Smart Health 23 (2022) 100233

Contents lists available at ScienceDirect

Smart Health

journal homepage: www.elsevier.com/locate/smhl

VoiceLens: A multi-view multi-class disease classification model through daily-life speech data

Soumyadeep Bhattacharjee [a,c,*], Wenyao Xu [b]

[a] Gifted Math Program, State University of New York at Buffalo, USA
[b] The State University of New York at Buffalo, NY, USA
[c] Williamsville East High School, East Amherst, NY, USA

ARTICLE INFO

Keywords:
Audio analysis
Speech processing
Multi-class classification
Disease prediction
Health care analytics

ABSTRACT

Biomarkers in the human voice can offer insight into neurological disorders because voice signals are influenced by the underlying cognitive and neuromuscular functions. In the past decade, there is an increasing research attention on voice-based neural disorder detection using machine learning techniques. However, existing works only attempt to detect a single neurological disorder (e.g., Parkinson's or Huntington's). In this work, we present the first computational model, namely *VoiceLens*, that detects multiple neurological disorders at the same time. The proposed *VoiceLens* framework combines the effectiveness of the powerful Mel-Frequency-Cepstral-Coefficients (MFCC) within a two-phase multi-class classification module to build an accurate voice-based disease prediction model. The first phase captures the fine-grained details of these disorders and their sequential variation patterns within a stacked Long-Short-Term-Memory (LSTM) network to make the baseline disease detection, i.e., healthy v.s. pathology. In the second phase, the detected pathology samples are further analyzed by a deep multi-layer learned descriptor to identify the disease types. The *VoiceLens* method is developed and evaluated using a large-scale Saarbruecken-Voice-Database comprising of samples from 2000 individuals with multiple disease patterns, including Laryngeal Cancers, Dish-Syndrome, and Parkinson's disease. Experimental results show the remarkable performance of *VoiceLens* by reporting *Accuracy* up to 97.5% in the disease detection, where the model also obtains 98.00% and 97.13% for *F1-Score* and *Recall*. Also, compared with existing machine learning models, the proposed *VoiceLens* system demonstrates around 15% (and 12%) average gain in the *Accuracy* (and *F1-score*) in a multi-disease identification test including six (6) different pathology classes and one (1) healthy class.

1. Introduction

Voice pathology is the study and diagnosis of the disease using voice as a biomarker. While it is well known that most vocal fold pathologies can be characterized by observing changes in the acoustic voice signal, distinguishing different pathology conditions within an integrated multi-class classification framework is indeed challenging. Early detection of a disease is important and this can be effectively achieved using voice signals. Voice usually provides an early sign or symptom related to a disease, such as Parkinson's, because multiple issues like tissue infection, mechanical stress, surface irritation, neurological and muscular changes, and several other disordered conditions (Al-Dhief et al., 2020) affect voice samples. The affected functionality and shape of the vocal folds result

* Corresponding author at: Gifted Math Program, State University of New York at Buffalo, USA.
E-mail addresses: sbhattac@buffalo.edu (S. Bhattacharjee), wenyaoxu@buffalo.edu (W. Xu).

https://doi.org/10.1016/j.smhl.2021.100233
Received 26 October 2021; Accepted 1 November 2021
Available online 28 November 2021

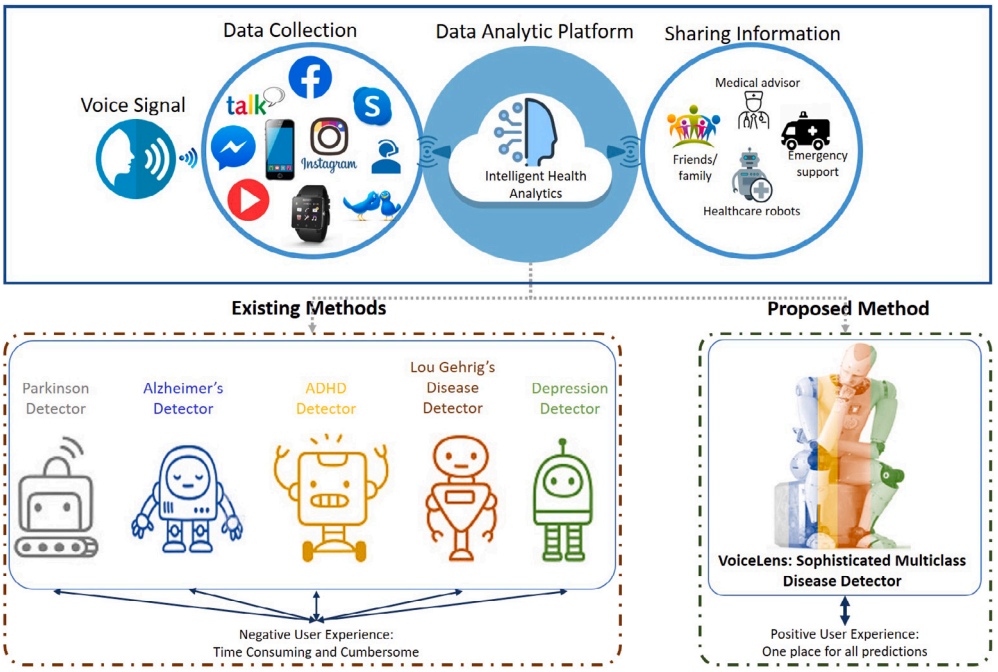

Fig. 1. Problem Overview.

in strained, harsh, weak, and breathy voices (Stathopoulos, Huber, & Sussman, 2011). Therefore, voice-based proactive multi-class disease classification is often an effective, convenient and pervasive approach, which may facilitate a more extensive follow-up diagnosis as well as a just-in-time treatment plan.

Existing works (Fang et al., 2019; Huiyi, Soraghan, Anja, & Gaetano, 2018; Wu, Soraghan, Lowit, & Di Caterina, 2018) focus on single disease detection, such as Parkinson's and ADRD, using speech, which require a separate disease-specific model to be created for every disease. Various parameters such as MFCCs, linear predictive cepstral coefficients (LPCCs), and higher-order statistics (HOS) have been used to develop machine learning and statistical methods for automatic detection of voice pathology conditions (Wang, Zhang, & Yan, 2011). A set of recent works have also employed deep learning based models for the task. However, as highlighted earlier, most of these methods still focus on addressing this simplified binary classification task and the performance deteriorates significantly when the problem setting turns more complex in a multi-class scenario. Moreover, as of now, existing voice-based pathology detection methods have some implicit subjective bias in the evaluation process (Hillenbrand & Houde, 1996). For example, (Markaki & Stylianou, 2009) uses auditory-perceptual assessments to evaluate the rate of severity of several clinically derived voice samples of Laryngitis, wherein the entire evaluation process is sensitive to the parameter choices, in addition to being time-consuming and expensive. Other invasive and manual procedures like Laryngostroboscopy require patients to be physically present at the clinic, which cannot be leveraged for regular monitoring of patients in critical conditions and can also be quite challenging for anyone in this COVID scenario.

As illustrated in Fig. 1, we aim to build an intelligent and proactive voice-based digital assistant, referred to as *VoiceLens*, which can perform a real-time screening of voice as a biomarker to evaluate the speaker's health condition, as they stay engaged in the daily-life conversations in an unconstrained environment. As illustrated in Fig. 2, given a voice signal collected via a front-end conversational chatbot (or a recording system), the proposed machine learning based intelligent model, VoiceLens, enables a real-time, non-invasive, early screening process of the individual's health condition to facilitate timely medical intervention, if required. The experimental results show remarkable improvement in *VoiceLens* performance by reporting an *Accuracy* of up to 97.5% in binary classification experiment setting, where the model also obtains 98.00% and 97.13% for *F1-Score* and Recall. The proposed *VoiceLens* system reports around 15% (and 12%) average gain in the *Accuracy* (and *F1-score*) in a multi-class scenario with seven (7) categories (i.e., six (6) different pathology classes and one (1) healthy class). The **primary contributions** of the paper are:

1. The proposed *VoiceLens* system develops the *voice-based classification model* that may *identify disease-specific pathology conditions* from both sustained phonations of vowels and the daily conversations by the speaker. Unlike many existing methods which formulate a simplified binary classification model (Healthy Vs. pathology), the proposed *multi-class classifier* offers a real-time disease-specific prediction to facilitate a more appropriate follow-up treatment plan.
2. By analyzing *the sequential utterance patterns of the conversation* within the LSTM model, the proposed method attains more detailed insights on the signal characteristics that may facilitate a more precise disease prediction by the consequential multi-class deep classifier of *VoiceLens*.

