


Deep Learning Approach for Voice Pathology Detection and Classification

Vikas Mittal, National Institute of Technology, Kurukshetra, India

 <https://orcid.org/0000-0003-3808-5057>

R. K. Sharma, National Institute of Technology, Kurukshetra, India

ABSTRACT

A non-invasive cum robust voice pathology detection and classification architecture is proposed in the current paper. In place of the conventional feature-based machine learning techniques, a new architecture is proposed herein which initially performs deep learning-based filtering of the input voice signal, followed by a decision-level fusion of deep learning and a non-parametric learner. The efficacy of the proposed technique is verified by performing a comparative study with very recent work on the same dataset but based on different training algorithms. The proposed architecture has five different stages. The results are recorded in terms of nine different classification score indices, which are mean average precision, sensitivity, specificity, F1 score, accuracy, error, false-positive rate, Matthews correlation coefficient, and the Cohen's kappa index. The experimental results have shown that the use of machine learning classifier can get at most 96.12% accuracy, while the proposed technique achieved the highest accuracy of 99.14% in comparison to other techniques.

KEYWORDS

Deep Learning, Fusion Classification, Mel Spectrogram, Voice Pathology, Wavelet Scattering

1. INTRODUCTION

Speech is the most basic form of communications known between two groups of living entities, including human beings, animals, and/or birds. As one of the fastest ways to express one's desire or to having a task performed, speech synthesis is, nevertheless, the result of a chain of complex processes.

In the eye of a specialist, speech reflects many of the speaker's vital traits, for example, cultural or mental health condition, physical trauma or affection, sex, various sorts of emotion, and more. Suggestively, if a person's normal voice deviates from those of the same sex-age group, then after a thorough examination, a voice pathologist may describe this to be a case of laryngitis, or other voice dysfunction. Dysfunction in voice may be attributed to a series of gradual alteration in the neurological, physical, or medicinal activities in the voice synthesis structure of the human body. In most cases, disorders such as laryngeal cancer, vocal cord cyst, fold, nodule, polyp, and unilateral nerve paralysis are due to prolonged and inappropriate usage of the vocal organ which eventually results in hoarseness in the voice. Researchers have also cited that as many as 25% of the world's population inevitably suffer from different types of vocal disorder due to the increasing trend of unhealthy lifestyle and self-abuse (Al-nasheri, Muhammad, Alsulaiman, & Ali, 2017; Hammami et al., 2020).

DOI: 10.4018/IJHISI.20211001.0a28

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

Some professionals, especially teachers, singers, and religious speakers, have a higher probability of being diagnosed with the vocal nodule pathology. The underlying reason being their nature of work as they are often required to utter a series of words for quite a long duration daily, which may also lead to an abuse of their vocal cords. As more and more swollen regions accumulate in the vocal folds, stiffness in the vocal cords will increase with time. A malfunction in the vagus nerve stimulating the larynx can also cause a disorder known as unilateral nerve paralysis, which is commonly observed as hoarseness in one's voice. Typically, an indicative breathy phonation in the voice is prominently noticed, which in most cases has been observed with symptoms such as having difficulty in swallowing, and signs of shortness-of-breath and mild cough. Fortunately, the observed condition may be reversed via expert counsel and treatment (Steffen et al., 2011).

1.1 Pathology Assessment

Still, an age-old yet prevailing worldwide practice is the invasive procedure for diagnosing pathologies in the human laryngeal set-up. Laryngostroboscopy and surgical micro-laryngoscopy are two endoscopic procedures which are required to insert large devices inside the human body for mapping the internal structure. They can cause innate distress and discomfort to patients under observation.

Over the years, researchers have devised non-invasive strategies to reduce such sufferings. The electroglottography (EGG) is one such instance where the assessment of a person's voices may be performed, alleviating the patient's discomfort. Yet, there is a cost issue in training individuals to use those machines, especially if the implementation is to be on a massive scale. As championed by some research enthusiasts, cost-effective, non-invasive, and automated diagnoses of pathologies in a voice signal are trending; clearly, such diagnostic procedures for performing in-depth analysis of glottal signals represent the future of speech pathology detection.

The glottal signal originates in between the vocal fold and the vocal tract. It is expected that research work in this field will heat up especially after having witnessed the tumultuous success of deep learning (Krizhevsky et al., 2012) in the different aspects of Artificial Intelligence (AI). Indeed, deep learning has quickly replaced the traditional non-parametric classifiers such as the support vector machines (SVM), discriminant analysis (DA), k-nearest neighbour (kNN), Naive Bayes (NB), conditional random field (CRF), random forest (RF), decision trees (DT) and more. Notwithstanding, the benefit of learning to capture handcrafted features which have been the prime focus for the past couple of decades cannot and should not be ignored completely. Whereas deep learning algorithms train on the basis of using its features, the non-parametric classifiers train via previously extracted or handcrafted features. Hence, a platform which can intrinsically and inherently harness the merits provided by both these systems for learning and decision making will be more beneficial than using either of them individually.

This article offers insights into a novel process of such a methodological fusion for voice pathology and detection. So far, no previous work advancing such a system tested on state-of-the-art datasets has been found. The rest of the discussion is as follows. Section 2 overviews the background, focusing on the art of speech production, related work and key contributions of the present effort. Section 3 discusses materials and methodological issues while Section 4 concentrates on results and discussion. Finally, Section 5 offers summary insights, study limitations and future research directions. It may be noted that the sort of computer-aided voice analysis system(s) advocated here may be further studied before employing them widely as a screening tool for detecting the onset of laryngeal pathology.

2. BACKGROUND

In order for IJHISI readers to better understand the proposed novel process of methodological fusion in the context of voice pathology and detection, we present here first, the art of speech production, then discuss a series of related work on the different kinds of computer-based, automated assessment

techniques that have emerged over the last several years, and finally, highlight the key contributions of our current work.

2.1 The Art of Speech Production

Speech production requires an articulate and systematic functioning of some vital organs with these organs responsible for three primary mechanisms, including creating air pressure, providing vibration, and resonance. The muscles in the abdomen and chest, together with the rib cage, diaphragms, and lungs are collectively responsible to provide an articulate air pressure mechanism which shakes the vocal folds repeatedly resulting in a “pitch”. The “pitch” or sound produced by the vibration of the vocal fold resonates all-around a person’s throat, then in the oral and nasal cavities. The characteristics of the vocal cords and tract in turn influence the sound that the human speech system produces. Via this process, a unique voice, representing a signature of these internal organ structures, is finally produced; thus, each individual on this planet has a distinct voice.

The minute details of the voice production process are automatically handled by the brain via an internal feedback. The central nervous system (CNS) through the superior laryngeal nerve (SLN) and vagus nerve stimulates the larynx muscles with the help of connections and signals from the brain. Mucosa, body, and the vocal ligament are three unique layers occurring within the vocal folds of the larynx. Owing to their tenderness, these layers are quite sensitively exposed to vibration which may occur repeatedly anytime in the vocal fold, thereby developing a mucosal waveform. The flexibility of the mucosa to create a normal mucosal wave is intertwined with its integrity to the “superficial lamina propria.” For a normal vocal sound to be produced, a fixed volume of air pressure sweeps over the vocal fold. The pitch, which is a frequency of the wave produced by the mucosa, determines whether a person’s laryngeal functioning is normal or not.

In a normal voice, breathiness is almost absent, as any kind of abnormality in the larynx to result in a harsh and breathy sound due to the accumulated swelling of the vocal folds that develop over time (and also due to its repeated abuse). Alternatively, a breathy sound is due to a partial loss of nerve input that results in air pressure leakage. The recurrent laryngeal nerve (RLN) and SLN stimulate the laryngeal muscles for providing the appropriate position of the vocal folds.

Pathology such as gastric reflux or antero-posterior (A-P) squeezing tends to alter the physical traits (e.g., elasticity, volume, shape, and more) of vocal folds structures, which may in turn induce a different kind of vibration. Constant A-P squeezing damages the muscle tension, while laryngitis and gastric reflux cause the larynx to be inflamed, swell-up, and change the elasticity of the left and right vocal folds.

2.2 Related Works

For decades, while the art of detecting pathologies in a voice existed, it is only in the last decade that a rapid rise in popularity of different kinds of computer-based, automated assessment techniques such as algorithms that made use of the wavelet family (Lin & Chen, 1997), fractals (. et al., 2000) and computation of neural maps and networks (Hadjitodorov et al., 2000) emerges. Eventually, researchers allude to either a long or short duration mode of analysing a voice signal (Al-nasheri, Muhammad, Alsulaiman, Ali et al, 2017).

Several descriptors are used by researchers for the long duration mode of analysis. These are based largely on the disruption of amplitude and frequency, fundamental frequency, pitch and amplitude perturbation quotient, harmonics to noise ratio (HNR), voice turbulence index, the normalized energy level of noise, excitation ratio of glottal to noise, amplitude and frequency vibration, and more (Boyanov & Hadjitodorov, 1997; Boyanov et al., 1993; Ebihara & Ogawa, 1986; Gavidia-Ceballos & Hansen, 1996; Hadjitodorov & Mitev, 2002; Michaelis et al., 1997). An issue with these descriptors is their reliance on the deduction of fundamental frequency entirely, which is quite challenging for some pathological voice. To remedy, methods have emerged that are devoid of the fundamental frequency, thereby giving birth and paving the way for a modern short duration mode of speech signal assessment

(Godino-Llorente et al., 2010). Here, promising results to determine the presence of pathology in a speech signal have been shown (Godino-Llorente & Gómez-Vilda, 2004).

For detecting disorder, a given speech signal may be divided up into frames via the windowing procedure. The classification of these frames as normal or abnormal may be determined by a threshold. For analysing a voice, the linear predictive cepstral coefficients (LPCC) and the Mel-frequency cepstral coefficients (MFCC) are two examples of such descriptors currently being used. These have been developed to replicate respectively the voice and auditory sensation, which are present in human beings. Even so, it is vital to state that the LPCC does not provide good results with pathologies as it is entirely a linear model, given that pathologies are intrinsically non-linear. The MFCC, conversely, has proven to be a good assessor for voice pathologies.

A key limitation with the short-term approach is its inability to be dexterous while working within a database. In other words, a classifier trained using features from a dataset will not classify another dataset properly. Hence, the lack of generality is a major issue for its massive deployment. The cepstral peak prominence smoothed (CPPS) and its mother version, the cepstral peak prominence (CPP), are other variants of the cepstral descriptors. Diagnosing a severely dysphonic voice is not complicated as these techniques are independent of the variation in time. In fact, they have proved to be quite useful (Ali et al., 2017; Benmalek et al., 2017). In the extant literature, multiple authors have classified laryngeal and physiological pathologies via MFCC features with three (3) different base classifiers: the SVM, DA, and the gaussian mixture model (GMM). Yet, when applied individually, these classifiers failed to yield good results; hence, a combined classification approach has now been proposed (Cordeiro et al., 2017).

