Toward assessment of human voice biomarkers of brain lesions through explainable deep learning

Benjamín Gutiérrez-Serafín[1], Javier Andreu-Perez[2], Humberto Pérez-Espinosa[1], Silke Paulmann[2], and Weiping Ding[3]

[1]*CICESE-UT3, México;* [2]*University of Essex, United Kingdom;* [2]*Nantong University, China. (e-mail: hperez@cicese.mx)*

Abstract

Lesions in the brain resulting from traumatic injuries or strokes can evolve into speech dysfunction in undiagnosed patients. Employing ML-based tools to analyze the prosody or articulatory phonetics of human speech could be advantageous for early screening of undetected brain injuries. Additionally, explaining the model's decision-making process can support predictions and take appropriate measures to improve patient voice quality. However, traditional ML methods relying on low-level descriptors (LLDs) may sacrifice detailed temporal dynamics and other speech characteristics. Interpreting these descriptors can also be challenging, requiring significant effort to understand feature relationships and suitable ranges. To address these limitations, this research paper introduces xDMFCCs, a method that identifies interpretive discriminatory acoustic biomarkers from a single speech utterance, providing local and global interpretations of deep learning models in speech applications. To validate this approach, it was implemented to interpret a Convolutional Neural Network (CNN) trained on Mel-frequency Cepstral Coefficients (MFCC) for the binary classification task to differentiate between patients from control vocalizations. The ConvNet achieved promising results with a 75% f-score (75% recall, 76% precision), comparable to conventional machine learning baselines. What sets xDMFCCs apart is its explanation through a 2D time-frequency representation that preserves the complete speech signal. This representation offers a more transparent explanation for differentiating between patients and healthy controls, enhancing interpretability. This advancement enables more detailed and compelling studies in speech acoustic traits of brain lesions. Furthermore, the findings have significant implications for developing low-cost and rapid diagnostics of unnoticed brain lesions.

Keywords: Intelligent audio analysis, acoustic features, traumatic brain injury, explainable machine learning.

1. Introduction

Traumatic brain injury (TBI) lesions is one of the major global causes of mortality and disability [1]. The first line of diagnostic is still and often expensive scanners [2] such as Magnetic Resonance Imaging (MRI) [3].

Simple measures denoting the rhythmic and intonation of the audio can be directly extracted from the speech signals, such as intensity (energy), gradient (pitch), and duration. Other standard audio features used in the analysis of the human voice are fundamental frequency, jitter, shimmer, and Harmonic to Noise Ratio (HRN). However, more conspicuous transformations of audio have denoted remarkable success in speech recognition tasks and other human audio analyses.

Mel-Frequency Ceptrals Coefficients (MFCCs) is a popular feature extraction method that was widely used in speech recognition systems [4]. MFCCS adjusts to the way humans perceive the loudness and frequency of sound, making it more suitable for speech analysis tasks than other general-purpose acoustic attributes. A key to its success is that it can represent the filter function of the human vocal apparatus, given that, in the source-filter model of speech, MFCC represents the filter part (vocal tract). The frequency response of the vocal tract is relatively smooth, whereas the source of voiced speech can be modeled as an impulse train. The result is that the vocal tract can be estimated by the spectral envelope of a speech segment.

The cepstrum concept is important for the MFCC calculation. The cepstrum of a signal is the result of calculating the Fourier transform of the spectrum of the signal studied on a logarithmic scale. The cepstrum gives us information on the rate of change of the different spectrum bands. According to the filter/source model of voice production, linguistic and paralinguistic information is contained in the vocal tract transfer function. In the cepstral domain, the influence of the vocal folds (source) and the vocal tract (filter) on a signal can be separated since low-frequency excitation and formant filtering of the vocal tract are located in different regions of the cepstral domain. To represent the human voice, the cepstrum is transformed using the melodic scale. The result of this transformation is the MFCC.

While several Cepstral-based alternative techniques for speech characterization exist, including CQCC (Constant Q Cepstral Coefficients), LFCC (Linear Frequency Cepstral Coefficients), CZT (Chirp Z-Transform based spectrum), and TECC (Teager Energy Cepstral Coefficients), no compelling evidence has emerged to establish the definitive superiority of any specific technique over the others. At best, research has shown that certain types of characterizations excel in particular tasks compared to others [5], [6], [7] [8]. The MFCC technique stands

Table 1: Subject characteristics

	Sex	Num.	Avg. age	Std. age
Patients	Male	8	47.86	8.41
	Female	8	48.29	14.69
Control	Male	9	46.67	15.26
	Female	7	46.2	16.41
Total		32	47.22	13.39

out as the most extensively tested and widely employed method for speech analysis. Considering its prevalence, it serves as an optimal foundation for developing an explanatory method that can be subsequently adapted to incorporate other characterization methods represented with spectrograms.

broposal of a frugal auto-assessment of patients' voice with brain lesion via a simplistic speech utterance only.

Evidence of the major relevancy of MFCC features for classifying patients with brain lesions versus healthy individuals.

A novel interpretative method, xDMFCCs, to discover acoustic biomarkers within MFCCs audio features.

The rest of the paper is organized as follows: Section 2 connects the presented work with past related work. Section 3 explains the data collection approach and the pre-analysis performed and describes xDMFCCs. Section 4 describes the benchmarks and results of applying the methods described in the previous section to the prosody data collected for brain disorders; Section 5 discusses clinical aspects and revises the impact of this work; finally, Section 6 concludes.

2. Related work

In the study of brain injuries, different machine learning techniques have been applied to analyze different sources of information and build models that automate the diagnosis and estimate the severity of the damage caused to people's motor skills. A well-studied source of information is Magnetic Resonance Imaging (MRI) [9, 10] with successful results. Sometimes they have considered interpreting the MRI images [11]. MRI can be considered a gold-standard diagnostic; however, it is an expensive modality of diagnosis that requires expensive specialized equipment, facilities, and trained personal

One cognitive function often affected by brain injuries is speech. Continuous monitoring of patients with this type of injury requires long recovery periods with events that lead to rehospitalization. In Ditthapron et al. [12], a study using full speech from patients with traumatic brain injury (TBI) caused by car accidents, falls, and runovers. The main objective of the mentioned work is to solve the problem of the scarcity of limited TBI voice data, exploring three Limited Learning Data (LLD) methods (transfer, multitasking, and meta-learning). The study demonstrated a considerable improvement in classification assessment metrics for patients with and without brain injury. Nevertheless, several longitudinal time points were considered, and full speech. Another study also considered a similar problem, but in this case, the approach was multimodal speech and gait [13].

Several works have considered using an end-to-end convolutional neural network (CNNs) that takes as input the transformation of audio signals into a spectrum [14, 15, 16, 17]. Several works have identified that the Mel Frequency Cepstral Coefficients or MFCCs contain essential information to identify people with a certain degree of depression [18, 19], even relevant coefficients, but not the time of speech when they are more relevant. However, most of these studies are applied to a population of patients with ongoing mental health problems. Still, little is known about recognition before this condition is triggered, for instance, due to a prior TMI or a strong concussion. Also, from the methodological standpoint, the computational prowess of CNNs and MFCCs have not been proposed combined in this type of application.