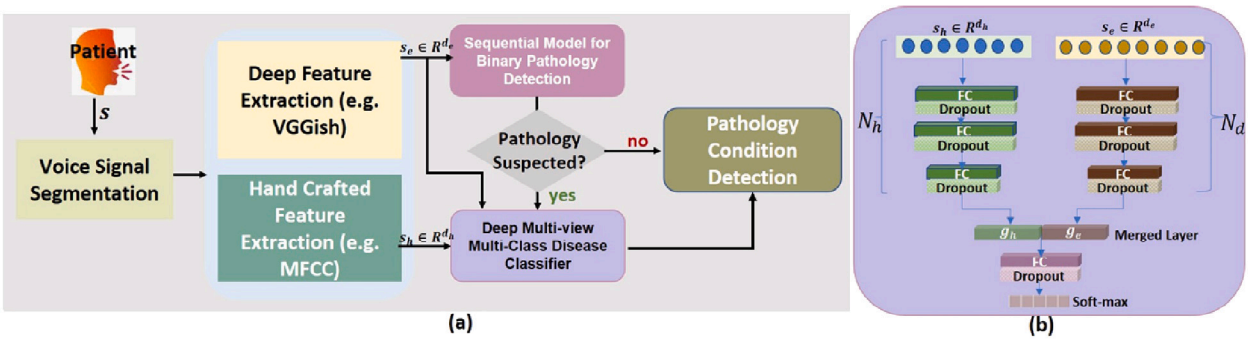

Fig. 2. An overview of the proposed *VoiceLens* method is shown in (a) and the detailed Architecture of the Proposed Multi-view Multi-Class Disease Classifier is shown in (b).

3. In contrast to the existing methods addressing a simplified binary classification task (healthy vs pathology), *an extensive experimental analysis on various large-scale datasets demonstrate the generalized capacity of VoiceLens in identifying different disease-specific pathology conditions.*

The rest of the paper is organized as follows: Section 2 briefly describes the related works. The proposed method is described in Section 3. Sections 4 and 5 respectively present the experimental results and conclusion.

2. Related work

Artificial Intelligence based methods have been effective in several real-life applications (Al-Dhief et al., 2020; Al-Nasheri, Muhammad, Alsulaiman, Ali, Malki, et al., 2017; Alhussein & Muhammad, 2018; Djenouri, Laidi, Djenouri, & Balasingham, 2019; Huiyi et al., 2018; Mohammed, Abd Ghani, Arunkumar, Hamed, et al., 2018; Mohammed, Abd Ghani, Hamed, Ibrahim, & Abdullah, 2017; Wu et al., 2018). To address the task of voice-based disease detection, various research has focused on using signal processing techniques that may automatically evaluate the voice-manifestations of the pathology conditions using machine learning based models (Harar et al., 2017; Hemmerling, Skalski, & Gajda, 2016; Mohammed, Abd Ghani, Arunkumar, Mostafa, et al., 2018). Various handcrafted features (both time and frequency domain representations) that include entropy, energy, time, contained Mel-frequency cepstral coefficients (MFCC), cepstral domains, frequency, harmonics-to-noise ratio, short-term cepstral parameters, normalized noise energy have been used to represent the signals (Al-Nasheri, Muhammad, Alsulaiman, Ali, Mesallam, et al., 2017; Wang & Jo, 2006). These feature vectors are then passed as an input to a separate classification model (Steidl, 2009; Ververidis & Kotropoulos, 2006; Yap, Epps, Ambikairajah, & Choi, 2011). Eskidere and Gürhanlı (2015) use multitaper MFCC features to build a Gaussian Mixture Model (GMM) based classifier to identify the disordered voice signal. Al-Nasheri, Muhammad, Alsulaiman, Ali, Mesallam, et al. (2017) use features extracted by investigating different frequency bands using correlation functions to build an SVM(Support Vector Machine) for binary pathology condition detection. Souissi and Cherif (2016a) utilize ANN(Artificial Neural Network) and SVM(Support Vector Machine) model for classification.

In addition to being single-disease specific, most research (Martínez, Lleida, Ortega, Miguel, & Villalba, 2012; Mohammed et al., 2020) focus on utilizing sounds of sustained vowel /a/ recorded in a clinical environment for their study, while some others (Al-Nasheri, Muhammad, Alsulaiman, Ali, Mesallam, et al., 2017; Hemmerling et al., 2016; Muhammad, Alhamid, et al., 2017; Souissi & Cherif, 2015) focus on the combination of vowels /a/, /i/ and /u/ to get high accuracy. Martínez et al. (2012) utilize 200 records of sustained vowel /a/ to represent a high value. Other studies (Al-Nasheri, Muhammad, Alsulaiman, Ali, Mesallam, et al., 2017; Muhammad, Alhamid, et al., 2017; Souissi & Cherif, 2016a) leverage the combination of vowels /a/, /i/ and /u/ to get high accuracy and do not focus on the pathology causes. In a binary classification framework, Muhammad, Alhamid, et al. (2017) build a dataset using three kinds of voice pathology samples. Important to note that while at one end, such a clinically informative data collection process may not be a feasible option to adopt in a home-like environment, in the absence of a health expert around, at the other end, these methods design a simplified binary classification task to identify only one disease-specific voice disorder pattern. Thus, many rare kinds of diseases are often ignored by the research community. In contrast, this project utilizes a large scale Saarbruecken Voice Database (SVD) (Muhammad, Alsulaiman, et al., 2017) comprising of both enunciations of vowels like /a/ and usual daily conversations by the speakers from 71 different disease-specific pathology conditions, which presents a relatively new and more challenging task of multi-class classification. Some diseases we are targeting in this work to build a multi-class classifier, have not been addressed before.

To improve the binary classification performance, a set of recent works have also leveraged deep learning based models (Lee, Jeong, Choi, & Hahn, 2008; Lee, Jeong, & Hahn, 2008). Fang et al. (2019) propose a deep neural network (DNN) to classify between normal and pathological classes. Wu et al. (2018) suggest a convolution neural network (CNN) model and for feature extraction, short-time Fourier transform (STFT) features for binary classification of voice samples. Huiyi et al. (2018) develop a convolution

deep belief network (CDBN) that uses spectrograms of normal and pathological speech as the input to identify pathology conditions in a binary classification framework. Mohammed et al. (2020) address the problem of voice pathology detection by designing a CNN. In the specific field of speech-based automatic depression detection, most of the existing methods (Asgari, Shafran, & Sheeber, 2014; Cummins et al., 2015; Quatieri & Malyska, 2012) rely on the importance of several acoustic characteristics like pitch, intensity, jitter, shimmer, harmonic-to-noise ratio, and rates of speech to identify the predicting depression state of an individual. These voice quality features are related to the observation that depressed speakers tend to speak in an unnatural and monotonous way. A set of recent works (Abdel-Hamid et al., 2014; Deng et al., 2013; Golik, Tüske, Schlüter, & Ney, 2015; Lee, Lee, & Chang, 2019; Vázquez-Romero & Gallardo-Antolín, 2020) have also employed deep neural network models to perform the depression detection. Important to note that the task of building a multi-class classifier that may discriminate the signals representing multiple disease conditions within a single model is still under-explored. Furthermore, while deep learning models for voice analysis may demonstrate an improved performance in drawing a high-level binary decision, the learned deep descriptors are typically best only in deriving an overall data characteristics.

Toward this, we argue that the complementary skills of both hand-crafted and deep features may be more effective to derive a comprehensive understanding of the fine-grained disease specific patterns and time-dependent variance of the samples. To enable a generic and early identification of different disease patterns, the model must jointly learn these discriminative data patterns from multiple views designed by its hand-crafted as well as deep feature vectors. Using just one of them in isolation may not be enough to represent the unique class-specific characteristics of the voice samples.

3. Proposed method

In this work, we propose a two-phased generic deep learning based multi-class classification model that can analyze a voice signal collected via a commodity microphone (e.g., a front-end conversational chatbot or a recording system) to enable a real-time, unobtrusive, early screening process of the individual's health condition. Different from existing machine-learning based voice analytical models, the proposed *VoiceLens* system presents a new processing scheme towards robust voice feature-based classification.

Specifically, given an annotated data collection $\mathcal{D} = \{(s_j, y_j)\}_j$, s_j represents a voice signal (which may either be phonation of the vowel /a/ or a usual daily conversation by the speaker) and the corresponding label $y_j \in C$ is the label specifying the pathology condition of the signal s_j, where C represents the set of all labels in the multi-class environment. Utilizing the multi-class annotated training collection $\mathcal{D}$, the proposed *VoiceLens* system is learned to predict the pathology condition of a new voice sample $s_{test} \in \mathcal{D}_{test}$, observed in test time, for which the underlying true label information (y_{test}) is unknown. The proposed method consists of three modules: (1) *Feature Extraction*, which can be divided into two parts namely handcrafted feature representation and deep learning based feature representation. Given an input signal s, the proposed feature extraction module derives its handcrafted feature vector $\mathbf{s}_h$ and deep learned feature vector $\mathbf{s}_d$; (2) the first phase *LSTM based Binary Pathology Detection* that learns to perform the initial pathology condition detection, wherein the objective is to evaluate a given voice sample as healthy Vs. unhealthy; and (3) The second phase *Deep Multi-class Multi-view Classification*, which jointly learns the fine-tuning of $\mathbf{s}_h$ and $\mathbf{s}_d$ within its two branches, to capture the disease-specific (like Parkinson's, Alzheimer's, Heart Diseases, Laryngeal Cancers, etc.) discriminative data patterns more effectively.