The voice pathology identification scheme employed in most approaches used the vowel /a/. Another body of research made use of both the sustained vowel /a/ and running speech with the Arabic voice pathology database (APVD) yielding an accuracy of over 99% (Mesallam et al., 2017; Muhammad et al., 2012). Prior to any kind of feature extraction, the voice activity detection (VAD) technique is typically performed in running-speech techniques. Its function is to automatically identify the unvoiced, voice or silent parts of a sound signal is a challenging and complicated operation. Authors of (Godino-Llorente et al., 2009) used MFCC with a VAD module for articulately segmenting the voice-region of a speech signal to obtain an accuracy of 96%. They implemented their algorithm on 23 normal and 117 pathological subjects' dataset, which is part of the Massachusetts Eye & Ear Infirmary (MEEI) database (Weber, 2010).

Finally, past research has also adopted the wavelet family for utilizing the frequency-time and localized data to classify voice disorders. For instance, authors of (Fonseca et al., 2007) used Daubechies discrete wavelet transform (DWT) along with SVM and linear prediction coding; more recently, MFCC-based features with three well-known machine learning (ML) techniques, namely, deep learning (DL), GMM, and SVM, have been used (Chen & Chen,). The use of MFCC-based descriptors with DL provided the highest accuracy. Yet, as per our knowledge, no current work has performed a voice pathology classification and detection via an approach which effectively combines DL and these conventional or non-parametric classifiers in one platform, which is the focus of the current effort in an attempt to close the noted research gap.

2.3 Contributions of the Current Work

In order to address the aforementioned limitations and contribute to the emerging knowledge in the field of voice pathology detection and classification, DL has become popular as it had produced several state-of-the-art results in other fields including autonomous vehicles, agriculture, and biomedical imaging. Still, a thorough implementation of various DL methods with different features and fusion with other parametric and non-parametric classifiers is yet to be demonstrated. The issue of general usability in which a trained network performs poorly when other datasets are tested against it is another concern of modern pathological voice detection technique. In light of this, our current work offers two key contributions.

Table 1. Descriptive data of participants of the experiment

Age	Healthy Samples	Pathological Samples	Pathology's category
18-60 Years	21(M)	52(M)	Hyperkinetic dysphonia, hypokinetic dysphonia and reflux laryngitis
	37(F)	98(F)	

1. A voice pathology detection and classification architecture which uses two different feature-fusion stages.
 The first stage is implemented inside a VAD module, while the second fusion takes place at the decision or prediction stage. During the second stage, the responses produced by a wavelet scattering based ensemble classifier, and a Mel-spectrogram based DL model are fused to provide a much more robust prediction.
2. A novel VAD module is proposed which will filter out non-speech or redundant region(s) in a speech signal before being adapted for training. Here, a concatenation of nine (9) different feature descriptors will be used for training a DL network.

This brings us to a discussion on the study materials and methods.

3. MATERIALS & METHODS

3.1 Dataset

VoiceICarfEDerico II (VOICED), that is, the pathological voice dataset as proposed by Cesari et al. (2018), will be used to demonstrate the functionality and efficacy of our proposed architecture.

Table 1 summarizes the descriptive data of experimental participants in *VOICED*. In this dataset, 150 pathological and 58 healthy voice recordings exist, each possessing a characteristic of 85kHz and a 32-bit resolution. 98 female v. 52 male candidates contributed to these 150 pathological recordings, whereas another 37 females v. 21 males contributed to the 58 healthy recordings. Participants belonged to the age group of 18-60 years.

Under the pathology's category, three (3) different types of disorders were grouped, namely, hyperkinetic dysphonia, hypokinetic dysphonia, and reflux laryngitis. The rigidity in vocal folds, nodules in the vocal fold, polyps, and more are types of diseases belonging to hyperkinetic dysphonia. Notably, voice disorders such as vocal fold paralysis, laryngitis, glottic insufficiency belong to the hypokinetic dysphonia category. Finally, pharyngitis, asthma, halitosis, and night-time cough are instances belonging to the third category. The signals were acquired by recording the voice obtained by uttering the vowel /a/ for five seconds without any interruption. The speech database has been developed by the "Institute of High-Performance Computing and Networking of the National Research Council of Italy (ICAR-CNR)" in collaboration with the hospital university of Naples "Federico II".

3.2 Proposed Architecture

A novel and robust architecture for detecting, filtering, and classifying a pathological voice signal is proposed herein.

Figure 1 highlights a graphical workflow utilizing five (5) different stages:

- Input sequence and pre-processing;
- Voice filtering and segmentation phase via a VAD module;
- Multiple feature extraction phase;

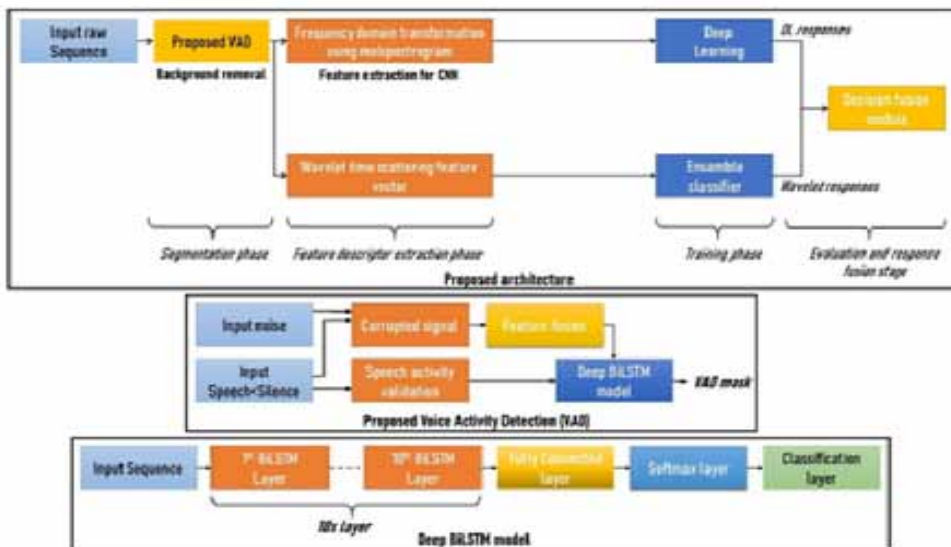
- Multiple training phases, and
- Decision Module.

The architecture utilizes two fusion modules, one of which is inside the proposed VAD module for fusing several feature descriptors prior to training the VAD-based filtering network while the second fusion module is applied at the decision module for integrating the varying-predicted-responses given by the DL model and the ensemble classifier. Owing to the complex nature of a pathological voice signal, it is quite important to identify and extract only the most useful signal component.

Simply, the speech sequence is first subjected to a VAD module, inside which lies a feature fusion module. Nine (9) different features from signals for training a bidirectional long-short term memory recurrent neural network (BiLSTM) are extracted. These varying types of features are then integrated before being used in the VAD’s training process. Subjecting a voice signal to this trained VAD network will help retain only the most useful component and will also pave the way for fruitful feature extraction in the remaining part of the main architecture.

Further, it can be seen from the topmost subfigure of **Figure 1** that the trimmed or filtered signal projecting from the VAD module is subjected to two different feature extraction modules. Here, two features are seen: (a) frequency domain transformation via Mel spectrogram; and (b) wavelet time scattering feature vector.

Figure 1. Graphical workflow of the proposed architecture and its sub-component.



The first set of features is trained via a DL network and the second set is trained via an ensemble classifier. The next phase is testing, during which an input signal is subjected to both the trained classifiers for producing two unique set responses, more specifically, a wavelet response, and DL response. As noted, the second fusion module (i.e. decision fusion module) is used for further fusing these two responses for providing a more robust classification.

The proposed architecture uses a DL-based VAD-based filter module, and learns a different kind of features using a decision-fusion of parametric and non-parametric classifiers.

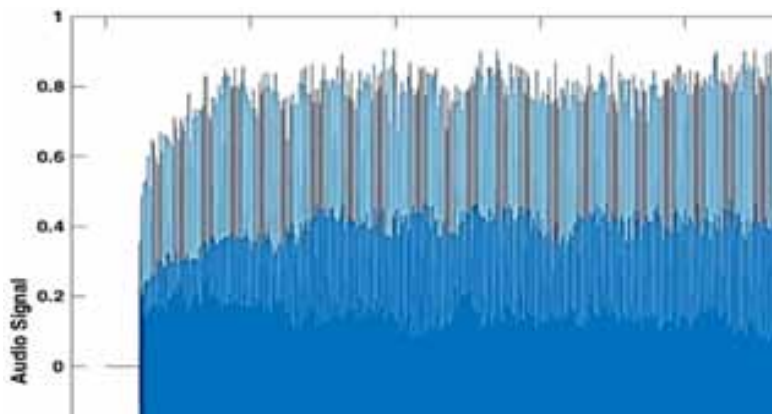
The detailed structure of all associated components is explained thoroughly in the following subsections.

3.3. Input Sequence and Pre-Processing

The pathological dataset (see *Section 3.1*) has an uneven number of recordings, that is, 150 and 58 for the pathological and healthy voice respectively. For the current work, 58 recordings from each of the two categories were selected so that training samples are evenly distributed during the training process. It is widely known that an uneven distribution of training samples can result in either overfitting or provide false-positive results. Considering both samples, a total of 116 voice recordings were selected for our analysis.

In each recording, the vowel sound \a\ can be heard to be pronounced by an individual for about 4.76 seconds at a sampling rate of 8000 Hz, with 38080 sampling points. **Figure 2** graphically illustrates the “voice003.dat” file derived from this dataset. All 116 samples were further broken up into ten (10) different segments (i.e. 0.476 second each). This kind of segmentation will safeguard any network against overfitting that may crop up due to lengthy training samples. Hence, the total number of resulting samples is 1160, in which 580 samples each belong to the healthy and pathological categories.

Figure 2. Sample number “voice003” pronunciation of the vowel \a\ for a duration of 4.76 seconds.



3.4 Voice Filtering-Segmentation Phase via a VAD Module

Past research has shown that using a VAD module is quite helpful although it is quite complex and challenging especially in low signal-to-noise ratio (SNR), or where noise is present in very less amount (Ramirez et al., 2004). Hence, a VAD-based speech detection and background (noise) signal filtering module that uses a fusion of several feature descriptors with a DL trainer is advocated herein. In a very low SNR environment, where the noise is present in extremely less amount, this approach has been found to be beneficial.

A long short-term memory (LSTM) type recurrent neural network (RNN) is suitable for the current application as it can study and remember long term dependencies of time-series data. While a LSTM layer observed only the future weights or the forward direction, a bidirectional LSTM (BiLSTM) can look into both the past and the future. A fourteen (14) layered BiLSTMDL network which uses 5000 number of hidden neurons is proposed for the training procedure. DL can assist in remedying the situation but there is need for high quality and a good number of training samples

especially for a speech-based network. A multi-pronged for extracting spectral features and harmonic ratio metric from an input signal is proposed here so that the network can learn different forms of the signal which will help in discerning noise from the good components. The second and third rows of **Figure 1** depict the simplified workflow of the proposed VAD module and the BiLSTMDL network. A detailed and sequential explanation of the procedure is then given.

