Automatic Voice pathology detection using Deep Learning Techniques

Robert-Valentin, Bencze[a], Coord. Asst. prof. PhD(c) Eng. Ana-Antonia Neacșu[a]

[a] *Politehnica University of Bucharest,*
Faculty of Electronics, Telecommunications and Information Technology,
Multimedia Technologies for audiovisual production and communications Master's programme,
Department of Telecommunications

Email addresses: benczejrobert@gmail.com *(*Robert-Valentin, Bencze*), (*Coord. Asst. prof. PhD(c) Eng. Ana-Antonia Neacșu*)*

Abstract

Pathologies that affect the vocal tract are known to alter the quality of the patients' speech in distinguishable or more subtle ways. This paper presents a non-invasive multi-class automatic vocal pathology detector that outperforms the mean accuracy of medical professionals by analyzing the less distinguishable features of pathological voices. The proposed system uses a deep learning algorithm capable to analyze multiple pathologies based on speech signals that consist in simple vowel utterances recorded in either audio or electroglottographic (EGG) format. The classifier obtained 86% and 75% accuracies for the audio and EGG separately, while their simultaneous analysis yielded 95% accuracy.

Keywords: *Machine Learning, Deep learning, Automatic Voice Pathology Detection, Multi-class Voice Pathology detection*

1. Introduction

1.1. Applicability, motivation and objective

The latest advancements in technology have driven massive changes and yielded multiple development opportunities in the healthcare system. Such opportunities include automatic diagnosis assistants that can aid medical professionals to faster identify symptoms and pathologies with improved accuracy. The pandemic of COVID-19 underlined the benefits of remote pathology diagnosis, hence the usefulness of an automatic pathology diagnosis system. Besides its uses in a pandemic context, an automatic voice pathology detection system can prove beneficial to automatically redirect patients to the corresponding medical speciality in medical call centers, thus reducing the queue waiting time.

Pathological voice can indicate a wide variety of underlying causes, sometimes on its own or alongside other symptoms. For instance, a dysphonic (hoarse-sounding) voice can be the result of multiple causes. Therefore arises the utility of integrating the proposed solution into a system composed of multiple classifiers that output a final diagnostic based on a cumulated analysis of the functions of multiple organs.

The motivation to implement an automatic voice pathology claassifier arised from its cost-effectiveness and the need of faster and more accurate diagnostics, doubled by the lack of research in the domain of multi-class voice pathology detection based on a multi-modal approach of the simultaneous analysis of multiple speech signal recording formats.

The main objective of this work is to outperform the average 71.4% accuracy of medical personnel. [1]

ISSN

1.2. Related work

The majority of research in vocal pathology detection was conducted on problems of binary classification, but significant progress is yet to be achieved in the field of multi-class pathology detection. This means that the diagnosis of the pathological recordings offers no information on the nature of the pathology, but rather only on its existence or absence. The best accuracies of the solutions proposed by the authors of [2], [3], [4] and [5] were 84.37%, 94.6%, 91.17% and 95% respectively.

The authors of [6] proposed a solution that achieved 94.4% accuracy based on an Artificial Neural Network that can distinguish multiple classes using audio and electromyographic (EMG) signals separately. However, their implementation did not treat case of the multi-modal approach. Moreover, the research paper did not specify how the training and test data were split, nor the method of feature extraction employed (i.e. if the features were extracted from subsections - known as windows - of the signals or from their entire duration).

