AVFAD	UA Speech	TORGO	SVD	TOTAL
ConVox	100.00%	99.72%	99.59%	99.86%	99.74%
[7]	—	—	—	80%	—
[14]	—	✓	✓	—	86%
[17]	—	—	—	71%	71%
[20]	✓	✓	—	—	93.36%
[21]	—	—	—	97.80%	—
[22]	—	—	97.73%	—	97.73%
[23]	—	—	—	93.90%	—
[24]	92.70%	—	—	90.90%	—

—	Dataset was not used or total accuracy was missing.
✓	Dataset was used, but accuracy was not provided for that dataset.

Table 1 shows ConVox's accuracy specific to each of the four datasets that we trained on. For the TORGO, Saarbrücken, and UA Speech Databases, the model was evaluated on all 6,602, 728, and 1,553 audio recordings, respectively. In addition, although the model was trained on 20% of the UA Speech Database, we evaluated it on all 75,364 audio recordings, finding an accuracy of 99.72%. ConVox achieved an accuracy of 99.59% on TORGO, 99.86% on Saarbrücken, and 100% on AVFAD. Other studies recorded accuracies in the range between 71% and 97.73%, lower than all of the recorded accuracies for our model.

Altogether, ConVox outperformed previous studies while being trained on more data than in prior research, making it more robust and accurate. Additionally, we trained our model on a corpus of data that contained samples of speakers in three different languages. Hence, through ConVox, we are moving towards language-agnostic voice disorder detection.