As discussed, although much effort has been dedicated to recognizing dysfunctional brain issues from speech cues, the noted algorithms suffer from the following limitations and challenges:

(a) Most work focused on detecting depression when diagnosed, but none have target recognition of brain lesions leading to depression, paving the way to early intervention.

(b) Most of the considered inputs correspond to full speech or multi-modal experiments, but little is known about using just short speech utterances for a more straightforward diagnosis.

(c) Explainability of the models is often overlooked or limited to mentioning a set of relevant features, but the biomarkers within are not determined.

Understanding and interpreting acoustic features in speech is a relevant issue for clinical practitioners.

For example, in a study aimed to detect unilateral vocal fold paralysis (UVFP) from voice recordings [20], the authors enabled the comparison of acoustic features, using a custom algorithm called Independence Factor to select a single feature from sets with similar information. Using this approach, they demonstrated the importance of checking for biases using explainable machine learning and clinician perceptual ratings. In another study [21] the authors applied machine learning methods to distinguish between two prevalent vocal pathologies, vocal cord polyp, and vocal cord paralysis. Acoustic and spectral features were extracted and various classifiers were compared using batched cross-validation. Explainable AI and feature interpretability analysis were conducted to identify important features for clinical care and planning. Octave-based spectral contrast and MFCCs 0 to 3 were identified as the most significant features. A convolutional neural network (CNN) was used to learn low-level speech descriptors for vowel classification [22]. The modified Local Interpretable Model-agnostic Explanations (LIME) method was employed to assess the impact of spectrotemporal vowel variation on decisions and observe temporal changes in depression likelihood. Using this analysis, they found that vowel-based information is more important than non-vowel segments in identifying depression. Therefore, the

findings obtained through explainable machine learning contribute to the development of interpretable decision-support systems for mental health diagnosis and care by enabling clinicians to better understand fine-grained temporal changes in speech data. In this paper, we worked on explaining the attributes and models learned from them, with which doctors can generate greater understanding and new knowledge.

3. Methods

3.1. Experiment and Collected Data

Data for this study come from a previous unpublished investigation that tested 32 individuals, of which 16 were patients with localized brain lesions in regions previously associated with perception deficits [23, 24]. The Max-Planck Institute approved the study for Human Cognitive and Brain Science Ethical Review Board. Following the guidelines of the Ethics Declaration of Helsinki, patient consent was sought before data collection began. Speech data were recorded from patients with lesions in the basal ganglia (n = 8), orbitofrontal cortex (n = 5), and anterior temporal lobe (n=3). These dysfunctions may have been caused for several reasons, including stroke, intracerebral bleeding, or traumatic brain injury. These three lesion sites were of particular interest given that the brain regions affected have been reported to form part of a cognitive paralinguistic network [25].

For all patients and controls with the same age and education, prosody production data were recorded as part of a spontaneous production task (patients were instructed to say the word "Anna" in a happy, sad or neutral way) or a modeled expression task (patients were instructed to repeat the word "Anna" after it had been spoken by a trained actor in a happy, sad, or neutral way). Each set of "Anna" was repeated three times (with 6 Anna repetitions). Thus, each participant has a set of 18 files per psychological condition (18 happy posed expressions and 18 happy, spontaneous expressions; same for sad and neutral). The data was recorded using a high-quality microphone, saved on a portable DAT recorder, and later digitized on a PC. For complete details on the experimental procedure applied to these data, see [26], who used the same paradigm but tested a different patient group. Table 2 outlines the background characteristics of our sample.

3.2. Proposed xDMFCCs method

The following few sections outline our approach: first, the extraction of acoustic features from the raw sound recordings; second, the methodology that was used to select the most indicative feature; third, a method to explain the biomarkers found in the sound voice to discriminate between patients and control. A sketch of the processing steps of this analysis is presented in Figure 1.

3.2.1. Acoustic feature extraction

Acoustic features were automatically extracted from full audio samples using the open-source tool Praat [27]. Praat is an open-source software suite for speech analysis mostly used for scientific research. A Praat script was programmed to iterate through all audio recordings and obtain acoustic information. A set of 72 low-level acoustic descriptors were extracted. This set of acoustic attributes included various aspects of the voice that, according to the literature, may be relevant for detecting markers related to acquired neurogenic voice disorders. This initial set of attributes included 17 prosodic attributes, 21 voice quality attributes, 17 articulatory attributes, and 17 spectral attributes. Prosodic attributes included nine statistics obtained from the pitch contour extracted using the autocorrelation method and eight statistics related to the intensity contour in dB. Voice quality attributes included three pulse-related attributes (fraction of locally unvoiced frames, number of voice breaks, and degree of voice breaks), five jitter-related statistics, six shimmer-related statistics, seven measures related to harmonicity and its relationship with noise. Articulatory attributes included 12 statistics related to formants 1 to 4, five attributes related to LTAS (Long-Term Average Spectrum) of the spectral envelope of the voiced parts. Finally, the spectral attributes included five attributes related to statistical measures of the spectrum (skewness, kurtosis, standard deviation, centroid), 12 related to spectral balances obtained through frequency band energy differences and ratios.

Additionally, MFCCs were also contemplated for this study. To compute this set of features, the Python package Librosa was employed. Our script was programmed to extract MFCCs in a sampling rate of 22050 Hz, 12 Mel filters to extract 12 MFCCs coefficients, an FFT window length of 2048, and a hop length of 512. To improve performance, both first-order and second-order derivatives were also contemplated. Afterward, descriptive statistics (mean, std, min, max, as well as 25, 50, and 75 percentiles) of the MFCCs, their delta, and the double delta of each sample were calculated. Overall, this method's whole feature extraction process adds up to 323 features.

3.2.2. Feature importance evaluation for brain-injury patient recognition

For the feature evaluation, we employed non-linear *Predictor Importance estimation via Random Forest*. A random forest (RF) classifier was applied to tackle the binary classification problem and discriminate between patient and control audio samples. RF permits analyzing the non-linear multivariate association among features that permits the observation classification. This is performed by calculating the Gini index, which measures the impurity level at each node split of the three [28]. The more the Gini index decrease at each node split accounting for a specific feature, the more important this later feature is.