2. Problem Formulation

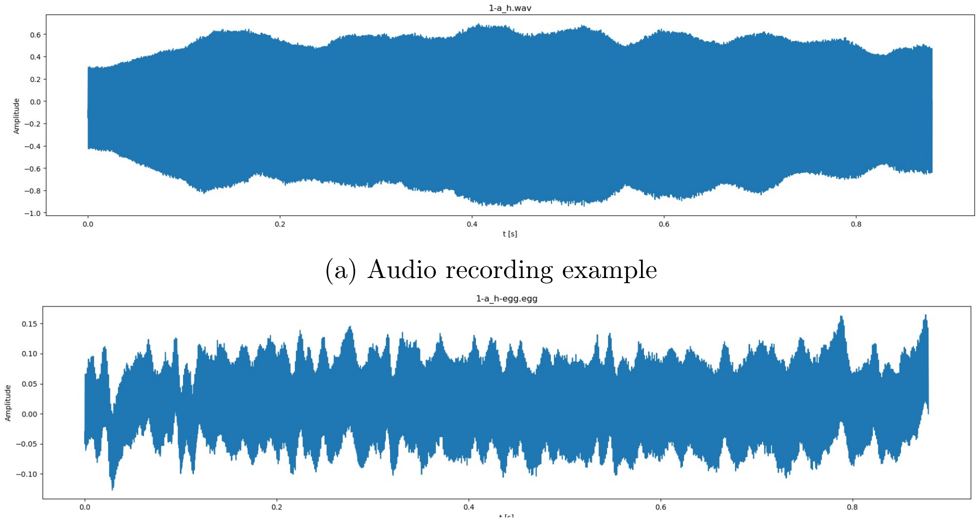
2.1. Theoretical aspects

The vocal tract is the biological structure that genreates speech and can be viewed as a sound source and a series of filters that can alter the resulting sound together or separately to achieve the desired phonemes, also known as the sounds that correspond to the letters.

Speech is an aperiodic phoneme sequence that results from rapid changes of the vocal tract, meaning that its corresponding signal can not be analyzed based on the assumptions that characterize stationary signals, which are generated by structures that do not change the properties of their filters. Every pathology that affects the vocal tract will also alter the speech signal in audible or non-audible ways that exhibit similar pathology-specific patterns. [7]

Speech can therefore be measured at different levels of the vocal tract and this research investigates the speech signals at the level of the vocal folds in electroglottographic (EGG) form and outside of the vocal tract in the familiar form of sound.

An EGG signal is recorded by an electroglottograph, a non-invasive device that measures the vocal fold vibration using two electrodes placed on the neck. [8]



(a) Audio recording example

(b) EGG recording example

Figure 1: Difference between the Audio and EGG recordings

2.2. The overall system architecture

The proposed system is a general-purpose flexible audio classifier that enables signal analysis based on multiple features that can be customized and enhanced depending on the domain of application.

The general structure of the system is presented in figure 2, where the input data is represented by audio or EGG recordings of speech that are split into partially overlapping windows. The windows are then saved to the storage of the computer that runs the experiment if there was no prior experiment, in order to minimize the processing time required for the next experiments. The signal are then processed by the feature extractors selected by the user who runs the experiment and are similarly saved to speed up further experiments. The next step in the pipeline is to split the saved features into training and test sets, where the test dataset will not be used by the neural network to learn and the data from the two sets are fed into the classifier.

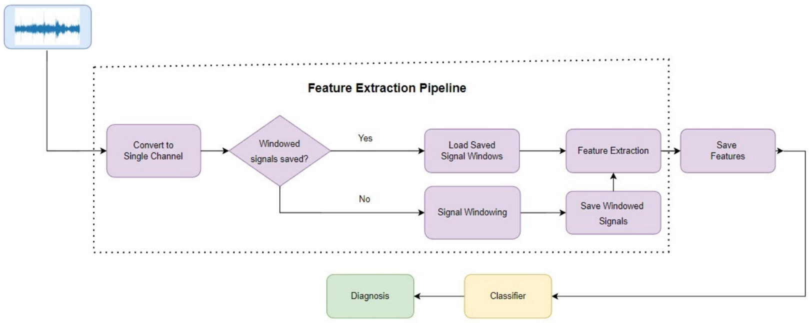


Figure 2: General overview of the system

2.3. Preprocessing and feature extraction

To obtain an accurate classification of the genres within the dataset, the classifier needs to be fed with relevant features extracted from the input signals or their windowed versions.

The features considered in this work are detailed below.

The Fourier Transform is a mathematical technique that can be used to decompose a signal into its pure-frequency sine wave components that represent the feature known as frequency spectrum of the signal. The Fourier Transform offers information about the global spectrum of the signal, meaning that it will not highlight frequency variations over time. The Inverse Fourier Transform is the mathematical technique that recovers the original signal from its frequency representation.

The Cepstral representation is an abstract representation of the "spectrum of a signal's spectrum" that is obtained by applying an Inverse Fourier Transform to the logarithm of the spectrum of a signal. This representation was proven useful in the context of audio analysis due to the logarithmic nature of the human hearing.

The spectrogram is a time-frequency representation that is used to analyze a signal's frequency variation over time. Essentially, a spectrogram is series of Fourier Transforms applied to subsequent time windows of a signal to capture the frequency spectrum variations of a signal.