- Dataset Preparation

The first stage is the dataset development stage. For this purpose, the 116 voice samples which is a collection of data from both healthy and pathological categories is used for representing the clean speech regions. A simple washing machine noise is used as the background noise. All the 116 voice samples are stitched-up together to create a long input audio sequence of length equal to $116 \times 4.76 \text{ seconds} = 552.16 \text{ seconds}$. A thresholding method is essential to filter out non-voice regions from this input clean audio. The regions of speech are annotated using a voice activity mask, and after which the second signal is used for corrupting the clean audio. The corrupted signal now has the useful regions being marked out. The process of thresholding and annotation are discussed as follows.

- Thresholding

The input speech is broken up into non-overlapping segmentations of window length (W) using relation, $W=L \times S$, where, L is the desired length in mS, and S is the sampling rate of the speech. The segments are then buffered as vectors. Then, the energy and spectral centroid associated with the segments are determined followed by normalization of the spectral centroid. The energy and spectral centroid are then smoothed by using two median filters consecutively, followed by the determination of the average mean of these smoothed components.

The next process is to determine the threshold of these two smoothed signal components consecutively using two common and simple steps: (i) Determining the histogram and its local-maxima; and (ii) If two local maxima values are uncovered, the threshold is the weighted average. In case these two conditions are met, then the presence of speech is confirmed. In this way, all the speech and non-speech regions in the entire clean audio length are determined. The final step in the thresholding process is to merge all the intersecting voice segments for removing the unwanted non-voice regions completely.

- Annotation of Voice Activity Mask

The new clean audio sequence developed after the thresholding operation is now labelled as L_2 . The speech and non-speech regions are assigned a binary mask, i.e. 1 and 0 respectively. The length of a silence or non-speech segment is now specified to be maintained at a maximum duration of 2 seconds.

A 10-seconds audio signal from the clean category showing the presence of non-speech regions is shown in **Figure 3**. VAD mask is allocated over the speech regions.

The next step process is to create a new voice activity sequence (L_3) by adding several random non-speech signal regions in L_2 , but not exceeding a duration of 2 seconds.

The final process is to create a 4th signal sequence (L_4) by adding the 8kHz washing machine noise signal to L_3 to create the corrupted dataset but useful components annotated by voice masks as shown in the third subfigure of **Figure 4**.

Figure 3. A 10-second audio signal from the clean category showing the presence of non-speech regions. Allocating VAD mask over speech regions.

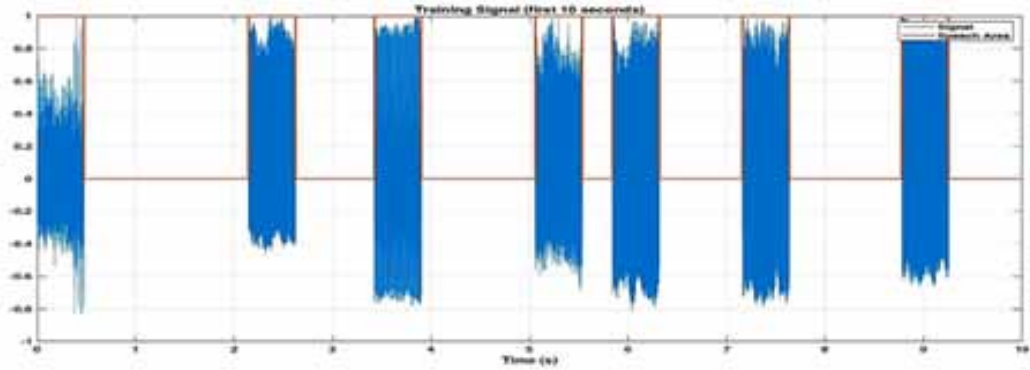
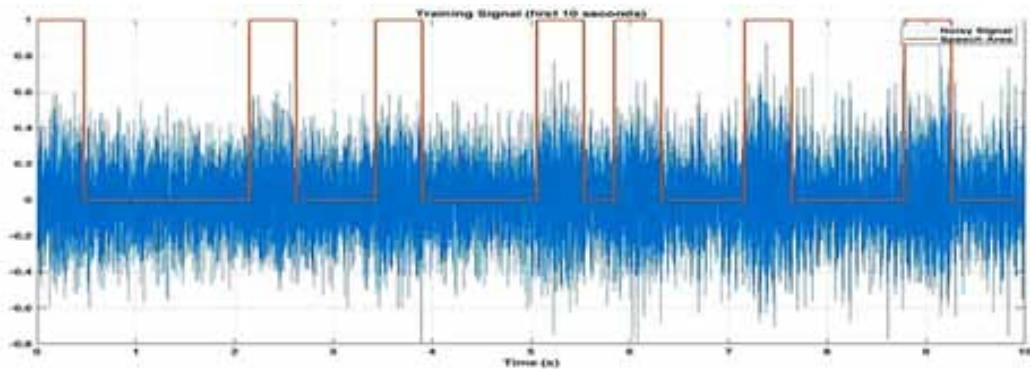


Figure 4. The ten seconds clean audio after being corrupted by noise. This signal uses a VAD mask to annotate the clean region.



3.5 Feature Fusion Module

It is responsible for extracting multiple type of features from the input signal followed by concatenation before subjecting it to a deep learning network. Nine (9) different type of features are considered for the current work as discussed below (Kim et al., 2006; Lerch, 2012; Peeters, 2004; Scheirer & Slaney, 1997; Smith, 2011).

- *Spectral Centroid*(SCen) is used for characterizing a spectrum and determines the position of the center of mass. It is computed by using a Fourier transform with the average of the frequencies present within the signal. The mathematical expression of a spectral centroid is

$$centroid = \frac{\sum_{k=b_1}^{b_2} f_k s_k}{\sum_{k=b_1}^{b_2} s_k} \quad (1)$$

where, f_k and s_k are the frequency and spectral value respectively for bin k, b_1 and b_2 are edges of band range within which spectral centroid is to be calculated (Peeters, 2004).

• **Spectral Crest (SCrest)** is mathematically expressed as

$$crest = \frac{\max(s_{k \in [b_1, b_2]})}{\frac{1}{b_2 - b_1} \sum_{k=b_1}^{b_2} s_k} \quad (2)$$

where, s_k is the spectral value respectively for bin k, b_1 and b_2 are edges of band range within which spectral crest is to be calculated (Peeters, 2004).

• **Spectral Entropy (SEn)** calculates the spectral power of an input signal using the Shannon entropy. It observes the normalized power of a signal as a statistical probability distribution. It can be mathematically expressed as

$$entropy = \frac{-\sum_{k=b_1}^{b_2} s_k \log(s_k)}{\log(b_2 - b_1)}, \quad (3)$$

where, s_k is the spectral value respectively for bin k, b_1 and b_2 are edges of band range within which spectral entropy is to be calculated (Peeters, 2004).

• **Spectral Flux (SF)** is used to observe the rate of change power spectrum between two adjacent frames. It is calculated by using the Euclidean distance metric of two adjacent and normalized spectra. It is mathematically expressed by

$$flux(t) = \left(\sum_{k=b_1}^{b_2} |s_k(t) - s_k(t-1)|^p \right)^{\frac{1}{p}}, \quad (4)$$

where, s_k is the spectral value respectively for bin k, b_1 and b_2 are edges of band range within which spectral flux is to be calculated, and the value of P will define the type of norm (Peeters, 2004).

• **Spectral Kurtosis (SKur)** is used for determining the occurrence and location of transient series in the frequency domain. It is also quite helpful in classifying power spectral density and capable of removing non-stationary information. It is expressed mathematically as

$$kurtosis = \frac{\sum_{k=b_1}^{b_2} (f_k - \mu_1)^4 s_k}{(\mu_2)^4 \sum_{k=b_1}^{b_2} s_k} \quad (5)$$

where, f_k and s_k are the frequency and spectral value respectively for bin k, b_1 and b_2 are edges of band-range within which spectral kurtosis is to be calculated. The terms μ_1 and μ_2 denote spectral centroid and spread respectively (Peeters, 2004).

- *Spectral Rolloff(SRoll)Point* is a quantity that is used for measuring the skewedness of a power spectrum. It extracts the spectral roll off point. It denotes the percentage of the power spectrum which is at lower frequency. It is expressed mathematically as

$$\text{SRollPoint} = I, \text{ such that } \sum_{k=b_1}^i s_k = K \sum_{k=b_1}^{b_2} s_k, \quad (6)$$

where, s_k is the spectral value for bin k , b_1 and b_2 are edges of band-range within which spectral spread is to be calculated. K denotes the above percentage between b_1 and i (Scheirer & Slaney, 1997).

- *Spectral Skewness(SSkew)*– The normalized skewness of a spectrum is the third central moment of this spectrum, divided by the 1.5 power of the second central moment. The spectral skewness of the signal over time can be expressed mathematically as

$$\text{skewness} = \frac{\sum_{k=b_1}^{b_2} (f_k - \mu_1)^3 s_k}{(\mu_2)^3 \sum_{k=b_1}^{b_2} s_k} \quad (7)$$

where, f_k is the frequency in Hz corresponding to bin k . s_k is the spectral value of bin k . b_1 and b_2 are the band edges, in bins, over which to calculate the spectral skewness. μ_1 and μ_2 are the spectral centroid and spectral spread respectively (Peeters, 2004).

- *Spectral Slope(SSlope)* is used to describe the amount of dependence that a reflectance can have upon the wavelength of a signal. It is also known by the term spectral gradient. It can be used for showing the level of affinity that an audio spectrum may have for high frequency signals. It is usually calculated through linear regression of a Fourier magnitude of a signal.

$$\text{slope} = \frac{\sum_{k=b_1}^{b_2} (f_k - \mu_f)(s_k - \mu_s)}{\sum_{k=b_1}^{b_2} (f_k - \mu_f)^2} \quad (8)$$

where, f_k and s_k are the frequency and spectral value respectively for bin k , b_1 and b_2 are edges of band range within which spectral slope is to be calculated. μ_s and μ_f are the mean frequency and spectral value respectively (Lerch, 2012).

- *Harmonic Ratio* is a ratio of the harmonic energy to the total energy of the audio. The autocorrelation is given by

$$\Gamma(m) = \frac{\sum_{n=1}^N s(n) s(n-m)}{\sqrt{\sum_{n=1}^N s(n)^2 \sum_{n=0}^N s(n-m)^2}} \text{ for } (1 \leq m \leq M) \quad (9)$$

where, s denotes an audio segment containing N number of elements, and the maximum lag is denoted by M .

For a given range, the maximum value of a normalized autocorrelation gives the desired harmonic ratio.

$$\beta HR = \max_{M_0 \leq m \leq M} \{\Gamma(m)\} \quad (10)$$

where M_0 is the minimum level of the searching region/range (Smith, 2011).