3.1. Feature extraction process

An important step in both the training and classification stages of the *VoiceLens* system is the selection and extraction of the features. Typically features used for audio classification tasks may either be defined in the time domain or the frequency domain. While the time domain features are more intuitive and easy to compute, frequency domain features are more discriminative and provide greater insight into the mutual connection between the signal and its articulation by the vocal organs. In this project, given an input signal s, we use two types of feature vectors: handcrafted Feature representation scheme (e.g. Mel-Frequency Cepstral Coefficients (MFCC)) to obtain a vector $\mathbf{s}_h \in \mathbb{R}^{d_h}$ and Deep Learning Based Features (e.g. VGGish features) (Gupta, Jaafar, Ahmad, & Bansal, 2013) to obtain a vector $\mathbf{s}_d \in \mathbb{R}^{d_e}$. Fig. 3 visualizes the features.

3.1.1. Handcrafted features

In this work, we have used three types of handcrafted features for audio signal representation: Mel-Frequency cepstral coefficients (MFCC), Log Mel-filter bank coefficients (logFBANK), and Spectral Subband centroids features. Their computation process is described below.

MFCC-based sequential feature extraction process. Mel-Frequency Cepstrum Coefficients (MFCC) are popular acoustic features representing the human auditory system. The advantage of MFCC is that it considers the perceptual characteristics of the human hearing system, which as per Psycho-physical studies, are computed in a nonlinear logarithmic scale rather than in a linear scale, and the frequencies are perceived in a nonlinear frequency binning based on critical ban filtering. While the nonlinear scale is represented by the Mel scale, the critical band filtering is distinguished by the Mel filter bank. The equation connecting frequency scale (f) and Mel-scale (f_{mel}) is defined as: $f_{mel} = 2595[log(1 + \frac{f}{1000})]$.

The MFCC feature extraction task is initiated by generating the frequency-domain representation of the signal using its Fourier Transformation. The Mel filtering in the frequency domain is then determined to get the Mel filter bank coefficients. Finally, the Discrete Cosine Transform (DCT) of the log-scale Mel filter bank coefficients are derived as the resulting MFCC coefficient. Each s is

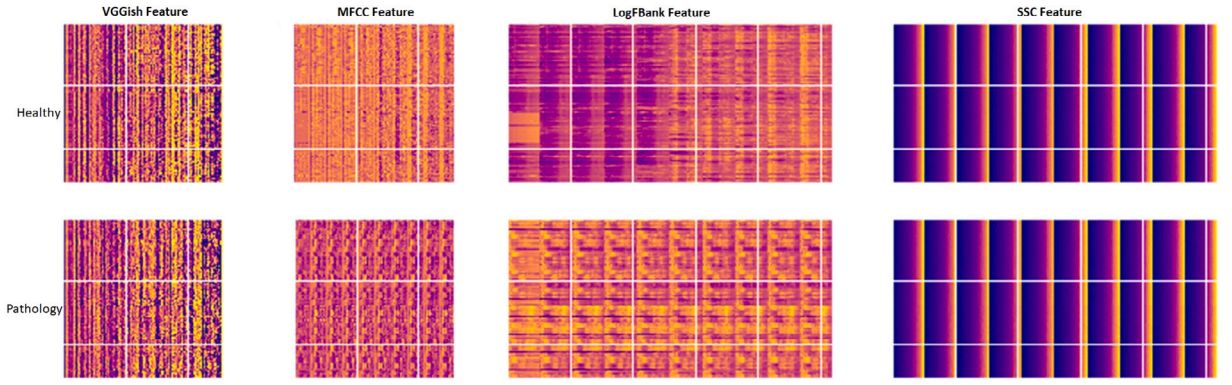

Fig. 3. The Visualizations of different Handcrafted (MFCC feature with 130 dimensions, LogFbank feature with 260 dimensions, and SSC feature with 260 dimensions) and VGGIsh deep features (with 128 dimensions) used for 128 audio signals.

represented in terms of a $d_c = 130$ dimensional MFCC features, which for the frequencies below 1 kHz, will be linear. For frequencies above 1khz, it will be logarithmic and defined as:

$$\mathbf{c}(l) = \sum_{m=1}^{M} ln((x)'(m))cos\left(\frac{l\pi(m-1/2)}{M}\right), \tag{1}$$

for $l = 1, \ldots, d_c$, where the spectral coefficients of the signal s, calculated by taking Discrete Fourier Transform (DFT) of the signal is denoted by x and $'$ presents the derivative operator.

Thus, utilizing the MFCC feature representation scheme, every voice signal s is represented in terms of a $d_h = 130$ dimensional MFCC features $\mathbf{s}_h = \mathbf{c}$.

Spectral Subband Centroid. Spectral Subband Centroid (SSC) is defined as the center of gravity of the spectral energy for the frequency band. This is defined as the unweighted sum of the power frequencies of the normalized power spectrum. As the frequency band is divided into a pre-defined number of subbands, the centroid between lower and higher edges of each subband is computed using the power spectrum of the audio signal (Paliwal, 1998; Seo et al., 2006). It is computed as:

$$\mathbf{SC}(k) = \frac{\int_{l^k}^{h^k} f w^k(f) P^\gamma(f) df}{\int_{l_j^k}^{h^k} w^k(f) P^\gamma(f) df}, \tag{2}$$

where given a signal s, $SC(k)$ is the spectral centroid of the kth band, l^k and h^k is the lower and higher edge of the frequency subband, $w^k(f)$ is the filter shape, $P(f)$ is the power spectrum, and γ is a constant controlling the dynamic range of the power spectrum with $\gamma < 1$. As reported by Thian, Sanderson, and Bengio (2004), the trajectory of SSCs in a given subband actually locates the peaks of the power spectrum in that subband. In this context, the representation is limited to one value per subband.

Thus, utilizing the SSC feature representation scheme, every voice signal s is represented in terms of a $d_h = 260$ dimensional SSC features $\mathbf{s}_h = \mathbf{SC}$ and we use $l^k = 0$ and $h^k = 25{,}000$.

Log Mel-filter bank energy. Log Mel-filter bank energy (LogFbank) via Mel-filter bank energy (Fbank) with the logarithmic transform is also used as a frequency domain feature representative. Important to note, MFCC is not robust in presence of additive noise, while LogFbank is found as a more noise-robust alternative in feature extraction (Wang, Fang, Li, & Chai, 2020). The LogFbank feature, a log power spectrum on a non-linear Mel scale of frequency, performs the logarithm transformation to control the effective strength of the Fbank energy. Therefore, the resulting features are more robust to the noise. The spectrum of a signal s denoted as $X(s)$ using Fast Fourier Transform (FFT). The power spectrum $|X(s)|^2$ with a filter bank with M filters, in which the filter $L^m(\omega)$ is a triangular filter (Sahidullah & Saha, 2012) defined as:

$$L^m(\omega) = \begin{cases} 0 & \omega < s(m-1) \\ \frac{\omega - s(m-1)}{s(m) - s(m-1)} & s(m-1) \le \omega \le s(m) \\ \frac{s(k+1) - \omega}{s(m+1) - s(m-1)} & s(m) \le \omega \le s(m+1) \\ 0 & \omega \ge s(m+1), \end{cases} \tag{3}$$

where the boundary points $s(m)$ are uniformly spaced in the Mel-scale and the **LF** is defined by taking the logarithm to the power spectrum $|X(s)|$ as:

$$\mathbf{LF}(i) = log\{\theta_1 \sum_{fl}^{fh} L_K(\omega)|X(\omega)|^2 + \theta_2\}, \tag{4}$$

for $1 \leq i \leq 260$ and we use $fl = 0$ and $fh = 25{,}000$. Thus, by utilizing the LogFbank feature representation scheme, every voice signal s is represented in terms of a $d_h = 260$ dimensional LogFBank features $\mathbf{s}_h = \mathbf{LF}$.

3.1.2. Deep learning based features

We extract short-term audio features from the VGGish model (Hershey et al., 2017), which is a deep audio embedding model, inspired by VGGNet (Simonyan & Zisserman, 2014). VGGish audio embedding network is pre-trained on a large YouTube-8M dataset (Abu-El-Haija et al., 2016) and produces a 128-dimensional audio feature. We utilize the VGGish opensource implementation[1] for our experiments. Thus, by utilizing the VGGish feature representation scheme, every voice signal s is represented in terms of a $d_e = 128$ dimensional deep features $\mathbf{s}_e$.