The Random Forest classifier was trained with acoustic features extracted with Praat. For our data splitting strategy, we emphasized avoiding filtering information from our training set to our validation set. For this reason, we adopt a leave-one-speaker-out scheme with the aim of training speaker-independent models. Moreover, speech samples spoken by males were segregated from female speaker samples given by the significant difference between the feature domain of speech signals. Taking this into account, optimal parameters for the female and male speaker-independent models were found using

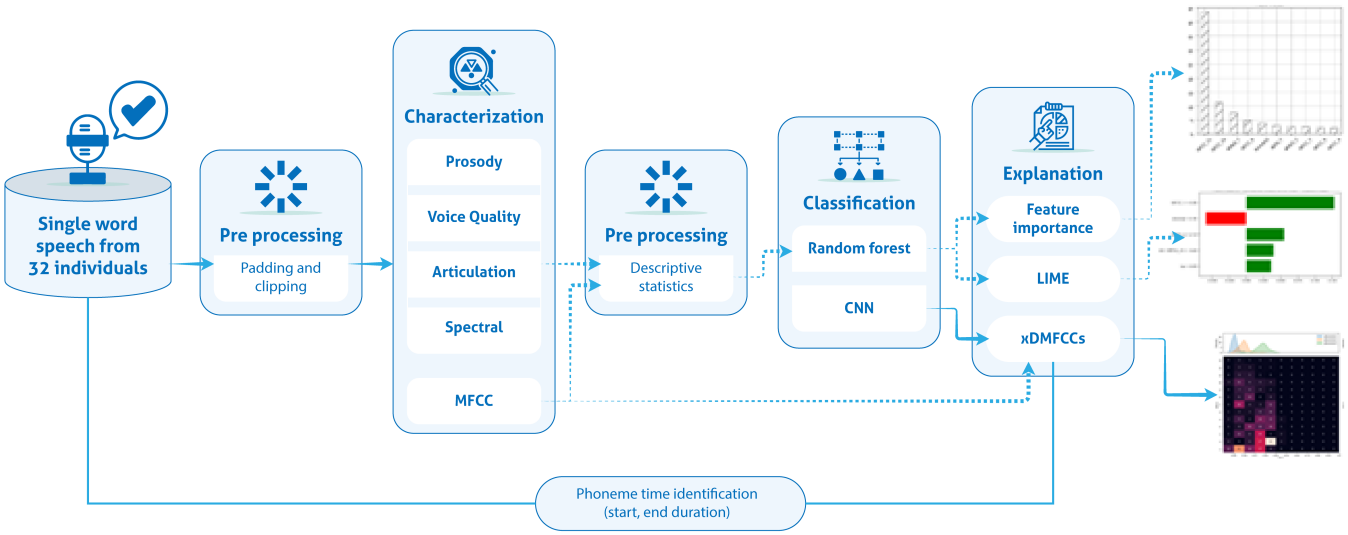

Figure 1: Ordered schema of the computational framework pipeline of the presented work.

grid search, providing fair results in both sets.

3.2.3. *xDMFCCs: Explaining Neural-Network reasoning on MFCCs*

To provide a more conclusive explanation, we proposed a global interpretable approach based on MFCCs and Convolutional Neural Networks to expand LIME image explainer capabilities [29]. Since the original LIME implementation is local in scope and supports the explanation of models with text, image, and tabular data, it is required to adapt the algorithm to work with audio data and provide a global explanation of the deep learning model for our speech application. Our method draws inspiration from the work proposed by Mishra et al. [30], which presented 3 different techniques to explain the decision process of algorithms trained to solve music content analysis tasks.

This sophisticated technique highlights the most significant regions based on their time and frequency locations that further determine the class to which an audio sample belongs. Consequently, a less intuitive explanation of the range of features has now been transformed into a more visual interpretation.

To begin with, phoneme start and end times are essential to relate the temporal component of the explanatory region to the emitted utterance. We took advantage of EasyAlign [31] plugin features to detect when a speaker starts and finishes producing each phonetic sound of the name ”Anna”.

As an additional pre-processing step, a homogeneous time duration is required. Therefore, audio samples longer than 1 second were clipped, while shorter samples were zero padded on the right side (see Figure 2). Then, the MFCCs transform was conducted with the same parameters as the Random Forest approach but added an additional component to the analysis. This combination of parameters, coupled with the uniform time duration, returns for each audio sample a two-dimensional array with the shape of 13 MFCCs and 44 frames.

As for the deep learning model, the architecture consists of three convolutional blocks formed by a convolutional layer,

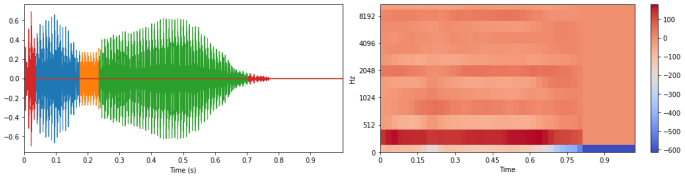

Figure 2: Waveform and time-frequency representation of sample audio. The utterance of each phoneme is highlighted in a different color.

max-pooling layer, and batch normalization, followed by two fully connected and drop-out layers. More exactly, the first block consists of a convolutional layer with 32 filters, a kernel size of (2x2), a relu activation function with zero padding, and a stride of (1x1) to keep the same size. It continues with a max pooling layer with both pool size and strides of (2x2) with no padding, followed by a batch normalization layer. For the next block, we doubled the convolutional layer filters and let the remaining components be similar to the first block. Next, we define the third block as the first one, but the convolutional layer without any padding. We then flatten the output of the convolutional base to train two fully connected layers with the relu activation function. The first one with 256 nodes, while 64 nodes have the latter. Each layer is followed by a drop-out layer with rates of 0.5 and 0.4, respectively. Finally, the top layer was declared with 2 nodes and a softmax activation function. The features were normalized and fed into the network in batches of 32 samples, and the model was trained with a learning rate of 1e-4. This model's results were comparable to those obtained with a Random Forest classifier.

Once the black-box model has been trained, we can implement our xDMFCCs method to obtain a global explanation for the CNN model, which relies on LIME for local interpretability in the time-frequency domain on a subset of speech samples. While we will provide a broad overview of local interpretability here, we encourage readers to refer to the LIME article for a

more detailed and comprehensive understanding.

For each correctly classified sample $x \in \mathbb{R}^{m*n}$ in our test set X_i^{tf}, a perturbed data set Z is generated. The appropriate selection of the number of samples to perturb, N, varies among the application and data dimensionality. Therefore, one should consider the trade-off between speed and maintaining explainability. In our case, we must get an explanation of every correctly classified sample in our test set, which could be time consuming if we select a large N value. Having experimented with different values of N, we concluded that the setting $N = 1430$ provided speed and consistent explanations for this specific task. First, a segmentation function is applied to divide the frequency and time axes of the original sample into 13 and 11 regular segments, respectively. For further reference, we denote the number of rows and columns (13, 11) in the frequency and time domain as (m, n). As a result, an explanatory region consists of 1 cepstral coefficient and 90 milliseconds in duration. These superpixel regions are treated as binary features. By means of these features, an interpretable representation for an image can be denoted as $x' \in \{0, 1\}^{d'}$. Thus, N perturbed samples $z' \in \{0, 1\}^{d'}$ are generated by randomly turning on and off superpixels of x' uniformly. This step of LIME is denoted *sample_around* and aims to build an interpretable model with a synthetic data set in the vicinity of x to approximate the black-box model.

