Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Unraveling the complexities of pathological voice through saliency analysis

Abdullah Abdul Sattar Shaikh [a], M.S. Bhargavi [a,*], Ganesh R. Naik [b]

[a] Department of Computer Science and Engineering, Bangalore Institute of Technology, Bangalore, 560004, Karnataka, India
[b] Adelaide Institute for Sleep Health, Flinders University, Bedford Park 5042, Adelaide, SA, Australia

ARTICLE INFO

Keywords:
Pathological voice
Saliency analysis
Autoencoders
Multi-class classification
UNet++

ABSTRACT

The human voice is an essential communication tool, but various disorders and habits can disrupt it. Diagnosis of pathological and abnormal voices is very important. Conventional diagnosis of these voice pathologies can be invasive and costly. Voice pathology disorders can be effectively detected using Artificial Intelligence and computer-aided voice pathology classification tools. Previous studies focused primarily on binary classification, leaving limited attention to multi-class classification. This study proposes three different neural network architectures to investigate the feature characteristics of three voice pathologies-Hyperkinetic Dysphonia, Hypokinetic Dysphonia, Reflux Laryngitis, and healthy voices using multi-class classification and the Voice ICar fEDerico II (VOICED) dataset. The study proposes UNet++ autoencoder-based denoiser techniques for accurate feature extraction to overcome noisy data. The architectures include a Multi-Layer Perceptron (MLP) trained on structured feature sets, a Short-Time Fourier Transform (STFT) model, and a Mel-Frequency Cepstral Coefficients (MFCC) model. The MLP model on 143 features achieved 97.1% accuracy, while the STFT model showed similar performance with increased sensitivity of 99.8%. The MFCC model maintained 97.1% accuracy but with a smaller model size and improved accuracy on the Reflux Laryngitis class. The study identifies crucial features through saliency analysis and reveals that detecting voice abnormalities requires the identification of regions of inaudible high-pitch sounds. Additionally, the study highlights the challenges posed by limited and disjointed pathological voice databases and proposes solutions for enhancing the performance of voice abnormality classification. Overall, the study's findings have potential applications in clinical applications and specialized audio-capturing tools.

1. Introduction

The most ideal form of communication has always been a voice due to its effectiveness in exchanging information between two entities through words, feelings, and situational context. However, voice can be susceptible to speech imbalance or auditory disorders, making proper voice function a major concern in the medical field [1–3]. Additionally, voice is vital in identifying characteristics such as gender, illness, and mental or physical trauma. Voice analysis, especially pathological voice, is considered a significant field in healthcare because of the increasing risks of pathological voice problems [4,5].

Various factors can affect an individual's voice, such as pathological conditions or lifestyle habits like smoking, bad vocal hygiene, or improper use of vocals [6]. These issues can disrupt an individual's ability to exchange information effectively. Common voice disorders can arise from vocal cords that do not close entirely or vibrate asymmetrically in an unideal manner. Clinical practitioners typically use diagnostic methods such as Stroboscopy, Laryngoscopy, and Endoscopy, which require specialized tools and an experienced professional. However, these diagnoses are invasive, time-consuming, and costly for the patient [3]. Hence, there is a need for effective yet efficient diagnostic methods that can automatically diagnose patients, saving time and money. Computer-aided architectures are ideal solutions for building non-invasive, accurate, and fast detection diagnosis systems that work in real-time and are cost-effective [3,7–9].

Artificial Intelligence (AI) has developed robust classification and detection models for pathological voices. Machine Learning (ML) algorithms such as Support Vector Machines (SVM), decision trees, clustering, and others have shown effective solutions in detecting and classifying pathological voices using handcrafted features [3]. Despite great success in detecting abnormal voice signals using various ML algorithms and signal processing techniques, a further classification of pathological voice samples into several different etiologies has yet to be attempted. Moreover, some studies have focused primarily on binary classification, leaving limited attention to multi-class classification [10, 11].