The parameters which are initialized for computing the above features are as follows. The length of the window is fixed at 256, and a periodic Hann (Hanning) window is used. The overlap length and the FFT length are set to 128 and 256 respectively. The range is maintained between 0 and half of the sampling rate. Finally, the power spectrum is used for extracting the above features. In the first step, a spectrogram which results in a short-time Fourier transform is used to compute the power spectrum of the noisy-training signal via a hann window of length 256 and an overlap length of 128. This operation results in a vector of frequencies along with their time instances. Next, the first eight (8) features are extracted via the information computed vis-à-vis equations E (1) through E (8). Let the extracted features be stored in the variables SCen, SCrest, SEn, SFl, SKur, SRoll, SSkew, SSlope. The next step is to use the noisy training signal, sampling rate, hann window with the fixed overlap length to extract the ninth (9th) feature descriptor, i.e. harmonic ratio. Let this feature vector be denoted by “ h_r ”. The final step in this feature fusion module (FFM) is to gather the descriptors extracted by all the nine (9) extractors into an array which is the required data that will be used to train a DL model defined below.

Thus, the training signal is given by:

$$\text{features} = [\text{SCen}, \text{SCrest}, \text{SEn}, \text{SFl}, \text{SKur}, \text{SRoll}, \text{SSkew}, \text{SSlope}, \text{Hr}] E \quad (11)$$

Let M and S denote the mean and standard deviation respectively. The new features may be obtained by normalizing the feature vector of E (12) using the following relation.

$$\text{feature} = \frac{(\text{feature} - M)}{S} \quad (12)$$

The concatenated and normalized features obtained in E (12) are used for training a DL network. The same process is also followed for extracting features from the validation dataset.

3.5 Deep BiLSTM Model

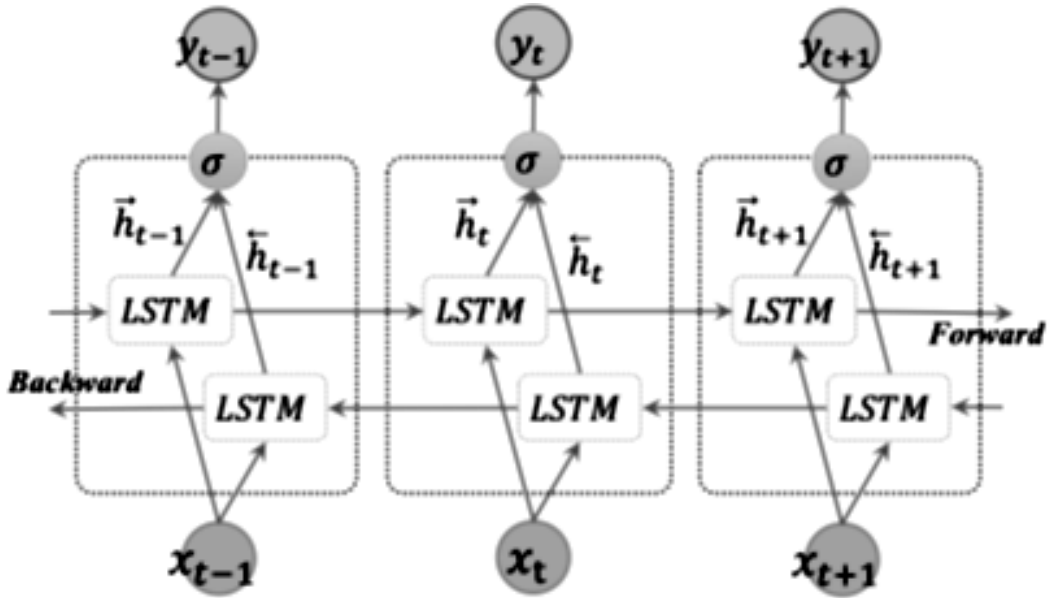
A DL model based on the LSTM network is used for differentiating between region of speech and background noise or silence. The LSTMs are a special type of DL architecture which has knowledge about weights that may further develop with a progress in time. It was first proposed in (Hochreiter & Schmidhuber, 1997), then upgraded and applied in various fields by researchers. A bidirectional LSTM network is where two LSTM layers moving in opposite direction is integrated together. Here, one of them moves in the forward direction and can use information from past to future, whereas the second one moves in the backward direction and can use information from future to past.

The BiLSTMs have shown more prowess for dealing with the task of sound classification in comparison with the unidirectional classifiers. However, it has been observed that BiLSTM network has not yet been implemented in voice pathology detection and classification (Glorot & Bengio,

2010). A BiLSTM-based architecture is used as the trainer for both the VAD module and the overall architecture for learning time-based dependencies of the voice signals.

A graphical structure of a BiLSTM network which has two LSTM layers pointed in the forward and backward direction is shown in **Figure 5**. In this network, a sequence running from time instance

Figure 5. Architecture of the BiLSTM layer used in the current work involving three consecutive steps.



T-n to T-1 is used to compute the output sequence \vec{h} of the forward layer iteratively while \overleftarrow{h} , which denotes the output sequence, uses the reverse inputs. These two outputs are determined using the basic equations of an LSTM network.

Let Y_T be an output vector realized from the BiLSTM layer whose contents are governed by the relation

$$y_t = \sigma(\vec{h}, \overleftarrow{h}) \quad (13)$$

where, the forward and backward output sequences are combined by the function σ . It can perform any function such as concatenation, multiplication, summation or mean. The output vector from a BiLSTM layer can also be expressed in the format $Y_T = [y_{T-n}, \dots, y_{T-1}]$ in a similar way as an LSTM layer (He et al., 2015).

The proposed deep BiLSTM model utilizes ten (10) BiLSTM layers with each of them having 300 numbers of hidden units for outputting sequences. The input length of the first BiLSTM layer can vary as per the application. For the current VAD network, the length is specified as 9. The 10th BiLSTM layer is connected to a fully connected layer having a size of 2 for classifying a voice as either pathological or healthy. It is finally followed by a softmax and a classification layer.

3.6 Feature Extraction via Mel-Spectrogram for the DL Classifier

To date, a vast majority of the algorithms that has been proposed for dealing with detection and classification of voice pathology used MFCC, or GMM. While some have used feature derived via zero crossing rate, SFI/SCen/SRoll, and linear prediction coefficients (LPC), Mel spectrogram (Rabiner & Schafer, 2010) is gaining popularity as an alternative feature descriptor for voice signal. An integration of this type of feature vector with a DL network is being investigated here.

- *Fourier spectral feature* is computed via a short-time Fourier transform whose energy $\left| x_{freq} [n, k] \right|^2$ can be expressed as -

$$\left| x_{freq} [n, k] \right|^2 = \left| \sum_{m=0}^{N-1} x(n-m)w(m) \exp \left(-j \frac{2\pi km}{N} \right) \right|^2 \quad (14)$$

where $x(n)$ is the discrete-time input speech sequence, N is the size of the window, and $w(m)$ is the sequence within the window.

From E (14), it can be further gathered that $\left| x_{freq} [n, k] \right|^2$ is a double-indexed function with time index n and frequency index k . Usually, the short-time framed Fourier energy $E_{FT}[n', k]$ will be collected for $n = 0, \Delta n, 2\Delta n, \dots, n' \Delta n$, and $n' \in Z^+ \cup \{0\}$ such that

$$E_{FT}[n', k] = \left| x_{freq} [n' \Delta n, k] \right|^2, \text{ for } n' \in Z^+ \cup \{0\}, \quad (15)$$

where $\Delta n > 0$ is the frame advance step size.

- *Mel spectral features* are finally obtained via the above Fourier spectral features with a Mel filter bank over a non-linear frequency scale, which is also commonly termed as Mel-scale. They can be acquired through the weighted Fourier spectral features via the Mel filter bank which is a uniformly spaced filter bank on a nonlinearly wrapped frequency scale, known as the Mel-scale, as illustrated in **Figure 7**.

An expressing comprising of both Mel-scale frequency (f_{mel}), and traditional frequency (f_{con}) in hertz can be stated as –

$$f_{mel} = 2595 \log \left(1 + \frac{f_{con}}{700} \right), \quad (16)$$

Without loss of generality, we chose a 128-band Mel filters throughout this work.

As depicted in **Figure 10**, the squared magnitude response of the i^{th} Mel filter, $\left| H_{mel}(i, k) \right|^2, 0 \leq k \leq N-1, 1 \leq i \leq 128$,

specifies the individual weighting factor for the k^{th} frequency component of the Fourier spectra.

According to E (15) and E (16), the short-time framed Mel-energy $E_{mel}(n', i)$ is given by

Figure 6. Block diagram of mel spectrogram (Rabiner & Schafer, 2010).

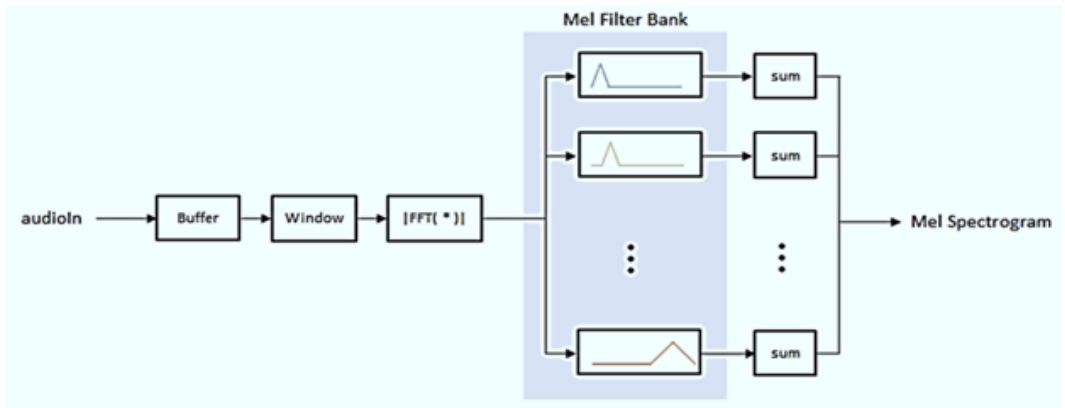
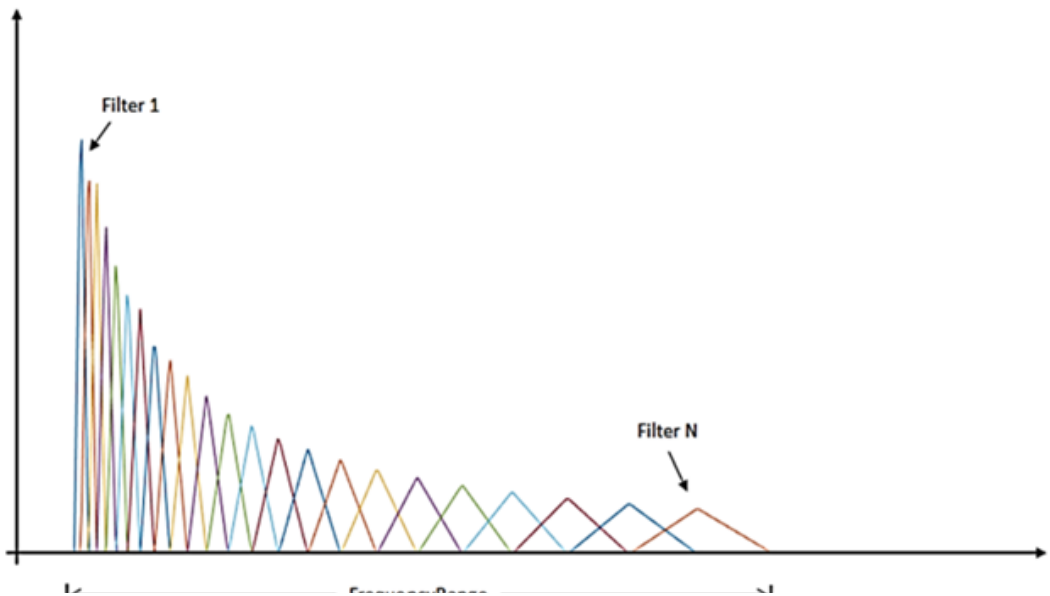