3.2. Proposed LSTM-based binary pathology detection

To obtain a detailed insight on the signal characteristic, each input audio signal s is divided into a sequence of n smaller time-stamped signal segments and presented as a sequence $\{s^{(t)}\}_{t=1}^{n}$ and each $s^{(t)}$ is represented in terms of a handcrafted MFCC feature $\mathbf{c}^{(t)}$, which appears to be most effective among the several handcrafted features that we have considered in the experiments of this paper. In this work, the Long-Short-Term Memory (LSTM) network model (Pascanu, Gulcehre, Cho, & Bengio, 2013), a variant of the Recurrent Network Model (RNN), is used for the first phase binary pathology condition detection module. RNNs form a chain-like neural network architecture that takes into consideration the current input in the context of the information from the past, to propagate the relevant historical information. While RNNs face a vanishing gradient problem and are unable to learn long-term dependencies, LSTM integrates the gating functions into its state dynamics to provide an efficient alternative. We use a stacked LSTM with L_0 layers (Ullah, Ullah, Khan, & Cheikh, 2019), in which the final hidden layer L_0, $\mathbf{n}^{(t)} = n_{L_0}^{(t)}$, depends on the input sequence and cell state is defined as follows:

$$
\begin{aligned}
\mathbf{n}_l^{(t)} &= \mathbf{h}_l^{(t)} \odot tanh(\mathbf{a}_l^{(t)}) \\
\mathbf{a}_l^{(t)} &= \mathbf{e}_l^{(t)} \odot \mathbf{a}_l^{(t-1)} + \mathbf{f}_l^{(t)} \odot \mathbf{b}_l^{(t)} \\
\mathbf{b}_l^{(t)} &= tanh\left(\mathbf{U}_l^{g}\mathbf{n}_{l-1}^{(t)} + \mathbf{V}_l^{g}\mathbf{n}_l^{(t)}\right) \\
\mathbf{e}_l^{(t)} &= \sigma\left(\mathbf{U}_l^{r}\mathbf{n}_{l-1}^{(t)} + \mathbf{V}_l^{r}\mathbf{n}_l^{(t)}\right) \\
\mathbf{f}_l^{(t)} &= \sigma\left(\mathbf{U}_l^{j}\mathbf{n}_{l-1}^{(t)} + \mathbf{V}_l^{j}\mathbf{n}_l^{(t)}\right) \\
\mathbf{h}_l^{(t)} &= \sigma\left(\mathbf{U}_l^{o}\mathbf{n}_{l-1}^{(t)} + \mathbf{V}_l^{o}\mathbf{n}_l^{(t)}\right)
\end{aligned}
\tag{5}
$$

where $\mathbf{n}_0^{(t)} = \mathbf{c}^{(t)}$, $\sigma()$ represents the logistic sigmoid function, and $\odot$ represents element-wise multiplication. The terms $\mathbf{a}_l^{(0)}$, $\mathbf{n}_l^{(0)}$ are set to zero vectors for all $1 \leq l \leq L_0$. The term $\mathbf{b}_l^{(t)}$ is a hidden representative based on the current input and the previous state. The terms $\mathbf{e}_l^{(t)}$, $\mathbf{f}_l^{(t)}$, and $\mathbf{h}_l^{(t)}$ determine the cell information flow with time, how the input is incorporated into the cell state, and the relation between the hidden state and the cell state, respectively. The recurrent learnable weights are depicted by $\mathbf{V}^g$, $\mathbf{V}^r$, $\mathbf{V}^j$, and $\mathbf{V}^o$, and the projection matrices by $\mathbf{U}^g$, $\mathbf{U}^r$, $\mathbf{U}^j$, and $\mathbf{U}^o$.

In this paper, every voice signal s represented as $\{\mathbf{c}^{(t)}\}_t$, is used as input to a stacked LSTM model with 2 LSTM layers. Each of these layers is followed by a drop-out layer. The number of hidden units in each of the LSTM layers is set to be 128, the drop-out ratio for each of their corresponding dropout layers is set as 0.2. The resulting output of the stacked LSTM sequence learning module is fed into a stack of FC layers for classification. The proposed model uses 2 FC layers, the activation of the last FC layer is fed into a soft-max layer to evaluate the pathology condition.

3.3. Deep multi-class multi-view disease classification

As illustrated in Fig. 2(a), a signal s, which is identified as 'unhealthy' by the initial LSTM based binary Pathology condition detection model, is further investigated for evaluating the exact disease condition that might have caused the changes in the individual's voice pattern.

3.3.1. Joint multi-view classifier learning

As shown in Fig. 2(b), the proposed deep multi-view classification network consists of two branches (namely $\mathcal{N}_h$ and $\mathcal{N}_d$), each representing one view (handcrafted feature and deep feature) of a signal.

The handcrafted branch $\mathcal{N}_h$ takes one d dimensional handcrafted feature vector $\mathbf{s}_h$ (e.g. MFCC feature where $\mathbf{s}_h = \mathbf{c}$) into a Multi-layer Dense Neural Network model, which is comprised of 3 fully connected (FC) layers with rectified linear unit (ReLU) activation function. Each of these layers is immediately followed by a drop-out layer. The number of hidden units in each of the first two Dense layers is set to be 256 and the number of hidden units in the last Dense layer is set to be 128. The drop-out ratio for each of their corresponding dropout layers is set to be 0.2. Given a signal s, the activation of the 3rd FC layer of $\mathcal{N}_h$ is represented as a compact learned feature representative with 128 dimensions $\mathbf{g}_h$. In our experiments, the handcrafted branch uses $d_h = 130$ dimensional input for MFCC and $d_h = 260$ dimensional inputs for both LogFBank and SSC (described in Section 3.1.1). The deep branch $\mathcal{N}_d$ takes one

[1] VGGish:https://github.com/tensorflow/models/tree/master/research/audioset/vggish.

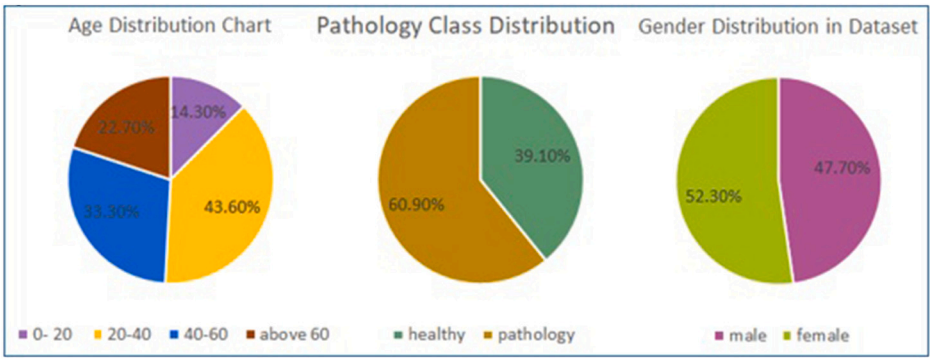

Fig. 4. Dataset Distribution for the Saarbruecken Voice Database (SVD) (Woldert-Jokisz, 2007).

d_e dimensional feature vector (e.g. Vggish feature $\mathbf{s}_e$) into an identical Multi-layer Dense Neural Network model (as of $\mathcal{N}_h$). Given a signal s, the activation of the 4th FC layer of $\mathcal{N}_d$ is also represented in terms of a compact learned feature representative with 128 dimensions $\mathbf{g}_e$. Both $\mathcal{N}_h$ and $\mathcal{N}_d$ are considered as a merged network, which is merged by concatenating the deep-learned features and feed to a new fully connected layer. As illustrated in the figure, during training, the features of $\mathcal{N}_h$ and $\mathcal{N}_d$ are concatenated as the input for fully connected layers and jointly learn the entire network for multi-class disease classification task. In order to guide this merged network to extract disease-specific data patterns, we pre-train the weights of both sub-networks ($\mathcal{N}_h$ and $\mathcal{N}_d$) separately with the corresponding annotated feature representations of D, except the 5th fully connected layer. To highlight the discriminative data patterns for each class, the loss function, $Loss$ is defined as follows:

$$Loss(\mathbf{W}) = -\frac{\sum_{y\in\mathcal{Y}}\sum_{j=1}^{|D|}(\mathbf{1}(y=y))log(p(y=y|s;\mathbf{W}))}{|D|}, \tag{6}$$

where $\mathbf{1}$. is the indicator function, $\mathbf{W}$ represents the neural network weight parameters, and $log(p(y = y|s;\mathbf{W}))$ computes the probabilistic score of the sample s_i for the class $y \in \mathcal{Y}$. The learning task is formulated as solving the minimization problem defined as: $\min_{\mathbf{W}} F(\mathbf{W})$. For pre-training $\mathcal{N}_h$ and $\mathcal{N}_d$, each signal s is represented using its corresponding feature descriptor.

4. Experiments

4.1. Dataset description

Saarbruecken Voice Database (SVD) (Woldert-Jokisz, 2007) is a collection of voice recordings and electroglottography (EGG) signals collected from more than 2 000 speakers. It contains recordings of 687 healthy persons (428 females and 259 males) and 1356 patients (727 females and 629 males) with one or more of the 71 different pathologies. One session contains the recordings of the following components: (a) vowels /i, a, u/ produced at normal, high and low pitch; (b) vowels /i, a, u/ with rising-falling pitch; and (c) sentence "Morgen, wie geht es Ihnen?" ("morning, how are you?"). All samples of the sustained vowels are between 1 to 3 s long, sampled at $fs = 50000$ Hz with 16-bit resolution [62]. All audio samples (healthy and pathological) in SVD were recorded in the same environment. Following Harar et al. (2018), we excluded samples that were shorter than 0.750 s in length (removed 319 recordings). We also excluded all recordings of speakers below the age of 19 and also above the age of 60 (it is known that the most significant changes of voice happen during adulthood until the voice matures at around the age of 20 and remains relatively stable until around the age of 60) (Stathopoulos et al., 2011). Fig. 4 provides more details on the dataset distributions. In the following section, we will describe the performance of the proposed method both in a binary class application setting (i.e., healthy Vs. pathology) and in a multi-class setting (i.e., classifying different disease conditions).