Deep Learning (DL) models have recently been used for pathological voice classifications. Utilizing deep learning (DL) architecture, which

* Corresponding author.
E-mail addresses: abdullahshaikh136@gmail.com (A.A.S. Shaikh), ms.bhargavi@gmail.com (M.S. Bhargavi), ganesh.naik@flinders.edu.au (G.R. Naik).

https://doi.org/10.1016/j.compbiomed.2023.107566
Received 2 June 2023; Received in revised form 14 September 2023; Accepted 10 October 2023
Available online 14 October 2023

trains on latent features, has simplified training complex and robust models [11]. The benefits of deep learning are its ability to handle nonlinear data representations and solve sparsity problems [12]. A better insight into what features a deep learning architecture learns can help us further optimize ML and DL models or create specialized tools for better pathological voice diagnosis [13,14]. Identifying latent features in deep learning has already proven helpful, with practical applications in CNNs [15]. Techniques like Transfer Learning [16], Latent Space Visualization, and Latent Space Interpolation [17] have been employed to enhance model performance.

This research explores the latent features learned by a deep learning architecture through visual saliency. Visual saliency highlights the parts of the input that provide the highest information gain for the model. The higher the magnitude of the saliency at a particular point in the input, the greater the degree of information gain and activation of the model for a specific input class. In this study, the focus is on four classes of voice, including Hyperkinetic Dysphonia [18], Hypokinetic Dysphonia [19], Reflux Laryngitis [20], and a healthy voice class from PhysioNet's Voice ICar fEDerico II dataset [21] for multi-class classification (the motivation for a multi-class classification is discussed in Section 2). The dataset is pre-processed by splitting each sample into equal parts to increase the dataset size. Then the UNet++ autoencoder [22] is used for denoising the dataset to extract accurate features. Various feature extraction and selection methods were employed to build accurate ML models for classification.

Three standalone neural network architectures are proposed with different data format inputs. The first architecture is a Multi-Layer Perceptron (MLP) trained on a handcrafted structured feature set from continuous voice samples. The second architecture uses Short-Time Fourier Transform (STFT) as input, while the third architecture employs Mel-Frequency Cepstral Coefficients (MFCC). Finally, common features and patterns learned by the deep learning architecture are investigated using visual saliency to demystify each pathological voice class's biomarkers. The research has implications for improving classification systems and specialized audio-capturing tools, paving the way for further enhancements in voice abnormality classification through focused feature analysis and advanced data collection techniques.

Our research aims to expand upon the multi-class classification of pathological voices by providing insights into the critical biomarker features of pathological voice classes through saliency maps. The study also presents a more straightforward approach for achieving high multi-class classifiers' accuracy by utilizing heavier open-source autoencoders and lighter MLP and 1D CNN networks. In addition, we aim to promote the reproducibility of our work by using readily available open-source resources, such as the UNet++ denoiser architecture [22] and the OpenSmile tool [23] for reproducible feature extraction.

We also aim to facilitate future studies to delve deeper into the specific biomarkers identified in our research for advanced data collection and feature extraction processes. This will enable researchers to enhance the unique biomarkers of each pathological signal, thus providing a more accurate and reliable classification of pathological voices. Overall, our contribution provides an important step towards improving the multi-class classification of pathological voices and advancing the understanding of the complex relationships and features among different pathological voice classes.

The significant contributions of this research are (i) Exploring latent features learned by a deep learning architecture through visual saliency to identify critical biomarker features of pathological voice classes. (ii) Proposing three standalone neural network architectures with different data format inputs (MLP, STFT, MFCC) for multi-class classification of pathological voices. (iii) Employing various feature extraction and selection methods to build accurate machine learning models for classification. (iv) Investigating common features and patterns learned by the deep learning architecture using visual saliency to demystify biomarkers of each pathological voice class. (v) Promoting reproducibility by utilizing open-source resources like the UNet++ denoiser architecture and OpenSmile tool for feature extraction. (vi) Facilitating future studies to delve deeper into the specific biomarkers identified for advanced data collection and feature extraction processes, enhancing the accuracy and reliability of pathological voice classification.