Figure 7. A Mel filter bank composed of the triangular band-pass filters each with a bandwidth and the spacing in accordance with the Mel-scale in the frequency domain (Rabiner & Schafer, 2010).



$$E_{mel}(n', i) = \sum_{k=0}^{N-1} E_{FT}[n', k] |H_{mel}(i, k)|^2, 1 \leq i \leq 128 \quad (17)$$

where, i is the Mel filter index. It is noted that the frequency dimensionality of $E_{mel}(n', i)$ is reduced from N to $E_{FT}[n', k]$ to 20.

The input voice signal segments, which are duly filtered by the VAD module, are initially buffered as frames or segments according to the specified length of the periodic hanning window. Adjacent segments are made to overlap one another by certain specified number of durations. After overlapping, the segments are transformed into the frequency domain with the number of points being deduced by the Fast Fourier Transform (FFT). According to the type of spectrum, power or magnitude can represent the frequency domain. The frequency-domain transformed signal sequences are now passed through a Mel filter bank, and the resulting spectral output terms are summed up. The channels are finally concatenated so that they are now transformed into a feature vector ready to undergo training.

- **Filter bank design** – The Mel filter bank used in our study has triangular filters which are made to overlap by 0.5 or half and spaced uniformly over the Mel frequency. As **Figures 6 and 7** show, the variable N is set as 128, implying that 128 numbers of Mel bandpass filters are used. Two (2) frequencies have been allocated to the first and the last filters to serve as a cut-off frequency of the filter bank. The filters are then normalized according to their available bandwidth.

3.7 DL Training Network

The Mel-spectrogram features, as extracted from the training samples, will be used for training a 33-layered DL network.

A repeated structure comprising eight (8) vanilla-CNN modules will be used for extracting deep features, which implies 8 convolutional layers with a 3x3 kernel, 1x1 stride sizes, and same padding arrangement. As **Table 2** shows, the number of filters or activation maps increases from 32 to 256 so that a vast number of multiple features can be learned. Three 3x3 max-pooling layers having 2x2 stride and with the same padding convolutional kernels are used after every two convolutional layers for reducing overfitting and speeding up the training by downsizing the number of training parameters. A 50% dropout rate is utilized to avoid overfitting the network with excess features. The final layers of the network consist of (a) a fully connected layer trained to recognize two classes, (b) a softmax layer, and (c) a cross entropy output layer.

3.8 Feature Extraction for the Ensemble Classifier

Wavelet features are extracted from all the training samples via the following procedure.

Initially, a filtered sample signal x is taken whose Fourier transform \hat{x} is given by

$$\hat{x}(w) = \int_R x(u) e^{-i w u} du \quad (18)$$

For a filter Ψ whose Fourier transform $\hat{\Psi}$ is located around the dimensionless frequency of 1, the wavelet filter bank which can be denoted by $\{\Psi_\lambda\}_{\lambda>0}$ is determined by dilating the mother wavelet Ψ and can be expressed as

$$\Psi_\lambda(t) = \lambda \Psi(\lambda t) \quad (19)$$

Table 2. The proposed deep learning network structure for training the mel-spectrogram features.

Sl.	Name Features	Type
1	imageinput	Image Input
2	batchnorm_1	Batch Normalization
3	conv_1 32 3x3x1 convolutions with stride [1 1] and padding 'same'	Convolution
4	batchnorm_2	Batch Normalization
5	relu_1	ReLU
6	conv_2 32 3x3x32 convolutions with stride [1 1] and padding 'same'	Convolution
7	batchnorm_3	Batch Normalization
8	relu_2	ReLU
9	maxpool_1 3x3 max pooling with stride [2 2] and padding 'same'	Max Pooling
10	conv_3 64 3x3x32 convolutions with stride [1 1] and padding 'same'	Convolution
11	batchnorm_4	Batch Normalization
12	relu_3	ReLU
13	conv_4 64 3x3x64 convolutions with stride [1 1] and padding 'same'	Convolution
14	batchnorm_5	Batch Normalization
15	relu_4	ReLU
16	maxpool_2 3x3 max pooling with stride [2 2] and padding 'same'	Max Pooling
17	conv_5 128 3x3x64 convolutions with stride [1 1] and padding 'same'	Convolution
18	batchnorm_6	Batch Normalization
19	relu_5	ReLU
20	conv_6 128 3x3x128 convolutions with stride [1 1] and padding 'same'	Convolution
21	batchnorm_7	Batch Normalization
22	relu_6	ReLU
23	maxpool_3 3x3 max pooling with stride [2 2] and padding 'same'	Max Pooling
24	conv_7 256 3x3x128 convolutions with stride [1 1] and padding 'same'	Convolution
25	batchnorm_8	Batch Normalization
26	relu_7	ReLU
27	conv_8 256 3x3x256 convolutions with stride [1 1] and padding 'same'	Convolution
28	batchnorm_9	Batch Normalization
29	relu_8	ReLU
30	dropout 50%	Dropout
31	fc 2 fully connected layer	Fully Connected
32	softmax	Softmax
33	class output cross entropy	Classification Output

The property of the filter Ψ is such that for negative frequencies, its Fourier transform is zero, which shows that it is analytic. Similarly, another filter $\hat{\Psi}$ revolving around 1 may be considered, and thus its wavelet filter bank $\hat{\Psi}_\lambda$ is also located around λ .

The pseudo-analytic Gammatone wavelet is compatible for the current case of Ψ , and is mathematically given by

$$\Psi(t) = ((n-1)t^{n-2} + it^{n-1})e^{-(b+i)t}1_{[0,\infty)}(t) \quad (20)$$

where b is the bandwidth which is approximately equal to $2^{-1/Q}$, Q is the quality factor whose value can be chosen as desired. Initially, the value of Q is chosen as 4.

Let x be any input signal decomposable via a wavelet filter bank to produce $x * \Psi_{\lambda_1}(t)$, for $\lambda_1 > 0$, here $*$ is a standard convolution operation. This process is called the wavelet decomposition of the input signal x .

The requirement to sample λ_1 regularly does not arise due to the dilation arrangement of the wavelet filter bank. Instead, it may be sampled using the relation $2^{j/Q}$, where Q is the mother wavelet Ψ 's Q -factor. In other words, sampling can be done uniformly using the $\log(\lambda_1)$.

A wavelet scalogram is thus created in this process which is expressed mathematically using the relation

$$x_1(t, \log \lambda_1) = |x * \Psi_{\lambda_1}(t)| \quad (21)$$

Unwanted time-shifting and warping during decomposition of wavelet is reduced by using a complex modulus. The wavelet scalogram depicts the frequency and time contained within a signal quite nicely. E (22) means that a wavelet scalogram for a combination of $(t, \log \lambda_1)$ can be obtained by using the information of x at the $\log \lambda_1$ and time t . Additionally, for imparting further stability and invariance, the scalogram is averaged in time by using $\phi_T(t)$, which is a low pass filter having a duration time T . Therefore,

$$S_1 x(t, \log \lambda_1) = x_1(\cdot, \log \lambda_1) * \phi_T(t) = |x * \Psi_{\lambda_1} | * \phi_T(t)| \quad (22)$$

For the current implementation, a Gabor filter is used as the low pass filter $\phi_T(t)$ whose frequency is centered at 0. T is the bandwidth in time. S_{1x} gives the coefficient of the first order wavelet scattering, which is like the coefficients of a Mel spectrogram. However, the weighted mean of $\phi_T(t)$ subtracted the fine-scaled temporal arrangement of the x_1 . To remedy this situation, a second wavelet decomposition is performed by using $Q=1$. The new expression is

$$x_2(t, \log \lambda_1, \log \lambda_2) = x_1(\cdot, \log \lambda_1) * \Psi_{\lambda_2}(t) = ||x * \Psi_{\lambda_1} | * \Psi_{\lambda_2}(t)| \quad (23)$$

E (23) is the new scalogram (x_2) but of second-order centered around $\log \lambda_1$ of x_1 . Similarly, for enhancing stability and invariance, the average of x_2 is performed in time using the same filter, and thereby resulting in

$$S_2 x(t, \log \lambda_1, \log \lambda_2) = x_2(\cdot, \log \lambda_1, \log \lambda_2) * \phi_T(t) = ||x * \Psi_{\lambda_1} | * \Psi_{\lambda_2} | * \phi_T(t)| \quad (24)$$

The above procedure can be performed successively for determining the third or higher order coefficients, however, second order is sufficient.

Now, all the resulting coefficients of the first-order are concatenated together into a single vector.

$$S_1 x(t) = \{S_1 x(t, \log \lambda_1)\}_{\lambda_1 > 0}, \quad (25)$$

The same is done for the coefficients belonging to second-order.

$$S_2 x(t) = \{S_2 x(t, \log \lambda_1, \log \lambda_2)\}_{\lambda_1 > 0, \lambda_2 > 0}, \quad (26)$$

The final step is to combine the vectors in E (25) and E (26) into a single vector for time t .

$$Sx(t) = \{S_1 x(t), S_2 x(t)\} \quad (27)$$

We use wavelet scattering for extracting feature descriptors from the 928 training sound signals. We use an invariance scale of 0.5 for all the training samples.

3.9 Ensemble Classifier via Random Subspace

The ensemble classifier is trained via the random subspace technique using very limited memory while providing a better technique than other ensemble techniques. It uses some important parameters such as m, d, x and n .

Here, m stands for the number of dimensions for sampling the learners, x is the feature vector matrix whose dimension is d , and finally, n is a number representing the quantity of learners in the ensemble classifier.

There are four simple steps for training it. Firstly, a set containing m number of predictors are chosen randomly from d number of possible values. Secondly, a weak learner is trained by utilizing the m predictors. Thirdly, the above two steps are repeated until n number of weak learners have been created. The final classification or prediction score is obtained by averaging all the scores.