4.2. Evaluation metrics

In this work, to compare the performance of the proposed multi-class disease classifier *VoiceLens* against state-of-the-art methods, a 10-fold cross validation technique is used. We use *F1-score* (the harmonic mean of the *Precision* and *Recall*) and *Accuracy* metrics as the metrics to report the performances (Powers, 2020). For multiclass classification, we use *Multiclass_Accuracy* as the metric, which are defined as

$$precision = \frac{tp}{tp+fp}.100\% \tag{7}$$

$$recall = \frac{tp}{tp+fn}.100\% \tag{8}$$

$$Accuracy = \frac{(tp+tn)}{|D_{test}|}.100\% \tag{9}$$

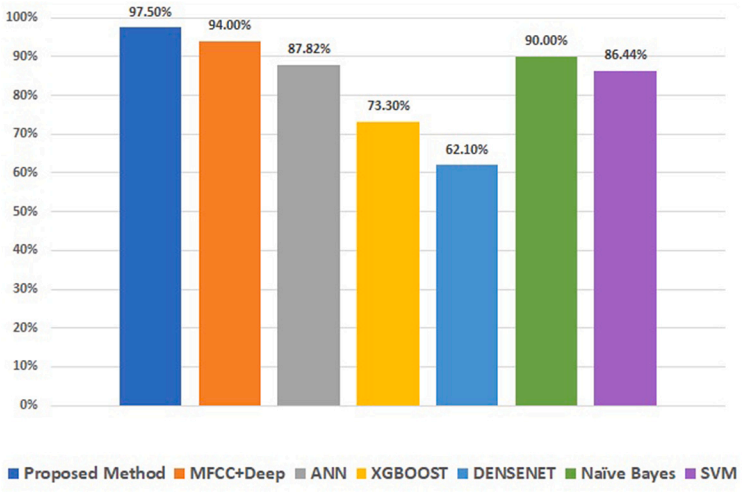

Fig. 5. Comparative Study on the Binary Pathology Detection task performed using the Saarbruecken Voice Database (SVD) (Woldert-Jokisz, 2007) dataset.

$$Multiclass\_Accuracy = \frac{(\sum_k tp_k)}{|D_{test}|}.100\% \tag{10}$$

where tp presents the number of true positives, tn presents the number of true negatives, fp presents the number of false positives, and fn presents the number of false negatives in the binary classification based experimental setting, where positive class represents a pathology condition and negative class represents the healthy condition. In the multi-class experiment setting, tp_k presents the number of identified true positive samples for the kth class in the entire test collection $\mathcal{D}_{test}$.

4.3. Results

The proposed *VoiceLens* was implemented and executed in a personal computer, which is a 1.6 GHz, quad-core Windows system with 16 GB RAM, I78550U CPU, 64Bit OS with Python 3.7, Tensorflow, and several other Machine learning related softwares installed.

Computational complexity analysis. Table 3 describes the average time complexity obtained for the proposed joint learning model. We have used Python library big_O[2]for the implementation. Since MFCC consistently reports improved performance compared to the other hand-crafted features used in our work, we have used MFCC feature representation to model the hand-crafted classifier $\mathcal{N}_h$ for this set of experiments. As observed in Table 3 the proposed method performs on the order of micro-seconds, where input signals are typically a few seconds long. This shows that the program performs in more-than-user-time as the decision time is significantly lower than length of input signal.

Binary pathology detection performance. The first set of experiments reports the performance of the initial LSTM pathology condition detection, described in Section 3.2. The experimental results show remarkable improvement in performance of *VoiceLens* by reporting an LSTM-based binary pathology condition detection *Accuracy* 97.5%, *F1-score* 98.00%, and Recall 97.13%. Fig. 5 reports the detailed comparative study on the binary classification task. The left side graph reports the performance using *Accuracy* (Eq. (9)) and the right side graph reports the performance using *F1-score*. We note that most of the existing models develop a vowel-based pathology detection scheme that uses single or concatenated signals to achieve high performance. As seen by comparing the results reported in the table, the proposed method shows an improved performance by reporting 3% improved *Accuracy* score compared to the various state-of-the-art existing supervised learning based methods. XG-Boost (Chen & Guestrin, 2016) (version 0.6) which is known for its effectiveness, efficient training phase and model interpretability. While a general heuristic is to have more training samples that trainable parameters in the learning architecture, DenseNet demonstrates promise to overcome the problem of having too many trainable parameters by densely connecting the convolutional layers. We have followed the implementation approach by Harar et al. (2018) that has adjusted Thibault de Boissiere's Keras implementation of the DenseNet (Keras framework Chollet et al., 2015, version 2.1.2), to process 1D signals treating raw audio as 1D vector. Spectrograms are processed as a matrix using the frequency bins as a stack of channels in the same way the 3 RGB channels are stacked in an image (Wyse, 2017). Souissi and Cherif (2016b) leverage MFCC, their first and second derivatives as features to detect voice impairment by using Artificial Neural Network (ANN). Dahmani and Guerti (2017) use MFCC, jitter, shimmer, and fundamental frequency as inputs to Naive Bayes classifier to

[2] https://pypi.org/project/big-O/.

Table 1
Comparative study on the first-phase LSTM based binary pathology detection task using the Saarbruechen Voice Database (SVD) dataset (Woldert-Jokisz, 2007).

Extracted features	Model method	*F1-score*	*Accuracy*
SSC	Gaussian Naive Bayes	89.88%	88.8%
	Support Vector Machine	86.8%	88.0%
	Decision Tree	82.7%	83.1%
	Random Forest	87.8%	94.8%
	Deep Neural Network	85.0%	88.0%
	Proposed LSTM-based Binary Pathology Detection	**88.0**%	**94.9**%
LogFBank	Gaussian Naive Bayes	90.5%	92.1%
	Support Vector Machine	90.6%	95.2%
	Decision Tree	90.0%	94.5%
	Random Forest	90.9%	**95.9**%
	Deep Neural Network	95.0%	94.6%
	Proposed LSTM-based Binary Pathology Detection	**95.0**%	95.7%
MFCC	Gaussian Naive Bayes	89.9%	95.4%
	Support Vector Machine	89.8%	95.4%
	Decision Tree	86.5%	94.4%
	Random Forest	91.1%	95.6%
	Deep Neural Network	93.5%	94.0%
	Proposed LSTM-based Binary Pathology Detection	**98.1**%	**97.5**%

Table 2
The performance study of the *VoiceLens* system following different configurations of its proposed Deep Multi-Class Multi-view Disease Classification module using the Saarbruechen Voice Database (SVD) dataset (Woldert-Jokisz, 2007).

Dataset distribution	Method	*F1-score* (Std. dev. over all trials)	*Accuracy* (Std. dev. over all trials)
$n_0 = 2$ diseased + 1 healthy	Single-view Deep Classifier($\mathcal{N}_d$ with VGGish)	92.2(±2.7)%	88.5(±2.26)%
	$\mathcal{N}_d$ with VGGish +$\mathcal{N}_h$ with SSC	**87.2**(±**5.5**)%	**88.1**(±**5.1**)%
	$\mathcal{N}_d$ with VGGish + $\mathcal{N}_h$ with LogFBank	**92.6**(±**3.5**)%	**90.2**(±**9.3**)%
	$\mathcal{N}_h$ with VGGish + $\mathcal{N}_h$ with MFCC	**94.5**(±**1.1**)%	**94.3**(±**1.2**)%
$n_0 = 3$ diseased + 1 healthy	Single-view Deep Classifier ($\mathcal{N}_d$ with VGGish)	88.8(±3.72)%	89.6(±2.77)%
	$\mathcal{N}_d$ with VGGish +$\mathcal{N}_h$ with SSC	90.9(±3.8)%	87.2(±3.6)%
	$\mathcal{N}_d$ with VGGish + $\mathcal{N}_h$ with LogFBank	84.5(±6.9)%	**83.1**(±**7.2**)%
	$\mathcal{N}_d$ with VGGish + $\mathcal{N}_h$ with MFCC	**90.9**(±**1.8**)%	**91.1**(±**1.7**)%
$n_0 = 4$ diseased + 1 healthy	Single-view Deep Classifier ($\mathcal{N}_d$ with VGGish)	85.43(±3.69)%	85.10(±3.54)%
	$\mathcal{N}_d$ with VGGish +$\mathcal{N}_h$ with SSC	**70.2**(±**9.7**)%	72.5(±10.4)%
	$\mathcal{N}_d$ with VGGish + $\mathcal{N}_h$ with LogFBank	**74.2**(±**6.3**)%	75.8(±7.2)%
	$\mathcal{N}_d$ with VGGish + $\mathcal{N}_h$ with MFCC	**87.6**(±**2.4**)%	**87.4**(±**2.3**)%
$n_0 = 5$ diseased + 1 healthy	Single-view Deep Classifier ($\mathcal{N}_d$ with VGGish)	82.50(±3.41)%	80.71(±3.87)%
	$\mathcal{N}_d$ with VGGish +$\mathcal{N}_h$ with SSC	**56.4**(±**10.1**)%	56.0(±8.2)%
	$\mathcal{N}_d$ with VGGish + $\mathcal{N}_h$ with LogFBank	**63.6**(±**7.6**)%	65.9(±10.8)%
	$\mathcal{N}_d$ with VGGish + $\mathcal{N}_h$ with MFCC	**83.2**(±**2.3**)%	**83.1**(±**2.4**)%
$n_0 = 6$ diseased + 1 healthy	Single-view Deep Classifier ($\mathcal{N}_d$ with VGGish)	80.75(±2.81)%	79.45(±2.23)%
	$\mathcal{N}_d$ with VGGish +$\mathcal{N}_h$ with SSC	**60.3**(±**9.8**)%	63.2(±10.1)%
	$\mathcal{N}_d$ with VGGish + $\mathcal{N}_h$ with LogFBank	**61.7**(±**9.1**)%	63.1(±8.6)%
	$\mathcal{N}_d$ with VGGish + $\mathcal{N}_h$ with MFCC	**82.2**(±**2.1**)%	**81.6**(±**1.8**)%