Then, the black-box model f is used to predict these perturbed samples' target variable $f(z)$. Next, weights are computed to measure the importance of each perturbation based on their proximity with the original sample given by the kernel function $\pi_z = exp(D(x, z)^2/\sigma^2)$, where D is the cosine distance function and σ set as 0.25. Subsequently, LIME fits a weighted *Ridge* regression model g with the data obtained in previous steps to get the K most important components of the audio sample, where $K = 5$. At this stage, LIME takes advantage of *Ridge* regression as a surrogate model to understand which variables are the most important for a given sample by contrasting the coefficients of the binary features.

For a global understanding of the model, we define significant superpixels as ssp. We use P to represent the patient class and C to describe the control class. Likewise, we denote the representative counts for the patient class as *CRP*, while *CRC* corresponds to the representative counts for the control class. We collect all ssp from each explained instance and group them accordingly into P or C based on their class. Afterwards, the counts of each ssp in P and C are stored in one-dimensional arrays, namely *CRP* and *CRC*, respectively, with a shape of (1, m*n). Next, we reshape the *CRP* and *CRC* arrays into a 2D time-frequency representation with dimensions (m, n). This transformation allows us to obtain a global interpretation of the most representative biomarkers for the patient and control classes.

Additionally, the phoneme time distribution was situated above each chart to facilitate the temporal localization of the utterance produced. If these 2D matrices of counts are visualized as heat maps, lighter regions symbolize a significant cepstral component at a specific point in time-related to a particular phoneme for most of the correctly classified samples for each class. On the other hand, darker areas are not significant for each category. To put it briefly, Algorithm 1 summarises the steps taken to obtain the global explanation of the model. Overall, the collective significance of each cepstral component and each temporal bin can be obtained by normalizing the count values of the explanation provided by xDMFCC to range from 0 to 1 and summing all elements over the axis.

Algorithm 1 Global interpretation using xDMFCCs

$P \leftarrow \{\}$
$C \leftarrow \{\}$
for all $x_i \in X, \ldots$ **do** ▷ Get local interp. using LIME
  $Z \leftarrow \{\}$
  for $i \in \{1, 2, 3..., N\}$ **do**
    $z_i' \leftarrow sample\_around(x_i')$ ▷ Synthetic sample
    $Z \leftarrow Z \cup \langle z_i', f(z_i), \pi_x(z_i) \rangle$
  end for
  $ssp \leftarrow Ridge(Z, K)$ ▷ Get Top5 binary features
  if $f(x_i) == patient$ **then** ▷ Save ssp based on class
    $P \leftarrow P \cup \{ssp\}$
  else
    $C \leftarrow C \cup \{ssp\}$
  end if
end for
$CRP \leftarrow \{\}$
$CRC \leftarrow \{\}$
for $j \in \{1, 2, 3..., d'\}$ **do** ▷ Get counts of all regions
  $CRP \leftarrow CRP \cup \textbf{sum}(P == j)$
  $CRC \leftarrow CRC \cup \textbf{sum}(C == j)$
end for
$CRP \leftarrow \textbf{reshape}(CRP, m, n)$ ▷ Reshape 1d to 2d array
$CRC \leftarrow \textbf{reshape}(CRC, m, n)$
return CRP, CRC

4. Results

In the next subsections, we will explain the resultant outcome from the analysis and the interpretation of the potential bio-markers as revealed by our proposed deep learning global explainer, xDMFCCs.

4.1. Feature analysis

Following the procedure explained in section 3.2.2, we took the set of most relevant features estimated by the estimation of the importance of random forest (Figure 3). Additionally, we employed Random Forest in conjunction with LIME which relies on local surrogate models to shed light on the rationale behind single predictions (Figure 4). Both feature approaches agree that MFCC related features are of crucial significance for this classification task. Therefore, we devoted greater efforts to find even more insights from these descriptors. From the distribution comparison in Figure 5 it can be seen that from a univariate perspective (diagonal), it is not possible to determine a suitable classification delimiter. However, in the multivariate

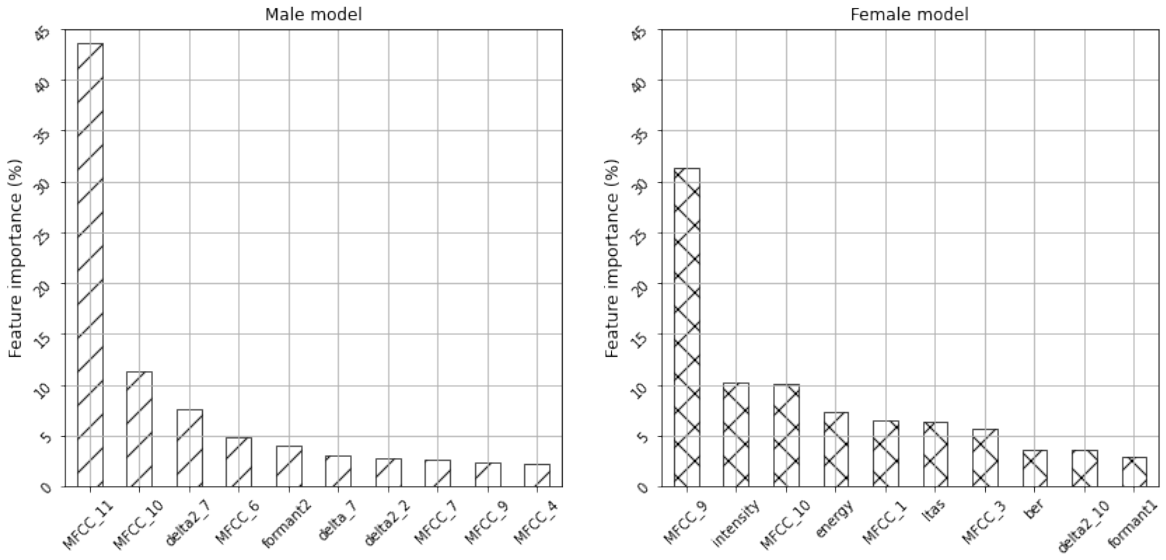

Figure 3: Top 10 most relevant features of both male and female Random Forest classifiers.

combination of several features, it was possible to determine some classification boundaries with the naked eye.

In particular, MFCC_1 demonstrated higher discriminability when combined with other features, while MFCC_10 provided a lesser degree of discriminatory power. The naming convention of *MFCC*, which stands for Mel Frequency Cepstral Coefficients, follows a numbering scheme where the coefficients are indexed from 1 onward, starting with the zero-order coefficient. Therefore, MFCC_1 represents the 0th coefficient, and MFCC_10 corresponds to the 9th coefficient.

Likewise, using RF as a baseline method, the LIME approach showed that MFCCs provided a high discriminative power. In light of the evidence from this analysis, we agree that MFCCs can encode prototypical discriminative patterns. Nevertheless, MFCCs together form a time-based spectrogram, denoted Mel-frequency cepstrum, which needs to be studied with spatial correspondence according to both dimensions: 1) cepstral component; 2) time point. For this endeavor, we subsequently used the proposed xDMFCCs method.