3.10 Decision Fusion Module

The decision fusion module is applied at the testing stage.

Using the test samples, two sets of feature vectors are extracted via Mel-spectrogram and wavelet scattering coefficient. The first set made up of Mel-spectrogram features are passed into the trained convolutional neural network (CNN) for extracting prediction scores whereas the second set is passed into the trained ensemble classifier. This process has resulted in the creation of two unique sets of prediction scores of the same test sample database. Here, a healthy test sample can be classified as pathological with a certain probability by the ensemble category, while the same sample can be correctly classified by the DL network. There may be many such occurrences, and therefore, a fusion of all the prediction responses derived from the two trained networks maybe combined to create a new prediction response.

Let W_R and D_R be the responses produced by the ensemble and DL classifiers respectively. The new response can be computed using the expression

$$\text{fused} = W_R \otimes D_R, \quad (28)$$

where, \otimes is the element-wise matrix multiplication operator.

4. RESULTS & DISCUSSION

Below, we present the implementation of the proposed work.

4.1 The Voice Activity Detector (VAD)

As noted previously, the length of the clean audio signal that is masked by 1 is 552.16 seconds, which is obtained by combining 116 samples. All the samples are broken up to ten (10) segments each to create 1160 samples to prevent overfitting during training. Roughly 80% or 920 samples were retained and then combined to create a $920 \times 0.476s = 437.92$ seconds long clean audio signal for training. The remaining 20% or 240 samples were retained and then combined to create a $240 \times 0.476s = 114.24$ seconds long test signal from which noise was to be removed.

The training was performed using the following set of parameters. In order to make the network iterate 100 rounds via the training data, the number of epochs was set as 100. The number of minibatch size was maintained at 64. The training segments were made to shuffle every epoch, and a piecewise learning rate schedule was employed for decreasing the learning rate by a factor of 0.1 after an elapse of 10 epochs. The adaptive moment estimation (ADAM) optimizer was used as it was more compatible with a RNN than the stochastic gradient descent (SGD) optimizer.

Figure 8 shows the convergence of accuracy with respect to validation accuracy, and the training loss with respect to the validation loss. The test sample of length 114.24 seconds is tested against the trained network, yielding an accuracy of 87.8% as shown by the confusion matrix of **Figure 9**. The result of testing a random 50-second portion of the validation signal is shown in **Figure 10**, where the background noise component is represented by blue color, and the speech region by red color. This trained VAD network will serve as a filter in the main architecture for safeguarding against unwanted noise or background components that may be present in the voice signal (whether healthy or pathological).

4.2 Classification Metrics

A comparison with recent analytic work using the same dataset is hereby presented. While six (6) metrics for classification have been suggested, we have used nine (9) metrics, which included popular ones such as mean average precision, sensitivity, specificity, F1 score, accuracy, and error.

Other popularly used metrics have also been considered, including false positive rate (FPR), Matthews Correlation Coefficient (MCC), and the Cohen's Kappa index (CKI) with brief summaries given below. Here, TP, TN, FP, FN stands for true positive, true negative, false positive, and false negative respectively.

- *Sensitivity*, also known as true positive rate (TPR), describes the actual number of positive samples as belonging to the true category.

$$Sensitivity = \frac{TP}{TP + FN} \quad (29)$$

- *Specificity*, also known as true negative rate (TNR), describes the actual number of negative samples as belonging to the negative category.

$$Specificity = \frac{TN}{TN + FP} \quad (30)$$

- *Precision* is a probability by which a network can make true positive classification.

$$Precision = \frac{TP}{TP + FP} \quad (31)$$

Figure 8. Training of the proposed VAD module with the BiLSTM network

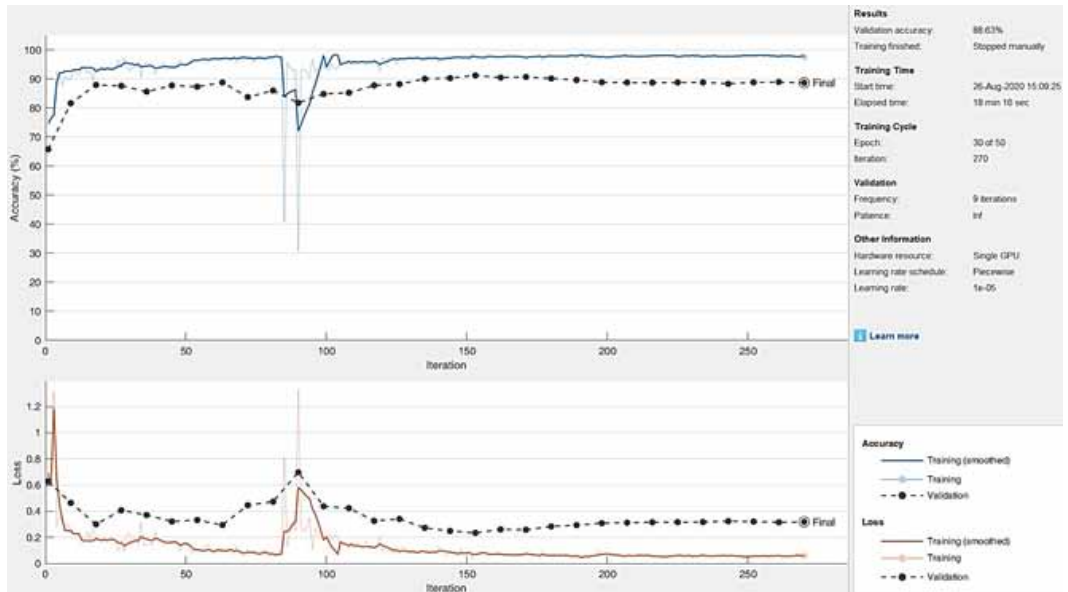


Figure 9. Confusion matrix showing classification of speech and non-speech regions in the input test signal. Here, 1 and 0 denotes the speech and non-speech regions respectively.

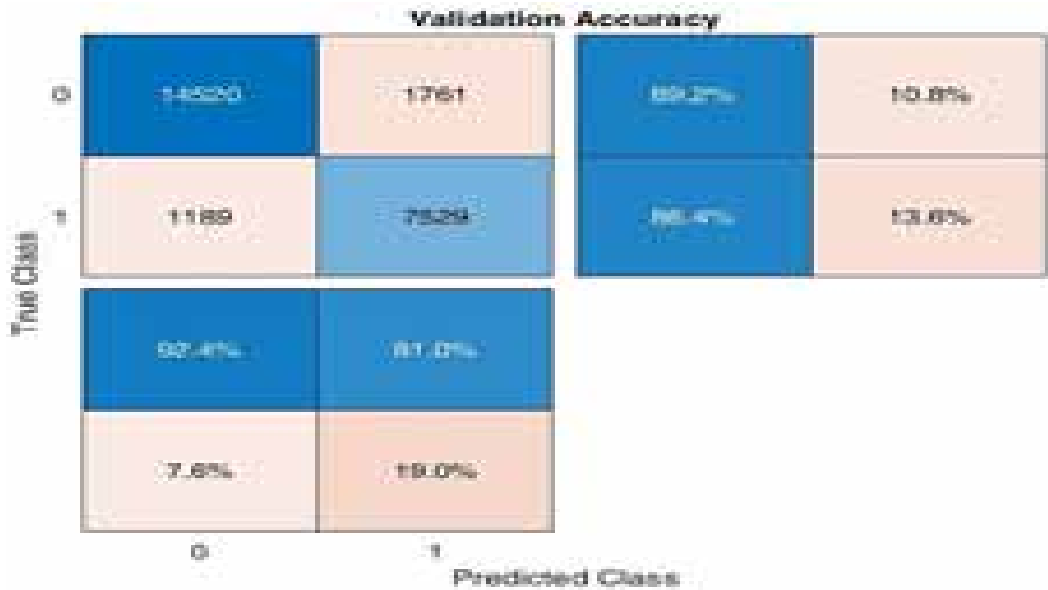
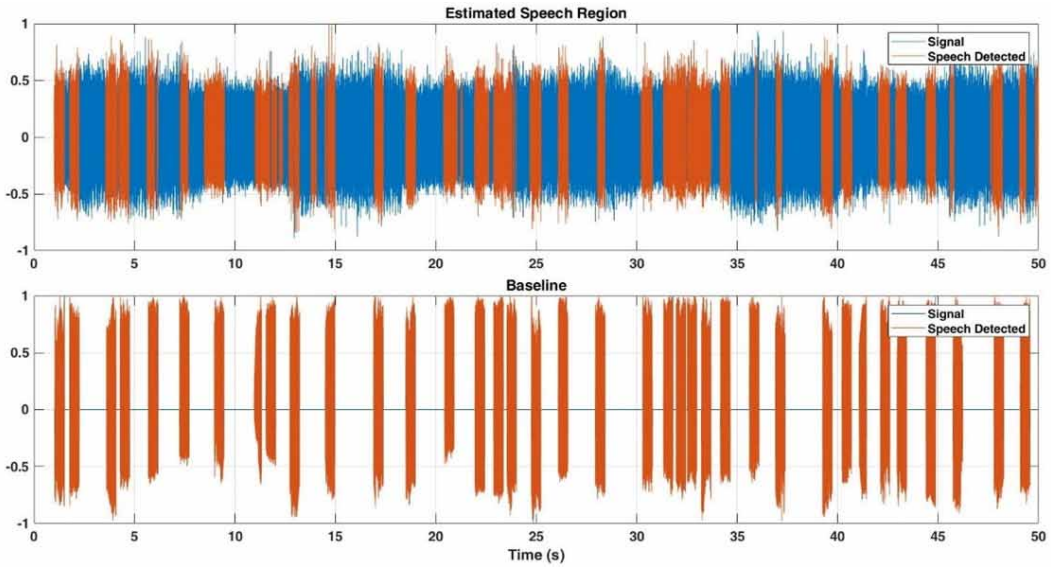


Figure 10. The top image shows the speech signals (red colour) being detected in the presence of noise (blue colour). The second image is the extraction of the baseline-speech signal from the noise corrupted signal.



- F_1 score is a harmonic mean of precision which is also known as Dice similarity coefficient (DSC).

$$F_1 = \frac{2TP}{2TP + FP + FN} \quad (32)$$

- FPR is the probability with which a trained network is capable of false rejecting a sample during test.

$$FPR = \frac{FP}{FP + TN} \quad (33)$$

- MCC , providing an index of a binary classification's quality, takes into consideration TPs, TNs, FPs, FNs, and is mostly highly regarded for being able to give good classification even in situation where classes are unbalanced.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (34)$$

- CKI measures the intra- and inter- rater's ability to provide good classification for categorical items. It is more robust than simple accuracy or most of the aforementioned metrics as it takes into account the chance of an agreement being reached by chance.

$$CKI = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}, \quad (35)$$

where, p_o and p_e are the observed and expected agreements respectively.