discriminate between three different groups: speakers with normal voice; speakers with spasmodic dysphonia; and speakers with vocal folds paralysis. Souissi and Cherif (2015) use dimensionality reduction for voice disorders identification system based on MFCC and Support Vector Machine (SVM). In this experiment setting, we also choose MFCC based feature vector as an input to a deep neural network with 3 FC layers (with 256 hidden units in the first two layers and 128 hidden units in the 3rd FC layer) immediately followed by a dropout layer (0.2 dropout ratio) and a Soft-max layer. Except for the final Soft-max layer, this deep architecture is identical with the branches ($\mathcal{N}_h$ and $\mathcal{N}_d$) of the proposed deep multi-view classification network. In Fig. 5, referred as MFCC+Deep, this deep neural network model reports *Accuracy* as 94%.

Table 1 reports the performances of the proposed binary classification module. To evaluate the effectiveness of several handcrafted features used in this work, we develop our in-house implementations of several machine learning based classifiers like Random Forest, Support Vector Machine, Decision Tree, Gaussian Naive Bayes, and Deep Neural Network, which use the same set of features are used as input and learn to make a binary decision. While MFCC demonstrates significantly better performance compared to other handcrafted features used in this work, the discriminability of SSC features appears to be considerably poor. As we make this observation, it also correlates with the visualization illustrated in Fig. 3. We note that the visualization of a randomly chosen healthy voice signal, as represented by its SSC feature vector, is very similar to its pattern observed for another randomly chosen pathology voice signal. A detailed analysis of these results prompts us to choose MFCC as the feature vector in the first phase pathology detection task of *VoiceLens*.

Table 3
Average time complexity details obtained from 10 iterations of training, where the hand crafted classifier $\mathcal{N}_h$ is modeled using MFCC feature and the deep branch $\mathcal{N}_d$ with VGGish feature of the signals.

Experiment setting	Avg. *complexity* (in seconds)
Binary classification	5.89×10^{-6}
Multi-class classification: $n_0 = 2$ diseased + 1 healthy	0.96×10^{-5}
Multi-class classification: $n_0 = 3$ diseased + 1 healthy	1.34×10^{-5}
Multi-class classification: $n_0 = 4$ diseased + 1 healthy	1.57×10^{-5}
Multi-class classification: $n_0 = 5$ diseased + 1 healthy	1.95×10^{-5}
Multi-class classification: $n_0 = 6$ diseased + 1 healthy	1.87×10^{-5}

Deep multi-class multi-view disease classification performance. The second set of experiments describes the performance of the deep learning based multi-class disease classification model described in Section 3.3. Table 2 reports the performance of *VoiceLens* in presence of up to 6 different multiple pathology classes, wherein the task is to predict the disease-specific pathology condition of the test sample. For a given $n_0 \in \{2, \ldots..6\}$ and each disease-specific pathology sample subcollection $P_k \subset \mathcal{D}_{test}$ $(k \in C)$, at a given iteration of testing, we randomly choose $(n_0 - 1)$ other pathology class-specific sample collections and the healthy sample subcollection to form the subset $\mathcal{D}_{test}^{sub} \subset \mathcal{D}_{test}$. Therefore, $\mathcal{D}_{test}^{sub}$ has samples from $n_0 + 1$ classes in total. For every choice of n_0, we reiterate the experiments with all possible choices of P_k for 20 times with 10 fold cross-validations. The table shows the performance of *VoiceLens* under different configurations of its multi-view multi-class classification module that is described in Section 3.3. Note that the proposed *Joint Multi-view Classifier* reports an improved performance over the single-view deep classifier $\mathcal{N}_d$ that leverages VGGish as a "warm start" for its lower layers and adds customized Fully connected layers for making a disease-specific classification decision. While LogFBank and SSC do not always positively contribute in terms of improving the classification performance, merging $\mathcal{N}_d$ and the handcrafted branch $\mathcal{N}_h$ using MFCC consistently improves *Accuracy* and *F1-Score* across several multi-class experiment settings with different choices of $n_0 \in \{2, \ldots..6\}$. In fact, on average, the proposed *Joint Multi-view Classifier* of *VoiceLens* system reports an improvement of around 2% over the single-view deep classifier $\mathcal{N}_d$.

Figs. 6 and 7 compare the performance of *VoiceLens* against several traditional machine learning models (Bishop, 2006), which include Random Forest, Support Vector Machine, Decision Tree, Gaussian Naive Bayes, and the single-view deep classifier ($\mathcal{N}_h$), as described in Section 3.3. To perform this set of experiments, we use LogFBank, SSC, and MFCC Features to describe the input sample to all the machine learning models mentioned earlier. Fig. 6 (and 7) reports the performance details using the *Accuracy* (and *F1-Score*) metrics. The results exhibit a more robust performance with a significantly lower standard deviation among *Accuracy* and *F1-Score* values and have an increase in the average by about $2 - 3\%$. As observed in both these figures, the proposed multi-class disease classifier module of *VoiceLens* consistently outperforms other machine learning models. The proposed joint multi-view classifier learning framework that integrates the finetuning process of $\mathcal{N}_h$ and $\mathcal{N}_d$ within a single neural network architecture also demonstrates promise in improving the performance of the single view deep classifier $\mathcal{N}_h$.

5. Discussion & conclusion

In this work, we have proposed Voice Lens, a two-phase multi-view multi-class classification module. The method leverages a powerful Joint Learning framework enhanced by the combination of handcrafted and deep features used on acoustic signals obtained from patient voice samples. As shown in the Experiment section, Voice Lens significantly outperforms traditional Machine Learning Models like Random Forest, Support Vector Machine, Gaussian Naïve Bayes, Decision Tree Classifier, and single view deep learning classifiers. However, one of the main roadblocks when addressing this problem include the lack of publicly available datasets with significant voice samples. Although the dataset used to evaluate Voice Lens, Saarbrucken Voice Database, is one of the largest databases with both healthy and pathological Voice samples available to the research community, some of the rare disease classes have a considerably sparse number of representative samples. This affects the very basic necessity of training various sophisticated machine learning including ours. Aside from this, the label information of the voice samples does not include the severity or stage of the disease in question. Therefore, it may be possible that some mild disease symptoms may be overlooked. Furthermore, some of the samples have been classified into multiple pathology classes, which can significantly affect training and testing processes. Moreover, the samples in this dataset are both male and female. This may have also affected the model as the acoustic signals produced by male and female participants are somewhat different. Additionally, the dataset in consideration is in the German Language. When deployed in a real-life scenario, the language factor may significantly affect performance. To overcome these impediments, we