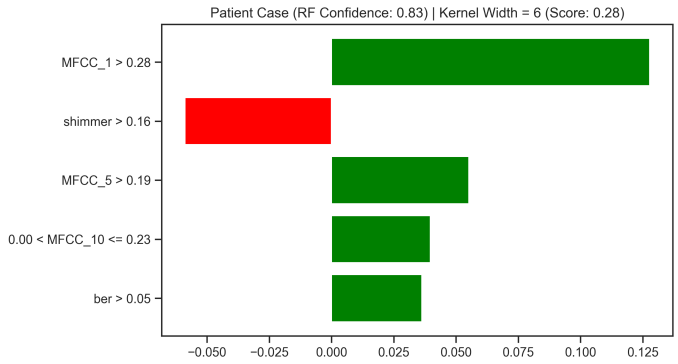

Figure 4: LIME interpretation of Random Forest classifier that predicted a sample as patient with confidence of 83%.

4.2. Classification

As a comparison baseline, we employed both Random Forest and Support Vector Machine classifiers. The classification accuracy reported in Table 2 shows a competing f score of 0.75-0.74 (max 1) in recognizing brain injury patients from healthy controls. This is encouraging, given the simplicity of the input given to the model. Specifically, we only provide an audio signal corresponding to the speech of three phonemes. Some classifiers are trained with male and female data independently, as recommended in speech analysis works [32]. Within the analysis, we found that recognition performance was similar in both, with marginally better recognition performance in females. Models based on deep learning can transfer knowledge learned from one sex to the other. Therefore, for training the CNN, we train with males and subsequently fine-tune (transfer) the model using the females' training data. Recognition of a dual CNN model (male + female) achieved similar performance to that of the RF trained just with the male model, but in the former, training data from females were also used. True positives for women and men reported 0.74 for both sexes. That means an increase in performance with respect to the sex-specific classifier for men while it remains at the same accuracy levels for women. From these results, we can intuit a superiority in the deep learning model at decoding patterns that can be equally discriminatory across the sexes.

4.3. Interpretation

The explainability of the biomarkers within the Mel-frequency cepstrum was determined by the proposed xDMFCCs method. For each prediction of the test, the xDMFCCs returned a map of coefficients based on the time domain; see Figure 6. In particular, the most influential regions were identified, and these were used to compute a global map, merging the results of each trial. This method is applied to the MFCCs

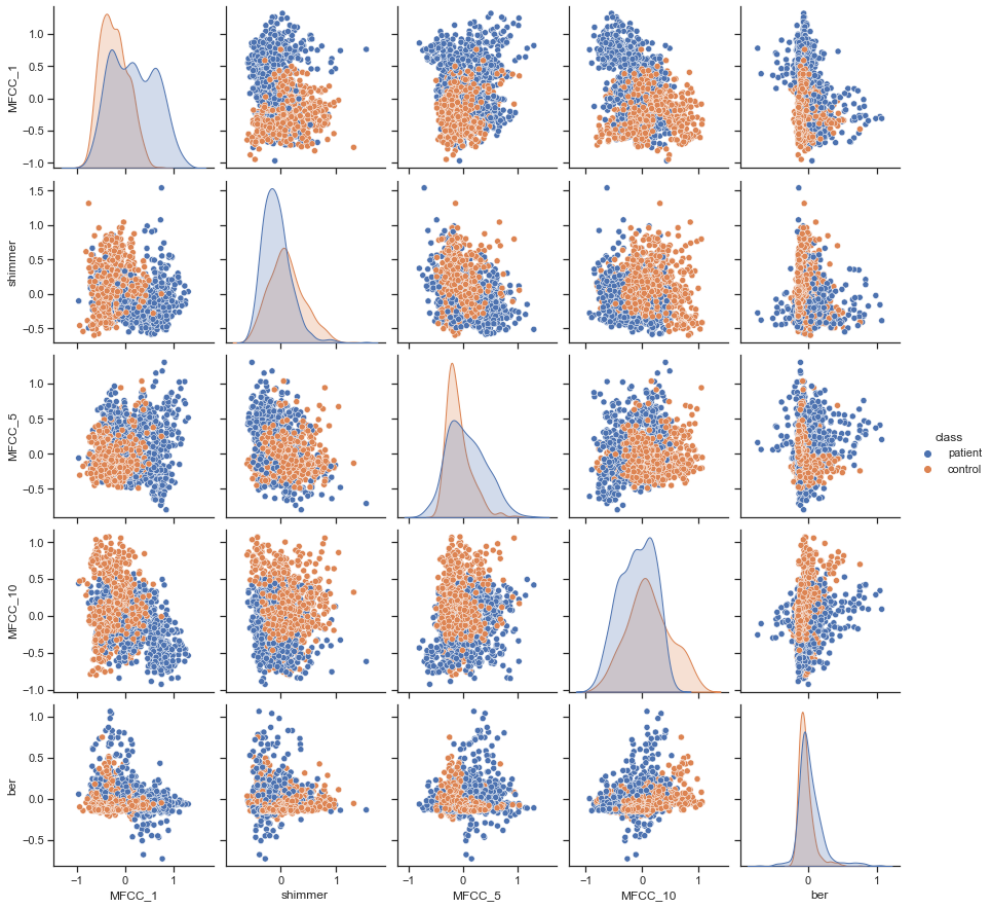

Figure 5: Distribution comparison between patient and healthy control samples of most significant features obtained through LIME algorithm.

Table 2: Performance of models trained to address the binary classification task of patients and healthy controls.

Model	Sex model	Acc	Spec	Prec	Recall	F-score
SVM	Male	0.56	0.62	0.62	0.72	0.67
RF		0.68	0.71	0.71	0.79	0.75
SVM	Female	0.72	0.73	0.73	0.70	0.72
RF		0.74	0.74	0.74	0.75	0.74
SVM	Male + Female	0.60	0.62	0.62	0.75	0.67
RF		0.63	0.65	0.65	0.69	0.67
CNN		0.73	0.75	0.76	0.75	0.75

as specified in Section 3.2.3, and the global, regional matrix is computed. The result for these matrices is shown in Figure 7. We have added the density of the location of the phonemes in sync with the time dimension on top of the matrix for reference purposes. The most prominent biomarkers for identifying a control speaker and a speaker with brain injury are shown in the left and right matrices, respectively. In this figure, several indicative patterns of their discriminatory speech can be interpreted. The first two MFCCs are presented in the research literature as indicators of the sharpness and clarity of speech sounds. In fact, the first MFCC is usually referred to as the average power of the input speech signal, and the second MFCC is the balance of the spectral energy distribution between the lower and higher frequencies [33]. B. Zhen et al. [34] found that MFCCs from 2 to 16 contain the most useful speaker information and MFCCs from 1 to 12 contain the most useful speech information.