- *Accuracy (Acc.)* demonstrates the level of comfort with which a trained classifier can predict positive and negative class. It is determined by dividing the sum of TP and TN by the total population.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (36)$$

- *Error* refers to the expression for error:

$$Error = 100 - Acc \quad (37)$$

4.3 Results Implementation

Altogether, 1160 of samples whose length are all 0.476 seconds were available for both training and testing as noted earlier. 80% or 928 records were set aside for training which includes 464 healthy and pathological samples each whereas the remaining 20% or 232 records, which includes 116 healthy and pathological test samples were kept for the testing the trained network.

The first task was to create a new healthy and pathological audio signal database by using all the 1160 samples with the trained VAD network. For all the 928 training samples, an overlap index of 0.5 was used to extract a 128-bin Mel spectrogram. This overlapping arrangement is beneficial on two accounts. First, the dimension of the feature vector was further reduced, and second, the spectral property was also maintained. The window size was fixed at 1024 with 512 points hopping size.

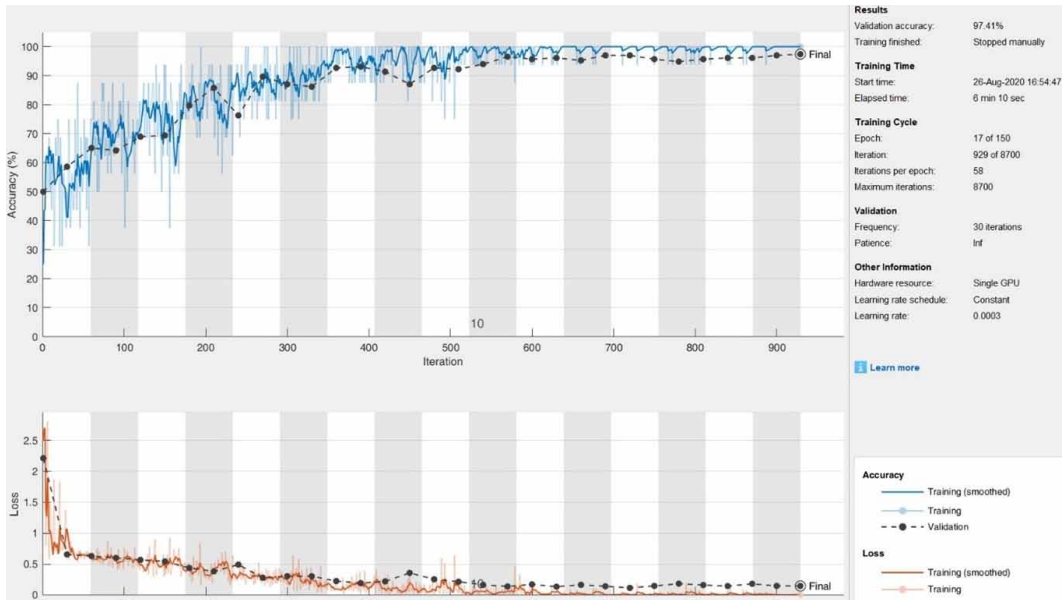
Additionally, an overlap of 512 per window, and 1024 points were used for the FFT. The Mel-spectrogram obtained was expressed in the logarithmic scale and normalized in the same way as in E (12). The resulting normalized feature vectors were made available for training the DL network. The network was then trained to learn and classify between healthy and pathological voice samples via the following parameters. The number of minibatch size was set as 32, maximum number of epochs was set as 150 but it was made to stop manually depending on the convergence of the training-progress.

Additional training parameters include the learning rate, which was maintained at a constant rate of 0.0003, and the use of stochastic gradient descent with momentum (SGDM) optimizer. The training progress of this Mel-spectrogram based DL network is shown in **Figure 11**.

The training was conducted smoothly with the accuracy converged to 100% after only 17 epochs; therefore, the training was made to stop manually. The next step is to train the ensemble classifier via the wavelet scattering features.

In order to implement the decision fusion module, it was desirable that another classifier be trained, but using a different set of feature extraction, specifically, using the wavelet scattering with an invariance scale of 0.5 features that were extracted from all the 928 training samples. The number of wavelet features which were deduced from the wavelet scattering coefficient that could be used for training was approximately 170 wavelet features per second, which amounted to a total of $928 \times 170 = 157760$ features. These features were used together with the random subspace method based ensemble classifier by fixing the number of learning cycle as 50.

Figure 11. Training progress showing the convergence of accuracy and loss.



The confusion matrices for testing the samples against the three trained networks are shown in **Figure 12a,b,c**. In **Figure 12a**, the trained-CNN classifier unveiled classifying 113 healthy v. pathological samples out of 116 test samples correctly, thereby providing a mean accuracy of 97.41%. The confusion matrix of **Figure 12b** reveals that out of 116, 115 healthy test samples are classified correctly. The same applies to 108 pathological voice test samples out of 116.

In comparison to the ensemble classifier, the CNN network gave a consistent performance over all the classification metrics, that is, 97.4% mean accuracy each for healthy v. pathological categories respectively. It also displayed consistent values over important metrics such as mean accuracy, sensitivity, specificity, precision and F1 score. However, there is still room for improvement as demonstrated by the values of FPR, MCC and CKI. Conversely, the proposed ensemble-CNN decision fusion technique yielded a comparatively higher mean accuracy of 99.14%. The remaining eight (8) metrics also demonstrated higher values. A combination of prediction responses obtained by testing the test samples against each of the two trained classifiers is thus capable of boosting the true positive classification rate.

Table 3 highlights a comparison of the performance achieved by the proposed fusion model, ensemble, and CNN classifiers' prediction scores. Out of the three trained classifiers, the ensemble classifier received the lowest mean accuracy, that is, 96.12%. The performance on the same dataset by a deep neural network (DNN), SVM, and RF classifiers which were obtained by authors of (Chen & Chen,) has also been recorded in this table. They used six(6) classification metrics, including accuracy, error, sensitivity, specificity, precision, and F1 score but left out three popular and important metrics which should also be used for judging network predictions, namely, FPR, MCC, and Cohen's kappa index.

Nevertheless, the results via three (3) different state-of-the-classifiers have yielded considerable accuracy. Specifically, the SVM classifier provided a high accuracy of 92.9% in comparison to the RF which yielded only 90.3% accuracy. These two techniques have shown somewhat similar performance in terms of F1 score and specificity. However, using a stacked auto-encoder based DL approach, the researchers achieved an astounding performance over their SVM and RF implemented counterparts.

Figure 12. Confusion matrix resulting from the testing of the a) trained CNN, b) ensemble classifier, and c) Proposed decision fusion.

True Class	Healthy	113	3	97.4%	2.6%
	Pathological	3	113	97.4%	2.6%
		97.4%	97.4%		
		2.6%	2.6%		
		Healthy	Pathological	Predicted Class	

Figure 12b. Confusion matrix resulting from the testing of the trained ensemble classifier

True Class	Healthy	115	1	99.1%	0.9%
	Pathological	8	108	93.1%	6.9%
		93.5%	99.1%		
		6.5%	0.9%		
		Healthy	Pathological	Predicted Class	

Figure 12c. Confusion matrix resulting from the testing of the trained Proposed decision fusion.

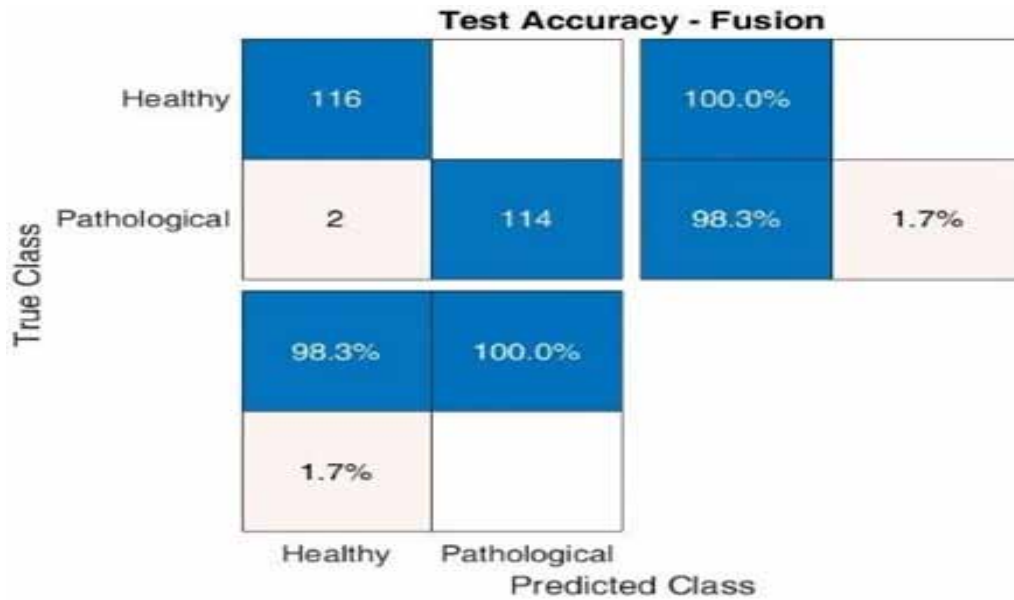


Table 3. Tabulation of the performance shown by various techniques and metrics. All values are recorded in percentage (%).

Method	Acc.	Error	Sen.	Spe.	Pre.	FPR	F1	MCC	CKI
Ensemble	96.12	0.38	99.14	93.10	93.50	6.90	96.23	92.41	92.24
CNN	97.41	2.59	97.41	97.41	97.41	2.59	97.41	94.83	94.83
Proposed fusion model	99.14	0.86	1	98.28	98.31	1.72	99.15	98.29	98.28
DNN (Chen & Chen,)	98.6	1.4	97.8	99.4	99.4	-	98.4	-	-
SVM (Chen & Chen,)	92.9	7.1	91.6	94.4	93.6	-	92.2	-	-
RF (Chen & Chen,)	90.3	9.7	90.3	94.1	93.8	-	92	-	-

Our proposed approach has nonetheless outperformed this DNN technique in seven (7) out of nine (9) metrics by yielding a mean accuracy of 99.14% in comparison to their 98.6%.

5. CONCLUSION

This work investigates improving the accuracy of the diagnosis of voice pathology in search of more robust solutions. The key contribution of this work is the deployment of a novel architecture to initially perform a deep learning-based filtering of the input voice signal followed by a decision-level fusion of deep learning and a non-parametric learner to provide highly precise voice pathology detection results. The efficacy of the proposed technique is verified vis-a-vis results of recently performed research on the same dataset but based on different training algorithms. Our results show that the use of ML classifier can reach up to 96.12% accuracy whereas the proposed fusion model with good selection of different features, filter, and the integration of DL and non-parametric learner provided the highest accuracy of 99.14%. The proposed approach has successfully produced remarkable and

convincing results that may benefit future researchers and practitioners in attempting to detect voice pathology non-invasively.