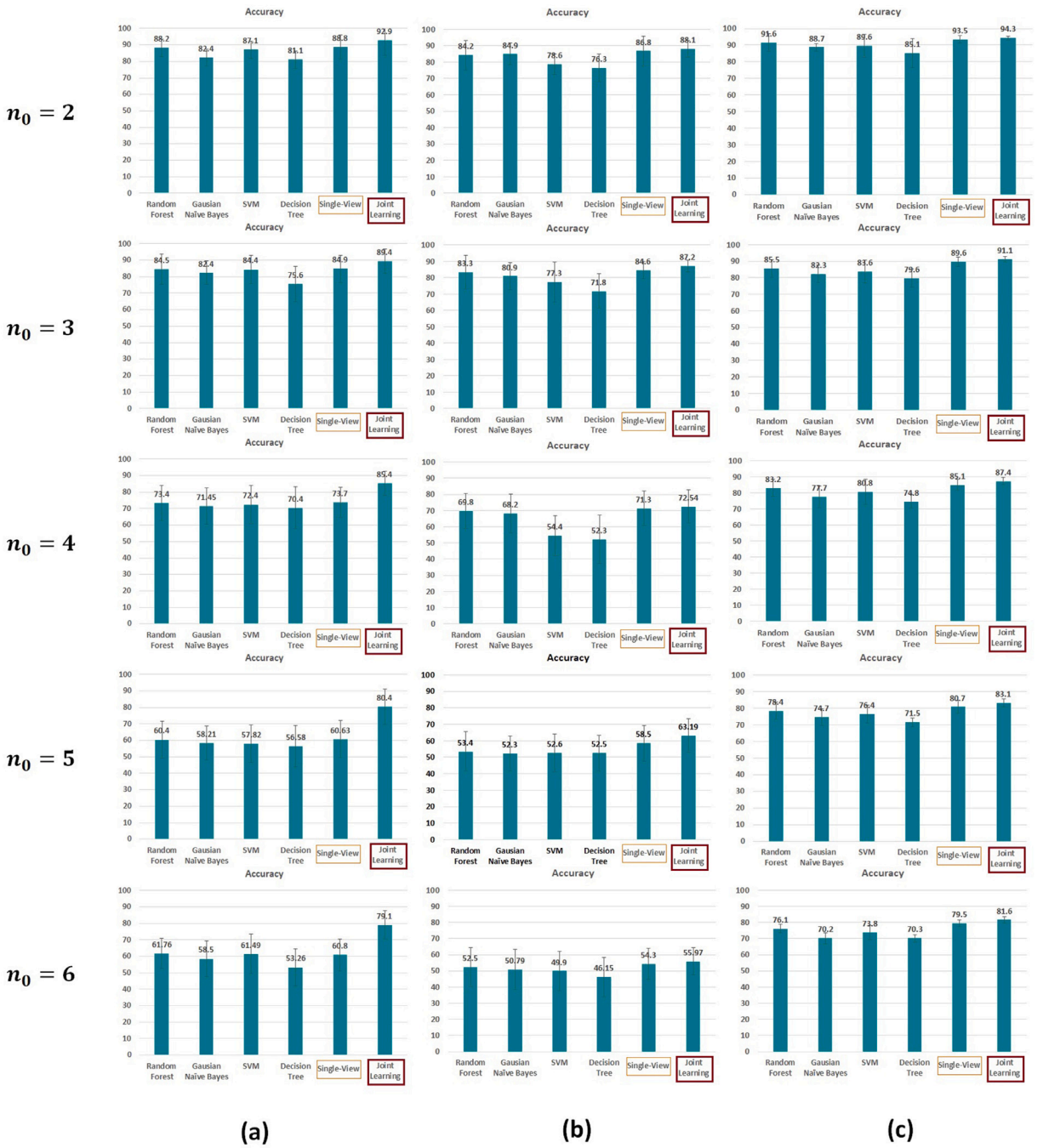

Fig. 6. Comparative Study on the proposed multi-view multi-class classification task using *Accuracy* in the Saarbruecken Voice Database (SVD) (Woldert-Jokisz, 2007) dataset. Columns (a), (b), and (c) show the results using LogFBank, SSC, and MFCC features respectively.

intend to collect data from patients both healthy and pathological under the consultation of healthcare professionals and separate this collected database based on gender, age, disease stage and ensure all categories have a similar and sufficient number of samples for training purposes. Moreover, given the sensitive nature of the problem being considered, the model must be explainable to the patient and caregiver. For example, the system must be able to identify which specific segment of the acoustic signal led the system to arrive at its decision. Toward this, one possible future extension would be to create a network architecture that may have the explainability capacity of its reasoning for its prediction.

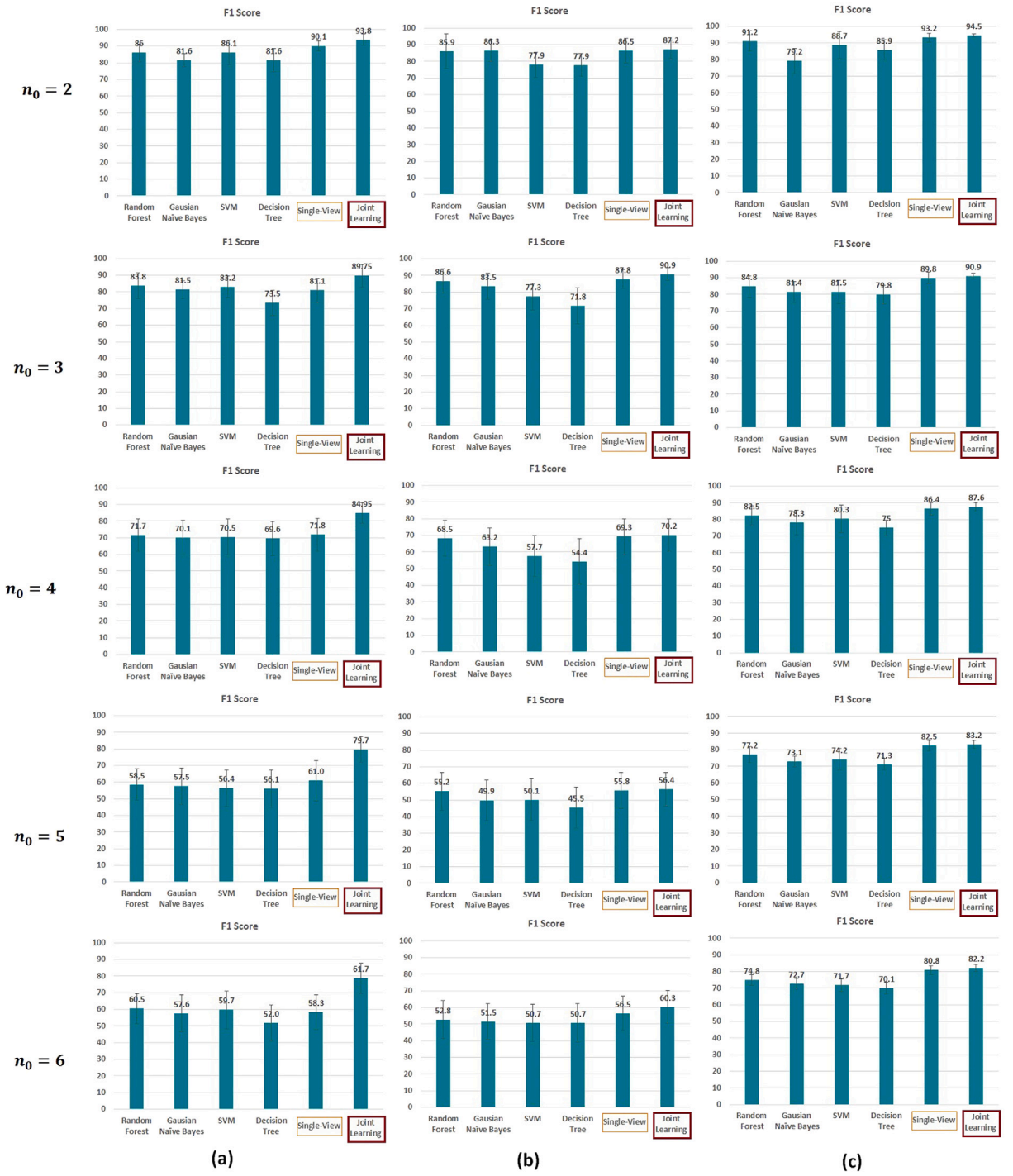

Fig. 7. Comparative Study on the proposed multi-view multi-class classification task using *F1-Score* in the Saarbruecken Voice Database (SVD) (Woldert-Jokisz, 2007) dataset. Columns (a), (b), and (c) show the results using LogFBank, SSC, and MFCC features respectively.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was in part supported by the U.S. National Science Foundation under Grant CNS-2050910 and U.S. National Science Foundation IEEE/ACM CHASE'21 Travel Grant Award.

References

Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(10), 1533–1545.

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., et al. (2016). Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675.

Al-Dhief, F. T., Latiff, N. M. A., Malik, N. N. N. A., Salim, N. S., Baki, M. M., Albadr, M. A. A., et al. (2020). A survey of voice pathology surveillance systems based on internet of things and machine learning algorithms. *IEEE Access*, *8*, 64514–64533.

Al-Nasheri, A., Muhammad, G., Alsulaiman, M., Ali, Z., Malki, K. H., Mesallam, T. A., et al. (2017). Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions. *IEEE Access*, *6*, 6961–6974.

Al-Nasheri, A., Muhammad, G., Alsulaiman, M., Ali, Z., Mesallam, T. A., Farahat, M., et al. (2017). An investigation of multidimensional voice program parameters in three different databases for voice pathology detection and classification. *Journal of Voice*, *31*(1), 113–e9.

Alhussein, M., & Muhammad, G. (2018). Voice pathology detection using deep learning on mobile healthcare framework. *IEEE Access*, *6*, 41034–41041.

Asgari, M., Shafran, I., & Sheeber, L. B. (2014). Inferring clinical depression from speech and spoken utterances. In *2014 IEEE international workshop on machine learning for signal processing* (pp. 1–5). IEEE.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining* (p. 785–794).

Chollet, F., et al. (2015). Keras: Deep learning library for theano and tensorflow. URL: https://keras.io/k, 7, (8), T1.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, *71*, 10–49.

Dahmani, M., & Guerti, M. (2017). Vocal folds pathologies classification using Naïve Bayes networks. In *2017 6th international conference on systems and control* (pp. 426–432). IEEE.