The research paper is organized as follows. Section 1 provides an introduction to the research emphasizing our contribution to the field. The next section presents a comprehensive survey of relevant literature. The Section 3 provides a concise overview of the dataset, research methodology, and fundamental aspects involved in data pre-processing and feature extraction. Model architecture and data classification are discussed in Section 4. In the Section 5, we assess the results and analyze saliency maps. The Section 6 draws noteworthy observations and discusses further correlations and insights in the study. In the Section 7, we present conclusions and propose potential avenues for future research utilizing the information presented in this study.

2. Related works

While reviewing literature in this domain, binary classification techniques have emerged as a critical approach to differentiate between healthy and pathological voice signals [24]. Traditional nonparametric methods and modern end-to-end techniques, such as neural networks [25] and wavelet decompositions [26] have been widely investigated. Audio signal pre-processing involves the use of either long-duration or short-duration modes, where the former analyzes voice signals using features such as disruption of amplitude and frequency [27], excitation ratio of glottal to noise [28], harmonics to noise ratio (HNR), voice turbulence index, normalized energy levels of noise, amplitude and frequency vibration, among others [27,29–31]. However, long-duration features may be less effective than short-duration features due to the dataset's frequency dependencies. Among the short-duration features, MFCC has shown promising results for voice datasets. Nevertheless, MFCC's performance is limited to the dataset on which it is trained, making it less generalizable [31,32]. Furthermore, it has been found that using large datasets improves generalization and consequently enhances the models' performance [31,33].

Procedural generation of audio [34,35] using techniques such as Generative Adversarial Networks (GANs) has been employed to achieve high accuracy in binary classification by [36]. They also explored comparative analysis on the VOICED Database and SVD database using techniques like conditional generative adversarial network (CGAN), improved fuzzy c-means clustering (IFCM), synthetic minority oversampling technique (SMOTE), which gave a sensitivity rating of 81.6, 79.2 and 86.5 respectively. Another promising approach is using wavelet families, such as the combinational approach of Daubechies Discrete Wavelet Transform (DWT) and Support Vector Machines [37], which achieved a 91.67% accuracy in binary classification.

Real-time Voice Activity Detection (VAD) has been used to classify audio as pathological or healthy by [38] using the Arabic Voice Pathology Database [39]. One combinational work by [40] utilized MFCC and VAD for segmenting voice regions in the Massachusetts Eye & Ear Infirmary dataset [41] and achieved an accuracy of 96%.

Various machine learning algorithms have also shown robust accuracies such as boosted trees [42] with 50%, threshold-based algorithms [40] with 91%, random forests with 87.4% [43], and K-nearest neighbors [43] with 93.3%. SVM with Gaussian Mixture model has been implemented in [44] to achieve a 94% accuracy. In addition, [45] utilized a CNN-LSTM (convolutional short-term memory) to achieve an accuracy of 71.36%, while [46] combined this network with SVMs to increase its accuracy to 78.37%. Finally, [11] achieved an accuracy of 97.8% on their binary classification using MFCC features combined with stacked and sparse autoencoders.

Recent research in the domain of voice classification has also produced significant results. For instance, Muraleedharan et al. [10] used phase space analysis to derive six measures and achieved an impressive

97% accuracy in distinguishing healthy and pathological voices on the VOICED dataset. Wang et al. [47] utilized multi-domain features and a hierarchical extreme learning machine (H-ELM) to classify healthy and pathological voices with accuracies ranging from 98.99% to 99.61%. In addition, Altayeb et al. [48] conducted binary classification experiments on the VOICED dataset using a feature set consisting of MFCC, ZCR, and DWT. Their work achieved 100% binary classification accuracy on the kurtosis feature group, with other classification results ranging from 60% to 97%. All these works are displayed in Table 1 for comparison.