According to the above interpretation of the coefficients, by looking at their relevancy in Figure 7 with respect to time, we can say that control individuals have a higher clarity on almost all phonemes at an earlier time of speech on-set. xDMFCCs reveals that healthy individuals are able to articulate the transition between the phonemes ”A” and ”NN” with a higher clarity with respect to patients. Also, healthy subjects denoted a larger number of discriminatory regions in the higher order coefficients (from 3 and above). Conversely, the temporal distribution of the phones is different for the two groups; in individuals with brain injury most clear common speech feature was the late emission of the second phoneme “A” at approximately half a second after the speech on-set. It is worth noting that brain injury patients delayed the emission of the “A” phoneme, which is a strong discriminatory feature. Patients speak at a slower pace than control subjects, beginning their vocalizations a little later and pronouncing their final syllables for longer, as revealed by xDMFCCs.

It is worth noting that the numeric collective significance of each MFCC (Table 3) for the sex-neutral model (female + male) is the intersection of the MFCCs for males and females, respec-

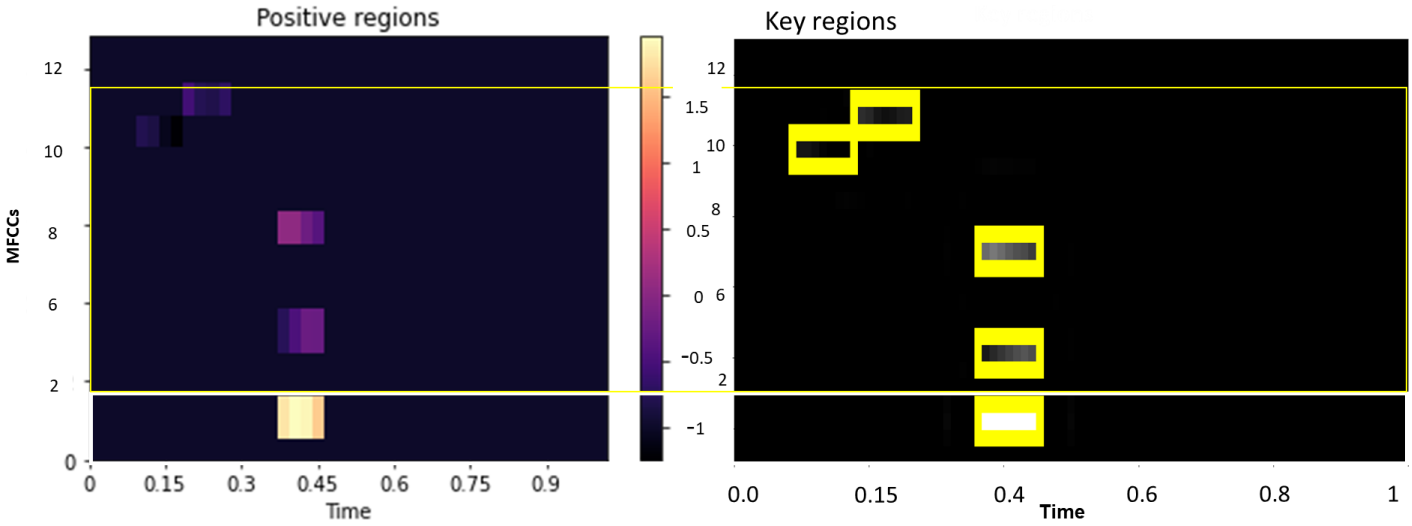

Figure 6: Explanation for a single control prediction where the 5 most influential regions are highlighted

tively. This intersection suggests again the highest discriminatory information in the coefficients related to acoustic clarity. However, this does not replicate the significance of the temporal bin (Table 4). Regarding patients, it can be observed that the last part of the utterance can get longer ($> t_6$).

5. Discussion

This study aimed to explore how deep learning can be used to identify measurable biomarkers within MFCCs to help assess speech production difficulties in patients with acquired brain injuries. The results show that MFCCs can successfully encode prototypical acoustic patterns that allow one to discriminate between speech provided by healthy controls and individuals with brain injuries. Specifically, findings highlight that speech from speakers without brain injury is marked through higher clarity.

Early findings from stroke patients who displayed brain lesions in the right hemisphere revealed difficulties in expressing their feelings through voice cues [35]. In particular, difficulties in controlling pitch and intonation variation have been described. Here, neither pitch nor intonation contour variables contributed significantly to the success of the discrimination model. However, current results add to existing evidence that speech from patients with acquired brain damage displays prototypical patterns that differentiate from healthy control speech even in the absence of diagnosed depression or speech and language disorders [36, 35]. Here, findings suggest that the effects of localized lesions can be fairly subtle as indexed through vocal clarity differences. Crucially, MFCCs have been linked to detecting small variations in terms of articulatory movements suggesting that the patient populations of interest here might express psychological states less successfully because of insufficient control over articulators, leading to the perception of reduced sharpness or clarity of a sound [37]. The basal ganglia have long been associated with motor movement disabilities in language production [38], and patients with lesions in the basal ganglia formed the largest subgroup of our patient population. MFCCs findings from our sample are thus complementing the existing literature.

The approach presented here has not only outlined that it is possible to differentiate between patients with lesions and healthy controls; it also shows promise in terms of helping clinicians diagnose neurogenic speech production problems in patients. This, in turn, can help to develop speech therapy approaches that aim to voice quality in patients better. In the absence of clear differences in pitch, loudness, or speech rate use between groups, the current results can be taken as the first indicators of how psychogenic expressions differ between groups. In the future, the approach might also be used to aid in the diagnosis of strokes as it allows the detection of abnormalities in speech pre-hospital admission.

Despite the limited number of subjects, the study employed a rigorous validation methodology to ensure scientific validity. The validation included appropriate controls, randomization where applicable, and robust statistical analyses. These measures aimed to mitigate bias and provide a reliable evaluation of trained models. While larger sample sizes are desirable to enhance generalizability, the contribution of this study lies in its pioneering nature and the generation of a unique dataset serving as a springboard for future studies in the field. The small sample size is acknowledged as a drawback, and future research should aim to replicate these findings with larger and more diverse cohorts. Additionally, conducting long-term follow-up studies can provide more robust and explainable brain lesion classification models.

6. Conclusion

In the presented paper and for the first time to the authors' knowledge, a single audio composed of three phonemes has been used to diagnose and interpret the phonetic characteristic of brain lesion patients at risk of developing mental illness.