Deng, L., Li, J., Huang, J.-T., Yao, K., Yu, D., Seide, F., et al. (2013). Recent advances in deep learning for speech research at microsoft. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 8604–8608). IEEE.

Djenouri, D., Laidi, R., Djenouri, Y., & Balasingham, I. (2019). Machine learning for smart building applications: Review and taxonomy. *ACM Computing Surveys*, *52*(2), 1–36.

Eskidere, O., & Gürhanlı, A. (2015). Voice disorder classification based on multitaper mel frequency cepstral coefficients features. *Computational and Mathematical Methods in Medicine*, *2015*.

Fang, S.-H., Tsao, Y., Hsiao, M.-J., Chen, J.-Y., Lai, Y.-H., Lin, F.-C., et al. (2019). Detection of pathological voice using cepstrum vectors: A deep learning approach. *Journal of Voice*, *33*(5), 634–641.

Golik, P., Tüske, Z., Schlüter, R., & Ney, H. (2015). Convolutional neural networks for acoustic modeling of raw time signal in LVCSR, In *Sixteenth annual conference of the international speech communication association*.

Gupta, S., Jaafar, J., Ahmad, W. W., & Bansal, A. (2013). Feature extraction using MFCC. *Signal & Image Processing: An International Journal (SIPIJ)*, *4*(4), 101–108.

Harar, P., Alonso-Hernandezy, J. B., Mekyska, J., Galaz, Z., Burget, R., & Smekal, Z. (2017). Voice pathology detection using deep learning: a preliminary study. In *2017 international conference and workshop on bioinspired intelligence* (pp. 1–4). IEEE.

Harar, P., Galaz, Z., Alonso-Hernandez, J. B., Mekyska, J., Burget, R., & Smekal, Z. (2018). Towards robust voice pathology detection. *Neural Computing and Applications*, 1–11.

Hemmerling, D., Skalski, A., & Gajda, J. (2016). Voice data mining for laryngeal pathology assessment. *Computers in Biology and Medicine*, *69*, 270–276.

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., et al. (2017). CNN architectures for large-scale audio classification. In *2017 IEEE international conference on acoustics, speech and signal processing* (pp. 131–135). IEEE.

Hillenbrand, J., & Houde, R. A. (1996). Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech. *Journal of Speech, Language, and Hearing Research*, *39*(2), 311–321.

Huiyi, W., Soraghan, J., Anja, L., & Gaetano, D. C. (2018). A deep learning method for pathological voice detection using convolutional deep belief networks. In *Interspeech*.

Lee, J.-Y., Jeong, S., Choi, H.-S., & Hahn, M. (2008). Objective pathological voice quality assessment based on HOS features. *IEICE Transactions on Information and Systems*, *91*(12), 2888–2891.

Lee, J.-Y., Jeong, S., & Hahn, M. (2008). Pathological voice detection using efficient combination of heterogeneous features. *IEICE Transactions on Information and Systems*, *91*(2), 367–370.

Lee, M., Lee, J., & Chang, J.-H. (2019). Ensemble of jointly trained deep neural network-based acoustic models for reverberant speech recognition. *Digital Signal Processing*, *85*, 1–9.

Markaki, M., & Stylianou, Y. (2009). Using modulation spectra for voice pathology detection and classification. In *2009 annual international conference of the IEEE engineering in medicine and biology society* (pp. 2514–2517). IEEE.

Martínez, D., Lleida, E., Ortega, A., Miguel, A., & Villalba, J. (2012). Voice pathology detection on the saarbrücken voice database with calibration and fusion of scores using multifocal toolkit. In *Advances in speech and language technologies for Iberian languages* (pp. 99–109). Springer.

Mohammed, M. A., Abd Ghani, M. K., Arunkumar, N. a., Hamed, R. I., Abdullah, M. K., & Burhanuddin, M. (2018). A real time computer aided object detection of nasopharyngeal carcinoma using genetic algorithm and artificial neural network based on haar feature fear. *Future Generation Computer Systems*, *89*, 539–547.

Mohammed, M. A., Abd Ghani, M. K., Arunkumar, N. a., Mostafa, S. A., Abdullah, M. K., & Burhanuddin, M. (2018). Trainable model for segmenting and identifying nasopharyngeal carcinoma. *Computers and Electrical Engineering*, *71*, 372–387.

Mohammed, M. A., Abd Ghani, M. K., Hamed, R. I., Ibrahim, D. A., & Abdullah, M. K. (2017). Artificial neural networks for automatic segmentation and identification of nasopharyngeal carcinoma. *Journal of Computer Science*, *21*, 263–274.

Mohammed, M. A., Abdulkareem, K. H., Mostafa, S. A., Ghani, M. K. A., Maashi, M. S., Garcia-Zapirain, B., et al. (2020). Voice pathology detection and classification using convolutional neural network model. *Applied Sciences*, *10*(11), 3723.

Muhammad, G., Alhamid, M. F., Hossain, M. S., Almogren, A. S., & Vasilakos, A. V. (2017). Enhanced living by assessing voice pathology using a co-occurrence matrix. *Sensors*, *17*(2), 267.

Muhammad, G., Alsulaiman, M., Ali, Z., Mesallam, T. A., Farahat, M., Malki, K. H., et al. (2017). Voice pathology detection using interlaced derivative pattern on glottal source excitation. *Biomedical Signal Processing and Control*, *31*, 156–164.

Paliwal, K. K. (1998). Spectral subband centroid features for speech recognition. In *Proceedings of the 1998 IEEE international conference on acoustics, speech and signal processing, ICASSP'98 (Cat. No. 98CH36181), vol. 2* (pp. 617–620). IEEE.

Pascanu, R., Gulcehre, C., Cho, K., & Bengio, Y. (2013). How to construct deep recurrent neural networks. arXiv preprint arXiv:1312.6026.

Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061.

Quatieri, T. F., & Malyska, N. (2012). Vocal-source biomarkers for depression: A link to psychomotor activity. In *Thirteenth annual conference of the international speech communication association*.

Sahidullah, M., & Saha, G. (2012). Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. *Speech Communication*, *54*(4), 543–565.

Seo, J. S., Jin, M., Lee, S., Jang, D., Lee, S., & Yoo, C. D. (2006). Audio fingerprinting based on normalized spectral subband moments. *IEEE Signal Processing Letters*, *13*(4), 209–212.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Souissi, N., & Cherif, A. (2015). Dimensionality reduction for voice disorders identification system based on mel frequency cepstral coefficients and support vector machine. In *2015 7th international conference on modelling, identification and control* (pp. 1–6). IEEE.

Souissi, N., & Cherif, A. (2016a). Artificial neural networks and support vector machine for voice disorders identification. *International Journal of Advanced Computer Science and Appllication*, *7*(5), 339–344.

Souissi, N., & Cherif, A. (2016b). Speech recognition system based on short-term cepstral parameters, feature reduction method and artificial neural networks. In *2016 2nd international conference on advanced technologies for signal and image processing* (pp. 667–671). IEEE.

Stathopoulos, E. T., Huber, J. E., & Sussman, J. E. (2011). Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4–93 years of age. ASHA.

Steidl, S. (2009). *Automatic classification of emotion related user states in spontaneous children's speech*. Logos-Verlag.

Thian, N. P. H., Sanderson, C., & Bengio, S. (2004). Spectral subband centroids as complementary features for speaker authentication. In *International conference on biometric authentication* (pp. 631–639). Springer.

Ullah, M., Ullah, H., Khan, S. D., & Cheikh, F. A. (2019). Stacked lstm network for human activity recognition using smartphone data. In *2019 8th European workshop on visual information processing* (pp. 175–180). IEEE.

Vázquez-Romero, A., & Gallardo-Antolín, A. (2020). Automatic detection of depression in speech using ensemble convolutional neural networks. *Entropy*, *22*(6), 688.

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, *48*(9), 1162–1181.

Wang, D., Fang, Y., Li, Y., & Chai, C. (2020). Enhance feature representation of electroencephalogram for seizure detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 1230–1234). IEEE.

Wang, J., & Jo, C. (2006). Performance of gaussian mixture models as a classifier for pathological voice, In *Proceedings of the 11th australian international conference on speech science and technology*, vol. 107 (p. 122–131).

Wang, X., Zhang, J., & Yan, Y. (2011). Discrimination between pathological and normal voices using GMM-SVM approach. *Journal of Voice*, *25*(1), 38–43.

Woldert-Jokisz, B. (2007). Saarbruecken voice database. Institut für Phonetik, Universität des Saarlandes.

Wu, H., Soraghan, J., Lowit, A., & Di Caterina, G. (2018). Convolutional neural networks for pathological voice detection. In *2018 40th annual international conference of the ieee engineering in medicine and biology society* (pp. 1–4). IEEE.

Wyse, L. (2017). Audio spectrogram representations for processing with convolutional neural networks. arXiv preprint arXiv:1706.09559.

Yap, T. F., Epps, J., Ambikairajah, E., & Choi, E. H. (2011). Voice source features for cognitive load classification. In *2011 IEEE international conference on acoustics, speech and signal processing* (pp. 5700–5703). IEEE.