The Mel-Frequency Cepstral Coefficients (MFCCs) are a another time-frequency representation of an abstract "spectrum of a spectrum" of the signal. The MFCCs are obtained by applying a Discrete Cosine Transform to the Inverse Fourier Transform of the logarithm of a Mel-scaled spectrogram calculated from the original signal.

The Mel Scale is a logarithmic frequency scale that more closely resembles the frequency behavior of human hearing.

The Discrete Cosine Transform is a mathematical technique that decomposes digital signals into cosine components, in a similar manner to the Fourier Transform, but with different properties.

2.4. Neural Network based classifier

A classifier is represented by an algorithm that is able to categorize input data as two (in the case of binary classification) or more (in the case of a multi-class problem) classes. Generally, a classification problem is solved (with limited accuracy) by a convoluted function with no known mathematical formula to create a general association between the input and the required output. Therefore we need a system capable of learning the function itself.

An example of such an algorithm is the Artificial Neural Network based classification. An Artificial Neural Network is a fully connected set of artificial neurons that are hierarhically organized in layers, meaning that each neuron in a given layer communicates with all the neurons of the previous and subsequent layer.

Each artificial neuron represents a computational unit which loosely models the behavior of its biological counterpart by a linear and a non-linear component. The linear component is the sum of all the inputs from the neurons of the previous layer or the direct inputs if the neuron is part of the input layer of the network. The non-linear part is represented by a non-linear mathematical function that is applied to the linear part of the neuron.

Such a classifier requires large amounts of data in order to correctly learn the function that maps the input to the desired output and therefore, more relevant data that contains new information will increase the performance of the classifier.

This is the main reason for the simultaneous analysis of both the audio and EGG signals, as the difference in figure 1 suggests.

The learning process of a neural network is represented by an optimization problem to find a minimum of a loss function that can be regarded as the penalty received by the network for a classification. Fundamentally, the loss function represents some kind of total difference between the output of the network and the expected value of its output.

An optimizer is an algorithm that modifies the parameters of the classifier with the aim of minimizing the penalty received by the neural network for all the classifications done in one epoch of the training process.

In the context of neural networks, an epoch represents a full pass through the training data examples.

3. Experimental Setup

3.1. Dataset

The Saarbrucken Voice Dataset (SVD) [9] is a set of speech signals recorded in audio or EGG format. The SVD consists of 869 healthy and 1356 subjects with one or multiple voice pathologies from the 71 voice disorders in the dataset. For the experiments presented in this work, an excerpt of three diagnostics was used: Healthy, Cysts and Ventricular Dysphonia. Each class was reduced to 54 recordings of 1 second each that correspond to 6 speakers with 9 utterances each, to avoid an unbalanced classification task, meaning that all the classes have an equal number of examples to facilitate a learning process that does not favor any of the diagnostics. The 9 utterances of a subject are the total utterances that result from the combination of either vowel and intonation detailed in the naming convention below.

The naming convention of each vocal recording respects the following pattern that enables the proposed system to perform an experiment on a filtered subset of desired vowels and intonations: $speaker\_no - VOW\_INT(-egg).extension$, where "-egg" is optional, $VOW \in \{'i', 'a', 'u'\}$, $INT \in \{'l', 'n', 'h'\}$, where 'h', 'n' and 'l' stand for high, normal and low respectively. Finally, extension can be either '.wav' or '.egg' for audio or EGG signals.

For instance, $2 - u\_h - egg.egg$ is the naming convention for the recording of the vowel 'u' that was pronounced with high intonation by the 2^{nd} subject and recorded in EGG format.

3.2. Audio Signal Analysis

The vocal recordings were divided into 20ms windows that partially overlap, as a means of data augmentation used to improve the learning of the neural network by essentially "teaching" it to recognize parts of the speech signals as speech signals and thus supplement the data used by the neural network to learn.

The duration of the windows was chosen 20 ms because, during speech, the shape of the vocal tract remains approximately the same during periods of 20 to 40 ms [10], meaning that the properties of the resulting signal will also negligibly change during those periods, facilitating its analysis.

Data augmentation and Windowing techniques

In the context of Deep Learning, signal windowing can be used as a means to artificially enlarge the number of training examples (known as data augmentation) for a neural network in order to achieve a higher accuracy and reduce the impact of real-life data variability. Another example of data augmentation is the addition of noise to the input signals.