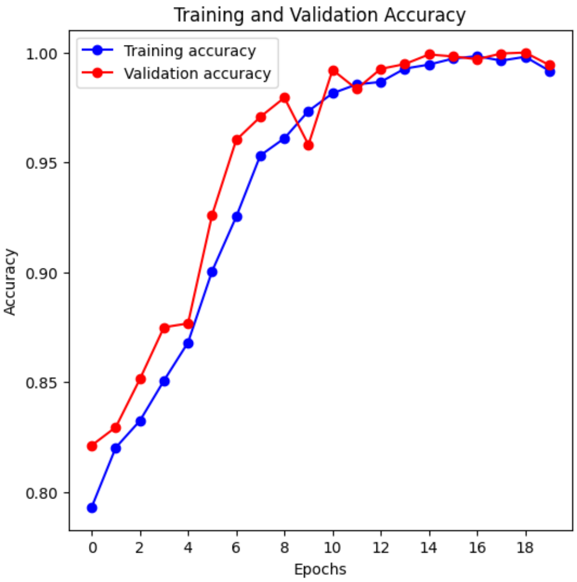


Fig. 3. Training and Validation Accuracy

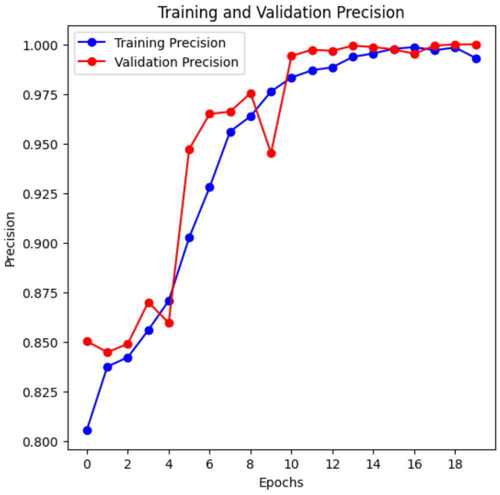


Fig. 4. Training and Validation Precision

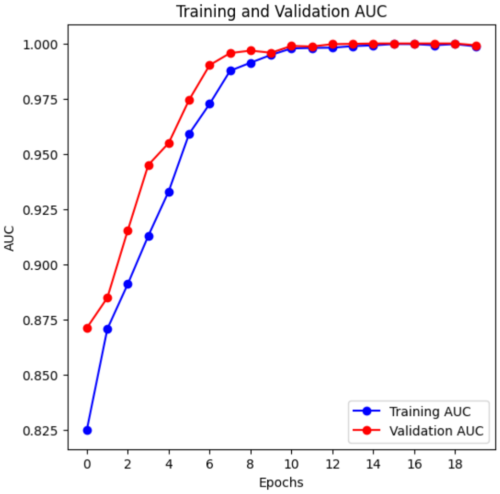


Fig. 5. Training and Validation AUC

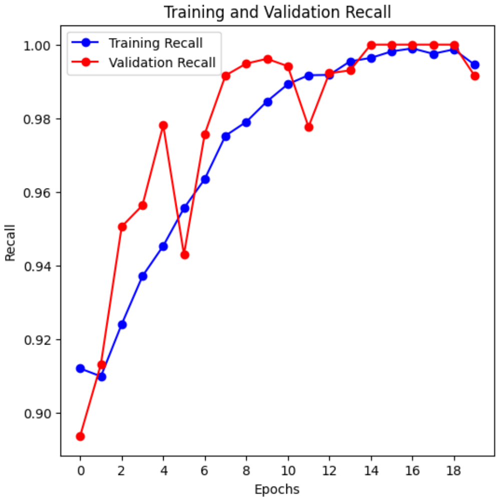


Fig. 6. Training and Validation Recall

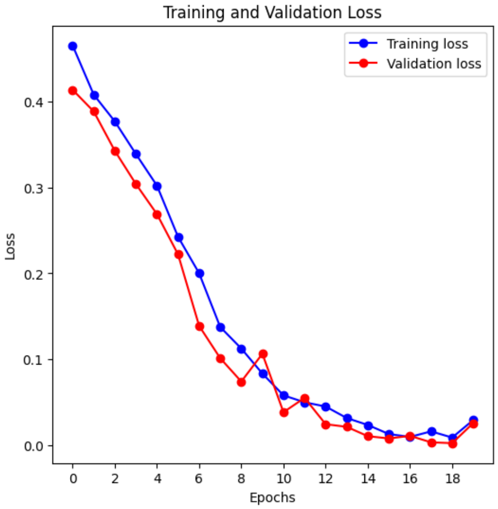


Fig. 7. Training and Validation Loss

Accuracy, AUC, precision, and recall all improved over the 20 epochs the model trained on, as shown in Fig. 3-6. The consistent improvement across these metrics in both the training and validation data indicates that the model not only fits to the training data well but also generalizes effectively to new, unseen data. This is crucial for ensuring that the model performs reliably in real-world applications, where it will encounter diverse voice samples beyond those seen during training. In addition, the binary cross-entropy loss decreased over the 20 epochs as illustrated in Fig. 7, suggesting that the model improves its ability to assign high probabilities to true positive samples and low probabilities to true negative samples. This reduction in loss reflects a more precise alignment between the model's predictions and the actual labels, enhancing the model's overall reliability and effectiveness in distinguishing between healthy and pathological voices.

V. Conclusion

Research on voice disorder classification as a tool to detect diseases has been the subject of scrutiny for quite some time now. With the advent of more robust machine learning architectures, it is more possible than ever to discern and diagnose based on a simple audio recording. In this study, we developed a novel sequential 1-D CNN that can effectively and accurately detect voice disorder through a short audio recording. Ultimately, we reached accuracies of above 99% in all the datasets, surpassing all existing attempts. Using four large datasets (AVFAD, SVD, TORGO, and UA Speech), we filtered for the MFCC as our feature being fed into ConVox and designed a pipeline that pre-processes the audio data, extracts features, and trains the model for high accuracy.