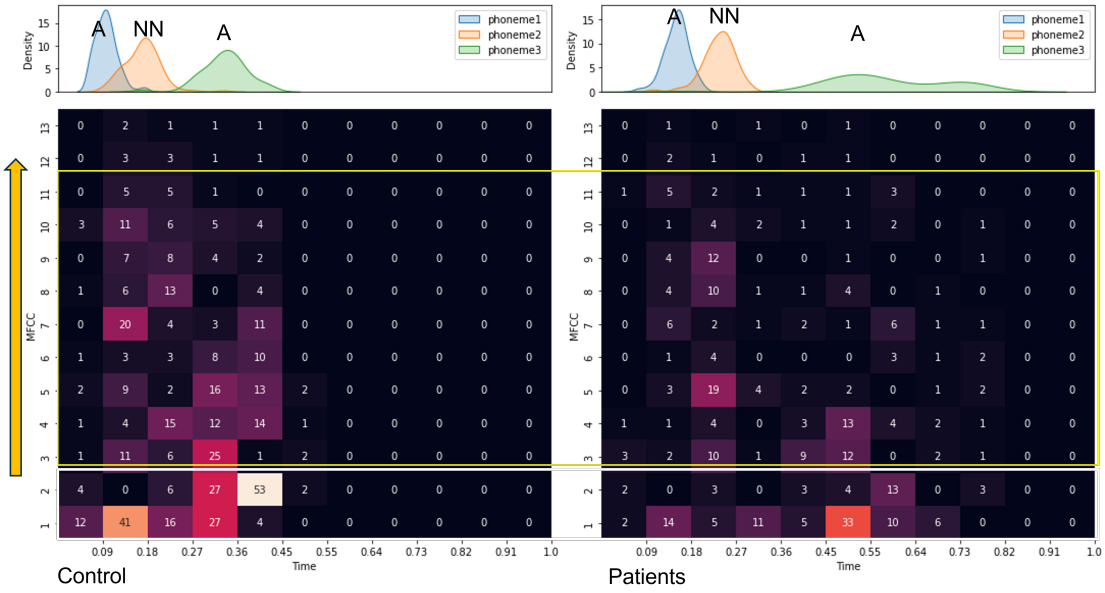

Figure 7: Interpretations provided by xDMFCCs. The heatmap on the left indicates significant regions for control samples, whereas the one on the right highlights significant areas for patient samples.

Table 3: Collective significance of each MFCC feature

Sex	Class	MFCC 1	MFCC 2	MFCC 3	MFCC 4	MFCC 5	MFCC 6	MFCC 7	MFCC 8	MFCC 9	MFCC 10	MFCC 11	MFCC 12	MFCC 13	R
Female	Control	**0.210**	**0.184**	0.094	0.107	0.098	0.058	0.066	0.045	0.045	0.053	0.015	0.012	0.006	r=0.83, $p < 0.05$
	Patient	**0.286**	0.086	0.110	0.089	0.092	0.029	0.083	0.065	0.050	0.038	0.047	0.008	0.008	
Male	Control	**0.187**	**0.136**	0.085	0.038	0.073	**0.155**	0.063	0.069	0.025	0.066	0.060	0.019	0.019	r=0.66, $p < 0.05$
	Patient	**0.286**	0.086	0.110	0.089	0.092	0.029	0.083	0.065	0.050	0.038	0.047	0.008	0.008	
Male + Female	Control	**0.201**	**0.165**	0.091	0.079	0.088	0.097	0.065	0.055	0.037	0.058	0.033	0.015	0.011	r=0.85, $p < 0.05$
	Patient	**0.285**	0.088	0.090	0.069	0.079	0.059	0.077	0.082	0.044	0.063	0.036	0.014	0.008	

Upon determination of the relevancy of the extracted features, it was found that the MFCCs yield a higher classification ability among all extracted features (section 3.2.2).

We proposed the deep learning global explainer xDMFCCs method, to diagnose and interpret the acoustic features of brain lesion patients at risk of developing mental illness based on a single audio composed of three phonemes, providing a novel approach to early screening and diagnosis. The RF and SVM classifiers achieved a competing F-score of 0.75 F-score in distinguishing brain injury patients from healthy controls. Both male and female data were used for training, with slightly better performance in females (0.75 F-score) than in males (0.74 F-score) using the RF algorithm. Recognition of a dual RF model (male + female) achieved a lower performance (0.65 F-score), but improved considerably when using CNN (0.75 F-score). These results are very encouraging due to the difficulty of the presented problem.

We highlighted the potential of speech analysis as a non-invasive and cost-effective method for detecting brain injuries.

The value of the presented innovation is twofold: 1) a simple and low-cost method for early diagnosis of brain lesions that can lead to mental illness; 2) an approach that can produce an interpretation of the most relevant MFFCs patterns that support the classifier consider the time dimension, and therefore correlation with the particular phonemes. The second value can be beneficial for empowering new neurogenic speech disorders studies, as experts can know how to tune their experiments based on the interpretation from xDMFCCs. This unique study, with a sensitive clinical population, provides hopeful evidence of a purely data-driven practical method whose result notably matches observational evidence of previous works in the acoustic assessment of brain lesions. We hope that the adoption of the proposed method in more extensive clinical trials will provide further insight into the usefulness of this proposed method.

Although our study acknowledged the limitation of a small sample size, it served as a pioneering effort. It generated a unique dataset that can be used as a foundation for future studies in the field. Future research should aim to replicate these findings with larger and more diverse cohorts, as well as conduct long-term follow-up studies to enhance the robustness and explainability of brain lesion classification models.

These conclusions highlight the potential of deep learning models and speech analysis in the early detection and diagnosis of brain lesions, while also emphasizing the need for further research and validation.

References

[1] A. A. Hyder, C. A. Wunderlich, P. Puvanachandra, G. Gururaj, O. C. Kobusingye, The impact of traumatic brain injuries: a global perspective, NeuroRehabilitation 22 (5) (2007) 341–353.

[2] S. Sophie, M. SchusterJames, H. SmithDouglas, C. SteinSherman, Cost-effectiveness of biomarker screening for traumatic brain injury, Journal of Neurotrauma (2019).

Table 4: Collective significance of each temporal bin

Sex	Class	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	R
Female	Control	0.051	**0.236**	**0.182**	**0.281**	**0.240**	0.006	0.000	0.000	0.000	0.000	0.000	r=0.32, $\boldsymbol{p = 0.33}$
	Patient	0.026	**0.140**	**0.214**	0.074	0.104	**0.229**	**0.128**	0.047	0.032	0.000	0.000	
Male	Control	**0.126**	**0.177**	**0.200**	**0.203**	**0.161**	0.130	0.000	0.000	0.000	0.000	0.000	r=0.86, $\boldsymbol{p < 0.05}$
	Patient	0.054	**0.153**	**0.189**	**0.145**	**0.164**	**0.191**	0.090	0.010	0.000	0.000	0.000	
Male + Female	Control	0.082	**0.213**	**0.190**	**0.250**	**0.209**	0.056	0.000	0.000	0.000	0.000	0.000	r=0.67, $\boldsymbol{p < 0.05}$
	Patient	0.045	**0.149**	**0.198**	**0.122**	**0.144**	**0.204**	0.103	0.022	0.010	0.000	0.000	

[3] S. Jayachitra, A. Prasanth, Multi-feature analysis for automated brain stroke classification using weighted gaussian naïve bayes classifier, journal of circuits, systems and computers 30 (10) (2021) 2150178.