Data augmentation is the concept of enhancing a dataset based on slightly modified versions of the data that it already contains, in order to diminish the effect of over-learning the trainig data (also known as overfitting). The most extreme example of overfitting is when a classifier obtains 100% accuracy on its training set but has 0% accuracy on the test set or on real-life data, meaning that its classification function is only useful for a very narrow set of examples.

Figure 3 depicts the most common windowing function types. The rectangular window applies no change to the signal's loudness in time, but heavily modifies the frequency spectrum of the analyzed signal due to its abrupt onset and ending. The Hamming and Hanning window types correspond to a smoother transition of the time-domain representation of the signal, by reducing the loudness of its onset and ending.

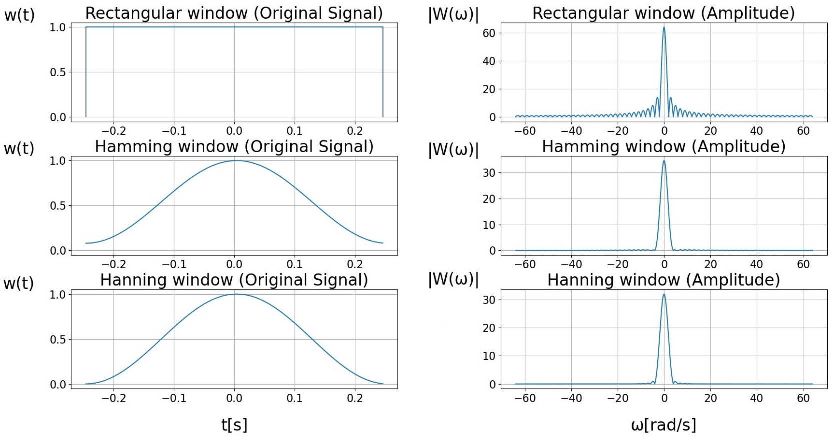


Figure 3: Commonly used windowing functions

Ideally, for no changes to the frequency spectrum, the central lobe of the windowing function's spectrum should be as narrow as possible and the ripples of non-zero frequency should also be as low as possible to minimize the impact of the windowing on the frequency spectrum of the initial signal. In reality, both constraints are not possible at the same time and there will be a trade off between the two constraints, depending on the application - as the above comparison indicates: the rectangular window has the narrowest central lobe but the most acute ripples, while the other two have a wider main lobe but milder ripples.

3.3. Proposed Structure of the Neural Netework

Figure 4 depicts the architecture of the neural network used for all the experiments, including the input and output layers. Every layer's activation function (except for the output layer) is the Rectified Linear Unit (or *ReLU* for short) that outputs the unaltered input if its value is positive or 0 otherwise. The non-linearity of the output layer was chosen to be the *Softmax* function that outputs a probability of the input signal being each pathology as the figure shows. The final classification of the network will be the most likely diagnosis, in this case "Healthy". *ReLU* was chosen due to its computational simplicity and because it facilitates a fast learning process.

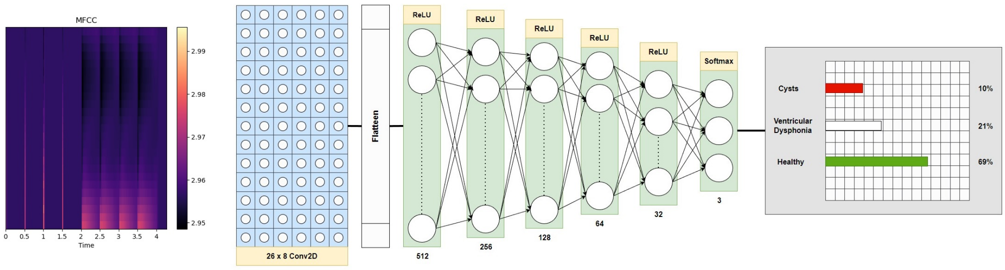


Figure 4: Proposed Architecture of the Network

The ADAM [11] optimizer was chosen for the training process for all the experiments due to its adaptive learning rate that yields the advantage of a faster learning at the cost of slightly decreased generalization capacities. Because of this downside, a 50% dropout rate was applied to each layer, meaning that for every learning epoch, approximately 50% neurons of each layer will be randomly ignored, thus reducing the network's overfitting.

The **categorical cross-entropy** was chosen as the loss function and is computed using the following formula:

$$CE = -\sum_{i=1}^{C} s_i log(\hat{s}_i), \tag{1}$$

where C is the number of classes, $\hat{s}_i$ is the score of the predicted class and s_i is the score of the i_{th} class.