Other recent research that has significantly contributed to voice pathologies includes several notable works. Kumar et al. [49] employed Videokymography to examine laryngoscopic images, effectively classifying various voice pathologies. Kim et al. [50] focused on patients suffering from dysphagia, which adversely affected the quality of their voices. Kim et al. proposed a non-invasive diagnosis methodology to address this issue by leveraging MFCC and STFT pre-processing techniques on the patients' audio. In another study, Huckvale et al. [51] evaluated the significance of contextual information in speech signals, emphasizing the importance of continuous speech for achieving high performance in classification models.

Furthermore, Han et al. [52] developed a novel approach utilizing self-attention Bi-directional Long Short-Term Memory (SA BiLSTSMs) to analyze voice characteristics such as grade, roughness, and breathiness. Dianat et al. [53] took a unique direction by utilizing pulmonary voices for diagnosing lung diseases through connective tissue. They achieved high accuracy by employing CNNs and tomography in their diagnostic framework.

After reviewing the literature in this domain, it is clear that considerable work has been done on the binary classification of healthy and pathological voices. Although binary classification is helpful for the initial diagnosis of pathological voice, it fails to capture the complex relationships and features among different classes that can be achieved through multi-class classification. The only works found that deals with multi-class classification of pathological voices is the research by [36,54]. The research by [54] attains highest performance in this task by using a light autoencoder to denoise the VOICED dataset's audio samples. It proposes a 3D CNN DL classifier to classify the input signal after pre-processing into three types of spectrograms. They propose a unique "group decision analogy" system that achieves an overall accuracy of 97.7% with 100% accuracy on hypokinetic dysphonia and reflux laryngitis classes.

3. Methods and data processing

The methodology of this study follows a 3-step process: pre-processing, training, and analysis. The first step involves addressing class imbalance, noise, and feature extraction. To effectively address class imbalance, data augmentation techniques are utilized to increase the dataset size and minimize the adverse effects of class imbalance. In order to mitigate the impact of noise present in the dataset, the audio signal is deliberately overlapped with a designated real-time noise signal. Subsequently, this combined signal is passed through an autoencoder, which effectively denoises the audio signal, resulting in a cleaner and more reliable representation of the data. This process helps to extract accurate features from the data. Feature extraction is done using the openSMILE library [23,55] in Python, which creates 6373 handcrafted features. Feature selection is then performed by rigorously testing ML algorithms to identify the most crucial features, resulting in a final set of 143 features. It is observed that 80% (115 features) of the features in the final set are derivatives of MFCC (52 features) and STFT (63 features).

Training is carried out on three different DL architectures to further improve the accuracy of the models. The first model uses an MLP model trained on the top 143 features, while the second model trains on the STFT transformation of the denoised audio signals using 1-D Convolutional Neural Network (CNN) architecture. The third model also utilizes 1-D CNNs but accepts MFCC coefficients from the denoised audio signals as its feature set [56]. Once the training phase is completed, the models are evaluated for performance on a test set. For the final step, Saliency maps are used for visual analysis to gain insights into the latent features and patterns learned by all the models.

3.1. Dataset description

Limited pathological voice databases, such as the Massachusetts Eye and Ear Infirmary Database (MEEI) [41], Arabic Voice Pathology Database (AVPD) [39], Saarbrucken Voice Database (SVD) [57], and The Advanced Voice Function Assessment Databases (AVFAD) [58], are currently accessible for research and analysis. Among these, the MEEI database is the most widely used due to its comprehensive collection of organic, traumatic, and psychogenic voice disorders. However, it remains inaccessible to the public, and its voice samples come from diverse frequencies and environments, presenting challenges for research and studies. In contrast, open-source databases like AVPD, SVD, and AVFAD offer multiple recordings of voice samples comprising complete sentences or sustained vowel sounds such as 'a', 'e', 'i', 'o', and 'u'. Past research indicates that these different voice databases do not merge seamlessly or perform well, primarily due to their insufficient size for generalization [31]. Moreover, certain pathology classes are mutually exclusive in various datasets, and an overabundance of pathology classes can drop the model performance and hinder feasible feature analysis.