[4] N. Desai, K. Dhameliya, V. Desai, Feature extraction and classification techniques for speech recognition: A review, International Journal of Emerging Technology and Advanced Engineering 3 (12) (2013) 367–371.

[5] H. A. Patil, A. T. Patil, A. Kachhi, Constant q cepstral coefficients for classification of normal vs. pathological infant cry, ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2022) 7392–7396.

[6] M. R. Kamble, H. A. Patil, Detection of replay spoof speech using teager energy feature cues, Computer Speech & Language 65 (2021) 101140.

[7] Z. Jiang, H. Huang, S. Yang, S. Lu, Z. Hao, Acoustic feature comparison of mfcc and czt-based cepstrum for speech recognition, in: 2009 Fifth International Conference on Natural Computation, Vol. 1, IEEE, 2009, pp. 55–59.

[8] K. Khoria, M. R. Kamble, H. A. Patil, Teager energy cepstral coefficients for classification of normal vs. whisper speech, in: 2020 28th european signal processing conference (EUSIPCO), IEEE, 2021, pp. 1–5.

[9] Y. Gu, M. Bahrani, A. Billot, S. Lai, E. J. Braun, M. Varkanitsa, J. Bighetto, B. Rapp, T. B. Parrish, D. Caplan, et al., A machine learning approach for predicting post-stroke aphasia recovery: A pilot study, in: Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments, 2020, pp. 1–9.

[10] T. M. Hope, M. L. Seghier, A. P. Leff, C. J. Price, Predicting outcome and recovery after stroke with lesions extracted from mri images, NeuroImage: clinical 2 (2013) 424–433.

[11] S. Pereira, R. Meier, R. McKinley, R. Wiest, V. Alves, C. A. Silva, M. Reyes, Enhancing interpretability of automatically extracted machine learning features: application to a rbm-random forest system on brain lesion segmentation, Medical image analysis 44 (2018) 228–244.

[12] A. Ditthapron, E. O. Agu, A. C. Lammert, Learning from limited data for speech-based traumatic brain injury (tbi) detection, in: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2021, pp. 1482–1486.

[13] T. Talkar, S. Yuditskaya, J. R. Williamson, A. C. Lammert, H. Rao, D. J. Hannon, A. T. O'Brien, G. Vergara-Diaz, R. DeLaura, D. E. Sturim, et al., Detection of subclinical mild traumatic brain injury (mtbi) through speech and gait., in: INTERSPEECH, 2020, pp. 135–139.

[14] D. M. Jeremiah, Detecting depression from speech with residual learning, Ph.D. thesis, Dublin, National College of Ireland (2020).

[15] M. Muzammel, H. Salam, Y. Hoffmann, M. Chetouani, A. Othmani, Audvowelconsnet: A phoneme-level based deep cnn architecture for clinical depression diagnosis, Machine Learning with Applications 2 (2020) 100005.

[16] M. Niu, B. Liu, J. Tao, Q. Li, A time-frequency channel attention and vectorization network for automatic depression level prediction, Neurocomputing 450 (2021) 208–218.

[17] N. Srimadhur, S. Lalitha, An end-to-end model for detection and assessment of depression levels using speech, Procedia Computer Science 171 (2020) 12–21.

[18] T. Taguchi, H. Tachikawa, K. Nemoto, M. Suzuki, T. Nagano, R. Tachibana, M. Nishimura, T. Arai, Major depressive disorder discrimination using vocal acoustic features, Journal of affective disorders 225 (2018) 214–220.

[19] J. Wang, L. Zhang, T. Liu, W. Pan, B. Hu, T. Zhu, Acoustic differences between healthy and depressed people: a cross-situation study, BMC psychiatry 19 (1) (2019) 1–12.

[20] D. M. Low, G. Randolph, V. Rao, S. S. Ghosh, P. C. Song, Uncovering the important acoustic features for detecting vocal fold paralysis with explainable machine learning, medRxiv (2020).

[21] N. Seedat, V. Aharonson, Y. Hamzany, Automated and interpretable m-health discrimination of vocal cord pathology enabled by machine learning, in: 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), IEEE, 2020, pp. 1–6.

[22] K. Feng, Toward knowledge-driven speech-based models of depression: Leveraging spectrotemporal variations in speech vowels, in: 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), IEEE, 2022, pp. 01–07.

[23] P. M. D. Paulmann, S., S. A. Kotz, Comparative processing of emotional prosody and semantics following basal ganglia infarcts: Erp evidence of selective impairments for disgust and fear, Brain research 1295 (2009) 159–169.

[24] S. S. Paulmann, S., S. A. Kotz, Orbito-frontal lesions cause impairment during late but not early emotional prosodic processing, Social Neuroscience 5 (1) (2010) 59–75.

[25] H. A. Kotz, S. A., S. Paulmann, On the orbito-striatal interface in (acoustic) emotional processing, Vol. 35, Oxford University Press, New York, NY, 2013.

[26] G. J. F. S. R. D. Moebes, Janine, C. Schroeder, Emotional speech in parkinson's disease, Movement Disorders 23 (6) (2008) 824–829.

[27] P. Boersma, Praat, a system for doing phonetics by computer, Glot. Int. 5 (9) (2001) 341–345.

[28] M. Saarela, S. Jauhiainen, Comparison of feature importance measures as explanations for classification models, SN Applied Sciences 3 (2) (2021) 1–12.

[29] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[30] S. Mishra, B. L. Sturm, S. Dixon, Local interpretable model-agnostic explanations for music content analysis., in: ISMIR, 2017, pp. 537–543.

[31] J.-P. Goldman, Easyalign: an automatic phonetic alignment tool under praat, in: Interspeech'11, 12th Annual Conference of the International Speech Communication Association, 2011.

[32] M. La Mura, P. Lamberti, Human-machine interaction personalization: a review on gender and emotion recognition through speech analysis, in: 2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT, IEEE, 2020, pp. 319–323.

[33] M. A. A. Wusu-Ansah, Emotion recognition from speech: An implementation in matlab (2019).

[34] B. Zhen, X. Wu, Z. Liu, H. Chi, On the importance of components of the mfcc in speech and speaker recognition, in: Sixth international conference on spoken language processing, 2000.

[35] B. Shapiro, M. Danley, The role of the right hemisphere in the control of speech prosody in propositional and affective contexts, Movement Disorders 25 (1985) 19–36.

[36] R. D. House, A., P. J. Standen, Affective prosody in the reading voice of stroke patients, Journal of Neurology, Neurosurgery and Psychiatry 50 (7) (1985) 910–912.

[37] L. M. A. M. P. E. S. J. Tsanas, A., L. O. Ramig, Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease, IEEE transactions on biomedical engineering 59 (5) (2012) 1264–1271.

[38] M. C. Silveri, Contribution of the cerebellum and the basal ganglia to language production: Speech, word fluency, and sentence construction—evidence from pathology, The Cerebellum 20 (2) (2021) 282–294.