The main metric used is the **accuracy** which can be described as the percentage of true positive predictions from the total number of predictions.

Figure 5 depicts the overview of the system that was used for the multi-modal approach. The *Clean Dataset* stages correspond to a functionality that allows dataset filtering based on the vowels and intonations of the patients. The datasets used for all the experiments were used entirely, without any filtering applied.

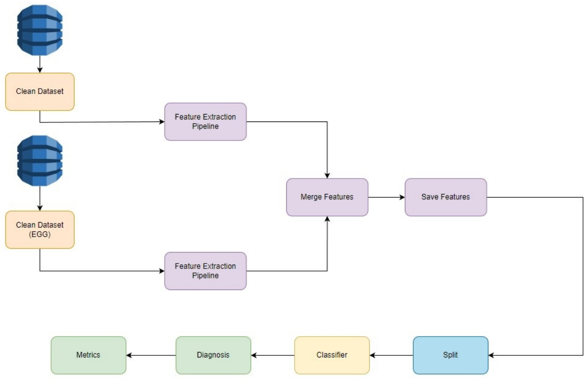


Figure 5: Multi-modal approach overview

The Audio, EGG and Multi-modal networks were trained for 130, 530 and 530 epochs respectively.

3.4. Performance Evaluation

Audio Format Dataset

Figures 6 and 7 depict the evolution of the training and validation losses and accuracies respectively for 130 epochs on the audio dataset:

The best model was saved at the epoch with number 117. The best accuracy for the Audio format dataset was 86%.

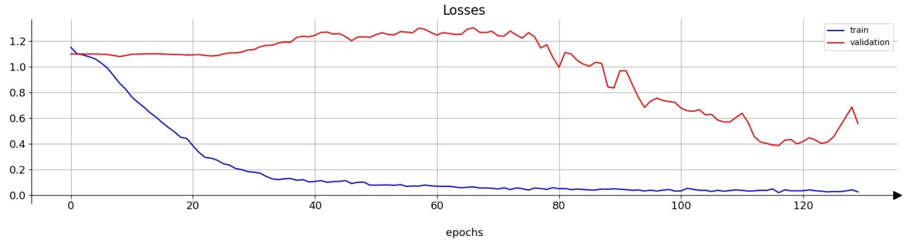


Figure 6: Loss for Audio Dataset

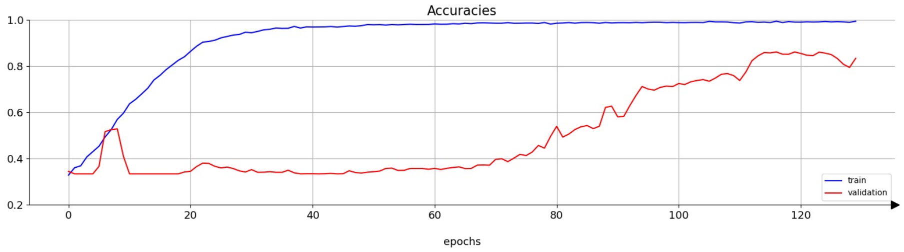


Figure 7: Accuracy evolution for Audio Dataset

EGG Format Dataset

Figures 8 and 9 illustrate the evolution of the accuracies and losses of the training and validation sets for a duration of 530 epochs on the EGG format dataset:

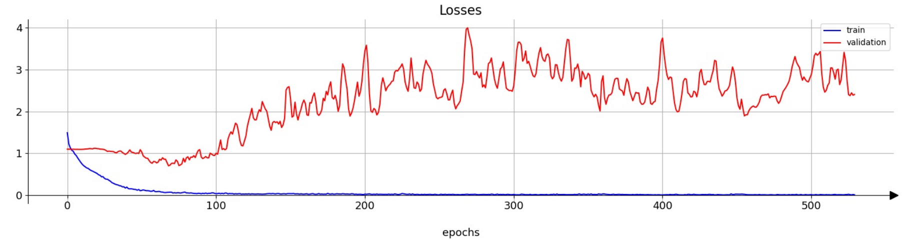


Figure 8: Loss for EGG Dataset

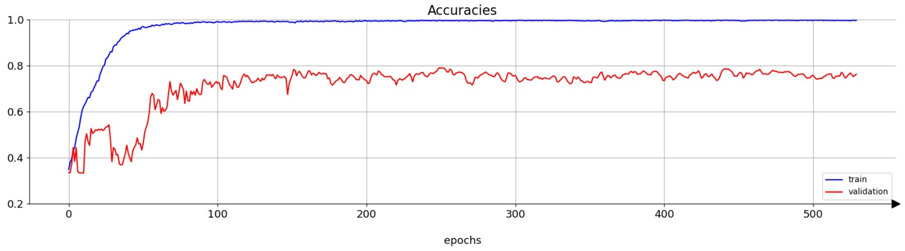


Figure 9: Accuracy for EGG Dataset

The best model was saved at epoch with number 77. The best accuracy for the EGG format dataset was 75%.