Despite promising results, we are still interested in other machine learning approaches such as traditional LSTMs to further enhance ConVox's performance. Even with our diversified dataset, future research should focus on expanding their data to include a wider range of voice samples and disorders, enhancing ConVox's robustness and generalizability. Additionally, it would be interesting to explore the integration of multimodal data, using not only voice analysis, but also other biometric indicators. We are currently focused on optimizing our architecture for both space and time efficiency. As we go forward, however, we will expand from binary classification to voice disorder categorization.

The implications of our results are profound. Early and accurate detection of voice disorders is imperative when it comes to treatment intervention and the patient's quality of life. Moreover, our machine learning model can be seamlessly integrated into Internet of Things technology, and more specifically, healthcare applications. We also seek to develop a user-friendly application for real-time voice disorder detection, providing a practical tool for clinicians and individuals alike to facilitate early diagnosis and intervention. Ultimately, ConVox provides a way for patients and at-risk individuals to find solidarity in knowing that their health can be monitored through an 8-second recording.

References

[1] Cleveland Clinic, “Voice Disorders: Types, Symptoms & Treatment,” Cleveland Clinic, Jun. 17, 2022. https://my.clevelandclinic.org/health/diseases/23339-voice-disorders

[2] ASHA, “Voice Disorders: Overview,” Asha.org, 2009, doi: https://www.asha.org/PRPSpecificTopic.aspx?folderid=8589942600.

[3] University of Michigan, “Spasmodic Dysphonia — University of Michigan Health,”, University of Michigan Health. https://www.uofmhealth.org/conditions-treatments/ear-nose-throat/spasmodic-dysphonia

[4] C. L. Ludlow et al., “Research priorities in spasmodic dysphonia,” Otolaryngology-Head and Neck Surgery, vol. 139, no. 4, pp. 495–505, Oct. 2008, doi: https://doi.org/10.1016/j.otohns.2008.05.624.

[5] D. Griffin and Jae Lim, “Signal estimation from modified short-time Fourier transform,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 2, pp. 236–243, Apr. 1984, doi: https://doi.org/10.1109/TASSP.1984.1164317.

[6] X. Peng, H. Xu, J. Liu, J. Wang, and C. He, “Voice disorder classification using convolutional neural network based on deep transfer learning,” vol. 13, no. 1, May 2023, doi: https://doi.org/10.1038/s41598-023-34461-9.

[7] L. Verde, N. Brancati, G. De Pietro, M. Frucci, and G. Sannino, “A deep learning approach for voice disorder detection for smart connected living environments,” ACM Transactions on Internet Technology, vol. 22, no. 1, pp. 1–16, Feb. 2022, doi: https://doi.org/10.1145/3433993.

[8] M. Chaiani, S. A. Selouani, M. Boudraa, and M. Sidi Yakoub, “Voice disorder classification using speech enhancement and deep learning models,” Biocybernetics and Biomedical Engineering, vol. 42, no. 2, pp. 463–480, Apr. 2022, doi: https://doi.org/10.1016/j.bbe.2022.03.002.

[9] S. Fang, Y. Tsao, M. Hsiao, J. Chen, Y. Lai, F. Lin, and C. Wang, ”Detection of pathological voice using cepstrum vectors: a deep learning approach,” Journal of Voice, vol. 33, no. 5, pp. 634-641, 2019, doi: https://doi.org/10.1016/j.jvoice.2018.02.003.

[10] A. A. Joshy and R. Rajan, ”Automated dysarthria severity classification using deep learning frameworks,” in 2020 28th European Signal Processing Conference (EUSIPCO), Amsterdam, Netherlands, 2021, pp. 116-120, doi: 10.23919/Eusipco47968.2020.9287741.

[11] T. Han, Hamid Sharifzadeh, and I. McLoughlin, “Automatic assessment of dysarthric severity level using audio-video cross-modal approach in deep learning,” Unitec Research Bank (Unitec Institute of Technology), Oct. 2020, doi: https://doi.org/10.21437/interspeech.2020-1997.