Notwithstanding, several limitations should be noted. First, the amount of data needed to train the model can be a challenge to this sort of research. While the PhysioNet's VoiceICar fEDerico II database may be comprehensive in terms of the number of people being recorded, yet in term of the number of samples of pathological patients, fewer samples of healthy people than desired were available. The statistically uneven distribution of individual pathologies is another issue, making the identification of voice pathology a complex issue. Finally, even though countermeasures to balance the classes with sample weights have been taken, we did not conduct our experiments separately on subsets of data for different genders.

Future work should consider the extraction of enhanced dimensions of the dataset and embedded quality attributes, including a new vowel combination and gender separation. In addition, the voice pathology identification technique may be further improved by reviewing various forms of CNN and training models. We believe that the use of DL methods for novelty detection, such as deep autoencoder, for modeling the normophonic voice can be an interesting idea for future investigation with prospect to identify challengingly disordered voices that are sparsely distributed across databases. Importantly, our work implies that the next step towards the goal of computerized acoustic analysis of voice signals can provide clinicians with fast, supportive methodology applicable to various state-of-the-art ML algorithms for massive datasets that could benefit from automated voice pathology detection.

ACKNOWLEDGMENT

The authors are thankful to the Special Manpower Development Program, Chip-to-System Design (SMDP-C2SD), initiated by the Ministry of Electronics & Information Technology (MeitY), Govt. of India, for providing research facilities in the School of VLSI Design and Embedded Systems, NIT, Kurukshetra.

REFERENCES

- Al-nasheri, A., Muhammad, G., Alsulaiman, M., & Ali, Z. (2017). Investigation of Voice Pathology Detection and Classification on Different Frequency Regions Using Correlation Functions. *Journal of Voice*, *31*(1), 3–15. doi:10.1016/j.jvoice.2016.01.014 PMID:26992554
- Al-nasheri, A., Muhammad, G., Alsulaiman, M., Ali, Z., Mesallam, T. A., Farahat, M., Malki, K. H., & Bencherif, M. A. (2017). An Investigation of Multidimensional Voice Program Parameters in Three Different Databases for Voice Pathology Detection and Classification. *Journal of Voice*, *31*(1), 113–118. doi:10.1016/j.jvoice.2016.03.019 PMID:27105857
- Ali, Z., Alsulaiman, M., Muhammad, G., Elamvazuthi, I., Al-nasheri, A., Mesallam, T. A., Farahat, M., & Malki, K. H. (2017). Intra- and Inter-database Study for Arabic, English, and German Databases: Do Conventional Speech Features Detect Voice Pathology? *Journal of Voice*, *31*(3), 386. doi:10.1016/j.jvoice.2016.09.009 PMID:27745756
- Benmalek, E., Elmhamdi, J., & Jilbab, A. (2017). Multiclass classification of Parkinson's disease using different classifiers and LLBFS feature selection algorithm. *International Journal of Speech Technology*, *20*(1), 179–184. doi:10.1007/s10772-017-9401-9
- Boyanov, B., & Hadjitodorov, S. (1997). Acoustic analysis of pathological voices: A voice analysis system for the screening and laryngeal diseases. *IEEE Engineering in Medicine and Biology Magazine*, *16*(4), 74–82. doi:10.1109/51.603651 PMID:9241523
- Boyanov, B., Ivanov, T., Hadjitodorov, S., & Chollet, G. (1993). Robust hybrid pitch detector. *Electronics Letters*, *29*(22), 1924–1926. doi:10.1049/el:19931281
- Cesari, U., De Pietro, G., Marciano, E., Niri, C., Sannino, G., & Verde, L. (2018). A new database of healthy and pathological voices. *Computers & Electrical Engineering*, *68*, 310–321. doi:10.1016/j.compeleceng.2018.04.008
- Chen, L., & Chen, J. (n.d.). Deep Neural Network for Automatic Classification of Pathological Voice Signals. *Journal of Voice*. PMID:32660846
- Cordeiro, H., Fonseca, J., Guimarães, I., & Meneses, C. (2017). Hierarchical Classification and System Combination for Automatically Identifying Physiological and Neuromuscular Laryngeal Pathologies. *Journal of Voice*, *31*(3), 384.e9–384.e14. Advance online publication. doi:10.1016/j.jvoice.2016.09.003 PMID:27743845
- Ebihara, S., & Ogawa, S. (1986). Normalized noise energy as an acoustic measure to evaluate pathologic voice. *The Journal of the Acoustical Society of America*, *80*(5), 1329–1334. doi:10.1121/1.394384 PMID:3782609
- Fonseca, E.S., Guido, R.C., Scalassara, P.R., Maciel, C.D., & Pereira, J.C. (2007). Wavelet time-frequency analysis and least squares support vector machines for the identification of voice disorders. *Computer Biology Med.* .10.1016/j.combiomed.2006.08.008
- Gavidia-Ceballos, L., & Hansen, J. H. L. (1996). Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection. *IEEE Transactions on Biomedical Engineering*, *43*(4), 373–383. doi:10.1109/10.486257 PMID:8626186
- Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010*, 249–56.
- Godino-Llorente, J. I., Fraile, R., Sáenz-Lechón, N., Osma-Ruiz, V., & Gómez-Vilda, P. (2009). Automatic detection of voice impairments from text-dependent running speech. *Biomedical Signal Processing and Control*, *4*(3), 176–182. doi:10.1016/j.bspc.2009.01.007
- Godino-Llorente, J.I., & Gómez-Vilda, P. (2004). Automatic Detection of Voice Impairments by Means of Short-Term Cepstral Parameters and Neural Network Based Detectors. *IEEE Trans Biomed Eng.*, *51*(2), 380–84. https://.10.1109/TBME.2003.820386
- Godino-Llorente, J. I., Osma-Ruiz, V., Sáenz-Lechón, N., Gómez-Vilda, P., Blanco-Velasco, M., & Cruz-Roldán, F. (2010). The Effectiveness of the Glottal to Noise Excitation Ratio for the Screening of Voice Disorders. *Journal of Voice*, *24*(1), 47–56. doi:10.1016/j.jvoice.2008.04.006 PMID:19135854

- Hadjitodorov, S., Boyanov, B., & Teston, B. (2000). Laryngeal pathology detection by means of class-specific neural maps. *IEEE Transactions on Information Technology in Biomedicine*, 4(1), 68–73. Advance online publication. doi:10.1109/4233.826861 PMID:10761776
- Hadjitodorov, S., & Mitev, P. (2002). A computer system for acoustic analysis of pathological voices and laryngeal diseases screening. *Medical Engineering & Physics*, 24(6), 419–429. doi:10.1016/S1350-4533(02)00031-0 PMID:12135650
- Hammami, I., Salhi, L., & Labidi, S. (2020). Voice Pathologies Classification and Detection Using EMD-DWT Analysis Based on Higher Order Statistic Features. *Journal of IRBM*, 41(3), 161-171. <https://10.1016/j.irbm.2019.11.004>
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE International Conference on Computer Vision*. doi:10.1109/ICCV.2015.123
- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780. Advance online publication. doi:10.1162/neco.1997.9.8.1735 PMID:9377276
- Huszar, F. (2017). *Mixup: Data-Dependent Data Augmentation*. <https://www.inference.vc/mixup-data-dependent-data-augmentation/>
- Kim, H. G., Moreau, N., & Sikora, T. (2006). *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. Wiley.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25(1).
- Lerch, A. (2012). *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. John Wiley & Sons, Inc. doi:10.1002/9781118393550
- Lin, Z., & Chen. (1997). Time frequency analyses of electrogastrogram. *Time frequency and wavelets in biomedical signal processing*, 147-181. .<ALIGNMENT.qj></ALIGNMENT>10.1109/9780470546697
- Lostanlen & Anden. (2016). *Binaural scene classification with wavelet scattering*. Technical Report, DCASE2016 Challenge.
- Mesallam, T. A., Farahat, M., Malki, K. H., Alsulaiman, M., Ali, Z., & Al-Nasheri, A. (2017). *Development of the Arabic Voice Pathology Database and Its Evaluation by Using Speech Features and Machine Learning Algorithms*. *J Health Eng*. doi:10.1155/2017/8783751
- Michaelis, D., Gramss, T., & Strube, H. W. (1997). Glottal-to-Noise Excitation Ratio - A New Measure for Describing Pathological Voices. *Acustica*, 83, 700–706.
- Muhammad, G., Mesallam, T. A., Malki, K. H., Farahat, M., Mahmood, A., & Alsulaiman, M. (2012). Multidirectional regression (MDR)-based features for automatic voice disorder detection. *Journal of Voice*, 26(6), 817.e19–817.e27. doi:10.1016/j.jvoice.2012.05.002 PMID:23177748
- Péan, V., Ouayoun, M., Fugain, C., Meyer, B., & Chouard, C. H. (2000). A fractal approach to normal and pathological voices. *Acta Oto-Laryngologica*, 120(2), 222–224. doi:10.1080/000164800750000964 PMID:11603777
- Peeters, G. (2004). *A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project*. Technical Report. IRCAM.
- Rabiner, L. R., & Schafer, R. W. (2010). *Theory and Applications of Digital Speech Processing*. Pearson.
- Ramirez, J., Segura, J. C., Benítez, C., de la Torre, Á., & Rubio, A. (2004). Efficient voice activity detection algorithms using long-term speech information. *Speech Communication*, 4(3-4), 271–287. doi:10.1016/j.specom.2003.10.002
- Scheirer, E., & Slaney, M. (1997). Construction and evaluation of a robust multifeature speech/music discriminator. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings*, 993-996. doi:10.1109/ICASSP.1997.596192

Smith, J. O. (2011). *Quadratic interpolation of spectral peaks*. W3K. https://ccrma.stanford.edu/~jos/sasp/Quadratic_Interpolation_Spectral_Peaks.html

Steffen, N., Vieira, V. P., Yazaki, R. K., & Pontes, P. (2011). Modifications of vestibular fold shape from respiration to phonation in unilateral vocal fold paralysis. *Journal of Voice*, 25(1), 111–113. doi:10.1016/j.jvoice.2009.05.001 PMID:20236796

Weber, E. D. (2010). The Massachusetts Eye and Ear Infirmary Illustrated Manual of Ophthalmology. *J Neuro-Ophthalmology*, 30(1), 106. doi:10.1097/01.wno.0000369166.94555.db

Vikas Mittal received his M.Tech in Electronics and Communication Engineering from Kurukshetra University Kurukshetra. Currently, he is pursuing Ph.D. at National Institute of Technology (NIT) in the School of VLSI Design and Embedded Systems. His research interests include Biomedical Signal Processing, VLSI Design and Embedded System.

R. K. Sharma, received his M.Tech in Electronics and Communication Engineering and PhD degree in electronics and communication from Kurukshetra University, Kurukshetra (through National Institute of Technology Kurukshetra), India in 1993 and 2007, respectively. Currently he is Professor with the Department of Electronics and Communication Engineering, NIT Kurukshetra, India. His main research interests are in the field of embedded applications, low power, digital design and disease/ stress detections using voice profiling of human beings.