Multi-modal Dataset

The pipeline for the multi-modal dataset is shown in figure 5. The features were concatenated along the time axis, with the MFCCs extracted EGG signals after the ones that correspond to the Audio signals.

Figures 10 and 11 illustrate the evolution of the accuracies and losses of the training and validation sets for a duration of 530 epochs on the multi-modal dataset:

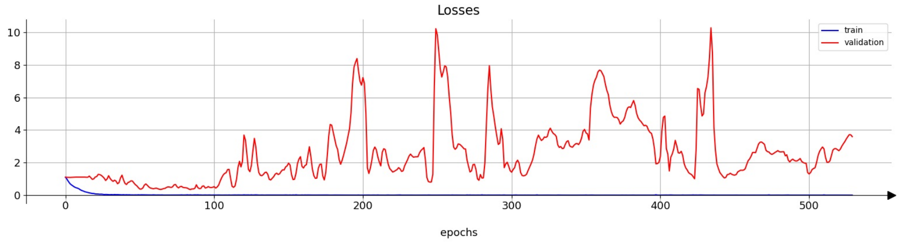


Figure 10: Loss for Multi-modal Dataset

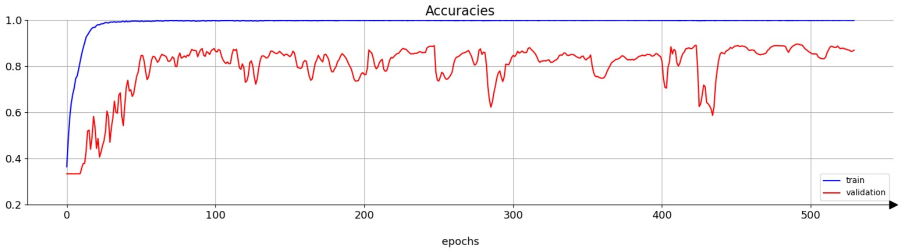


Figure 11: Accuracy for Multi-modal Dataset

The best model was saved at epoch with number 80. The best accuracy for the multi-modal dataset was 90%.

3.5. Real Life Experiment

The author tested the reliability of the proposed system by recording two vocal signals to run a real-life experiment. The first recording consisted of the utterance of vowel "a" with normal intonation with one's vocal folds, while the second recording was a simulation of a dysphonic voice by using the false vocal chords to utter a hoarse sounding "a" of normal intonation.

The results are shown in figure 12 where a confusion matrix shows that all the sub-windows of both recordings have been classified as healthy, meaning that the algorithm is robust to fraud. The folder Cysts contains the dysphonic vocal recording simulation as well. The confusion matrix is a tabular performance metric that displays the per-class accuracy of the classifier, as well as the wrongly associated labels with the input examples and the number of such misclassifications.

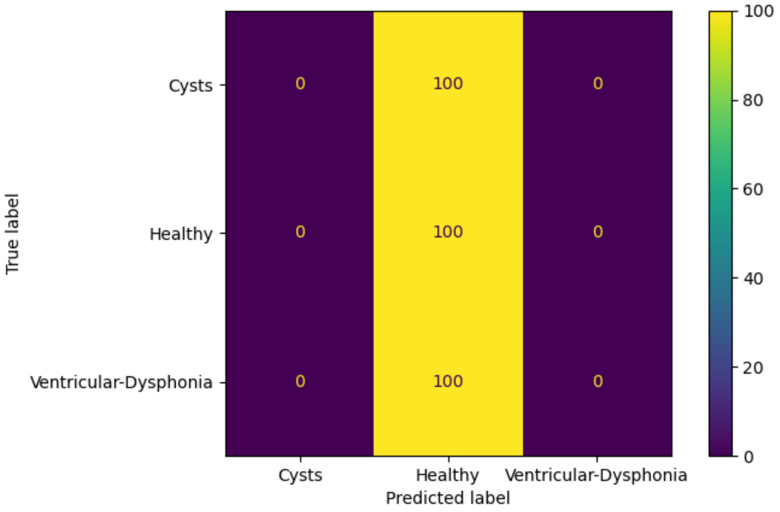


Figure 12: Live Experiment Results

This experiment can be run by downloading the code and following the instructions found at the address in the footnote of this page.