[12] I. Miliaresi, K. Poutos, and A. Pikrakis, “Combining acoustic features and medical data in deep learning networks for voice pathology classification,” in Proc. 28th Eur. Signal Process. Conf. (EUSIPCO), Jan. 2021, pp. 1190–1194.

[13] Y. Takashima, R. Takashima, T. Takiguchi, and Y. Ariki, “Dysarthric speech recognition based on deep metric learning,” Interspeech 2020, Oct. 2020, doi: https://doi.org/10.21437/interspeech.2020-2267.

[14] S. Venugopalan et al., “Speech intelligibility classifiers from 550k disordered speech samples.” Accessed: Jul. 19, 2024. [Online]. Available: https://arxiv.org/pdf/2303.07533.

[15] Olha Pronina and Olena Piatykop, “The recognition of speech defects using convolutional neural network,” CTE workshop proceedings, vol. 10, pp. 153–166, Mar. 2023, doi: https://doi.org/10.55056/cte.554.

[16] Z.-Y. Chuang, X.-T. Yu, J.-Y. Chen, Y.-T. Hsu, Z.-Z. Xu, C.-T. Wang,F.-C. Lin, and S.-H. Fang, “DNN-based approach to detect and classify pathological voice,” in Proc. IEEE Int. Conf. Big Data, Dec. 2018, pp. 5238–5241.

[17] H. Wu, J. Soraghan, A. Lowit, and Gaetano Di-Caterina, “A deep learning method for pathological voice detection using convolutional deep belief networks,” Strathprints: The University of Strathclyde institutional repository (University of Strathclyde), Sep. 2018, doi: https://doi.org/10.21437/interspeech.2018-1351.

[18] B. F. Zaidi, S. A. Selouani, M. Boudraa, and M. Sidi Yakoub, “Deep neural network architectures for dysarthric speech analysis and recognition,” Neural Computing and Applications, vol. 33, no. 15, pp. 9089–9108, Jan. 2021, doi: https://doi.org/10.1007/s00521-020-05672-2.

[19] D. Mulfari, G. Meoni, and L. Fanucci, “Machine learning in assistive technology: A solution for people with dysarthria,” in Proc. 4th EAI Int. Conf. Smart Objects Technol. Social Good, Nov. 2018, p. 308.

[20] D. Ribas, M. A. Pastor, A. Miguel, D. Martínez, A. Ortega and E. Lleida, ”Automatic voice disorder detection using self-supervised representations,” in IEEE Access, vol. 11, pp. 14915-14927, 2023, doi: 10.1109/ACCESS.2023.3243986.

[21] J. Reid, P. Parmar, T. Lund, D. Aalto, and C. Jeffery, “Development of a machine-learning based voice disorder screening tool,” American Journal of Otolaryngology, vol. 43, no. 2, 2022, doi: https://doi.org/10.1016/j.amjoto.2021.103327.

[22] S. Sekhar, G. Kashyap, A. Bhansali, A. A., and K. Singh, “Dysarthric-speech detection using transfer learning with convolutional neural networks,” ICT Express, vol. 8, no. 1, pp. 61-64, 2022, doi: https://doi.org/10.1016/j.icte.2021.07.004.

[23] M. Alhussein, G. Muhammad, “Voice pathology detection using deep learning on mobile healthcare framework,” in IEEE Access, vol. 6, pp. 41034-41041, 2018, doi: 10.1109/ACCESS.2018.2856238.

[24] A. Koudounas, G. Ciravegna, M. Fantini, G. Succo, E. Crosetti, T. Cerquitelli, and E. Baralis, “Voice disorder analysis: a transformer-based approach,” Audio and Speech Processing, 2024, doi: https://doi.org/10.48550/arXiv.2406.14693.

[25] A. Shenfield and M. Howarth, “A novel deep learning model for the detection and identification of rolling element-bearing faults,” Sensors (Basel), 2020, doi: 10.3390/s20185112. PMID: 32911771.