[1]

4. Conclusions and further development

The 75% accuracy obtained on the dataset by the system trained on recordings of EGG format, compared to the 86% accuracy of the audio dataset indicates that the use of other features could be explored to attain an improved performance than the performance obtained with Mel-Frequency Cepstral Coefficients. Thus, in order to increase the accuracy of the system on signals with different formats than audio, a method that uses differentiated features for each type of signal might prove beneficial. Moreover, the fact that the multi-modal dataset approach resulted in greater performance than either previous methods, separately, is an indicator that the two types of signals contain complementary information about the pathological nature of the recordings.

Additionally, the implemented algorithm might also prove its uses in applications such as non-invasive vocal recovery monitoring after surgery by measuring the difference between the neural network's current output and a healthy, target output example.

The advantages of the system include ease of use, as patients only have to speak, non-invasivity and objective diagnosis that is based on vocal parameters.

The obtained results suggest that further development is of interest and can be employed in several directions.

To further develop the system, data augmentation techniques could prove useful, such as noise addition or artificial example generation by Generative Adversarial Networks [12]. Noise addition could prove especially useful given that real-life applications imply audio recordings with consumer-grade microphones in various environments that likely differ from the equipments and environments used at the creation of the Saarbrucken Voice Dataset.

Another possible development direction would be the addition of features extracted from EMG signals to the multi-modal approach or to any of the EGG and Audio approaches to determine whether a new type of signal would help performance improvements.

[1] `https://drive.google.com/drive/folders/1fKaaZbTVSJtIjPSj_RUyvXSSTiTCv4JH?usp=share_link`

References

[1] Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, 11(1):1–9, 2020.

[2] Fahad Al-Dhief, Nurul Muazzah Abdul Latiff, Marina Mat Baki, Nik Noordini Nik Abd Malik, Naseer Sabri, and Musatafa Albadr. Voice pathology detection using support vector machine based on different number of voice signals. 11 2021.

[3] Meisam Khalil Arjmandi, Mohammd Pooyan, Hojat Mohammadnejad, and Mansour Vali. Voice disorders identification based on different feature reduction methodologies and support vector machine. In *2010 18th Iranian Conference on Electrical Engineering*, pages 45–49. IEEE, 2010.

[4] Fahad Taha Al-Dhief, Marina Mat Baki, Nurul Mu'azzah Abdul Latiff, Nik Noordini Nik Abd Malik, Naseer Sabri Salim, Musatafa Abbas Abbood Albader, Nor Muzlifah Mahyuddin, and Mazin Abed Mohammed. Voice pathology detection and classification by adopting online sequential extreme learning machine. *IEEE Access*, 9:77293–77306, 2021.

[5] Everthon Silva Fonseca, Rodrigo Capobianco Guido, Sylvio Barbon Junior, Henrique Dezani, Rodrigo Rosseto Gati, and Denis César Mosconi Pereira. Acoustic investigation of speech pathologies based on the discriminative paraconsistent machine (dpm). *Biomedical Signal Processing and Control*, 55:101615, 2020.

[6] Farika Putri, Wahyu Caesarendra, Elta Diah Pamanasari, Mochammad Ariyanto, and Joga D Setiawan. Parkinson disease detection based on voice and emg pattern classification method for indonesian case study. *JEMMME (Journal of Energy, Mechanical, Material, and Manufacturing Engineering)*, 3(2):87–98, 2018.

[7] Ben Maassen, Raymond Kent, and Hermann Peters. *Speech motor control: In normal and disordered speech*. Oxford University Press, 2007.

[8] ICspeech. Portable electroglottography system, https://icspeech.com/electroglottography.html, 2009.

[9] Bogdan Woldert-Jokisz. Saarbruecken voice database. -, 2007.

[10] Olaide Agbolade. Vowels and prosody contribution in neural network based voice conversion algorithm with noisy training data. *arXiv preprint arXiv:2003.04640*, 2020.

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*, 2018.