The dataset used in this study is the Voice ICar fEDerico II (VOICED) dataset [21,59], which was published in 2018. It includes individuals ranging in age from 18 to 70 years old. The voice recordings were obtained using a system called "Vox4Health", which utilized a Samsung Galaxy S4 smartphone microphone placed at a 45-degree angle and 20 cm away from the speaker. The recordings were conducted in a quiet medical room with humidity levels between 30% and 40% at the "High-Performance Computing and Networking (ICAR-CNR) in collaboration with the University Hospital of Naples". This dataset contains 206 five-second audio recordings, sampled at 8000 Hz, where patients vocalized a continuous vowel sound "a". Fig. 1 provides the audio with file sample visualization. The dataset includes recordings of patients diagnosed with three pathology classes: Hyperkinetic Dysphonia, Hypokinetic Dysphonia, Reflux Laryngitis, and a Healthy voice class. The distribution of the samples per class in the dataset is shown in the first two columns of Table 2 (the third column is explained further in Section 3.3 in Data Pre-Processing Techniques under Audio data augmentation subsection). In our scenario, we will also be utilizing the SVD database for comparative analysis as it is the only database with similar pathologies and audio properties to the VOICED Database.

3.2. Visual saliency

Saliency maps are a visualization technique utilized in DL to select features and identify the most relevant segment of the feature set for a given task. These maps are generated by taking the gradient of the output of the selected model with respect to the input signal. The gradients react according to the output of the model with the input signal, so their absolute values can represent the sensitivity of the input signal with the output extracted by the DL model.

Visualizing the gradients overlapped with the input signal provides information about the most relevant part of the signals activated for the current task. This makes saliency maps ideal for DL models with many layers and convolutions. It is not easy to pinpoint which features the neural network is learning and whether they are substantial.

Understanding the learning process of the neural network model through saliency maps can have implications for identifying potential areas of improvement in increasing model performance or modifying the feature extraction process.

Table 1
Selected previous studies in pathological classification.

Problem	Author	Dataset	Methodology	Result (Acc) %	Limitations
Binary classification of healthy vs. pathological voice	Fonseca et al. (2007) [37]	Private Dataset	DWT + SVM	91.67	Insufficient evaluation metric reports for reliability testing
	Muhammad et al. (2012) [38]	AVPD	Multi directional Regression (MDR) model	97.48	Experiments were not generalized; No cross validation experiments were performed
	Godino Llorente et al. (2009) [40]	MEEI	MFCC + VAD	96.00	Healthy vs. Pathological class imbalance was not addressed
	Amara et al. (2016) [44]	SVD	SVM + Gaussian Mixture mode	94.00	Small sample size; No cross validation experiments were performed
	Harar et al. (2017) [45]	SVD	CNN + LSTM	71.36	Insufficient metrics reports and experiments
	Kadiri et al. (2020) [46]	HUPA, SVD	Glottal flows estimate features on SVM	78.37	Limited exploration of feature extraction and engineering methods for enhanced classification performance
	Chen et al. (2022) [11]	VOICED	MFCC features on DNN	97.80	Complex implementation
	Muraleedharan et al. (2023) [10]	VOICED	Phase space features on SVM	97.00	Work lacks crucial sensitivity, specificity data. Feature interpretability missing for proposed phase space characteristics in medical voice analysis.
Multiple binary classification of healthy vs. different pathological diagnosis classes	Altayeb et al. (2022) [48]	VOICED	MFCC + ZCR + DWT features on SVM	60.00 to 90.00	Abstract overlooks dataset-specific influence on chosen features (MFCCs, ZCR, DWT) and their applicability to different datasets.
	Chen et al. (2021) [43]	VOICED	HTT + KNN	93.30	No comparative analysis with similar methods related to KNNs and Hilbert–Huang Transform
Multi-class classification on different pathological diagnosis classes	Chui et al. (2020) [36]	SVD, VOICED	CGAN-IFCM	90.1% (sensitivity)	Lacks imbalanced data strategy, skips CGAN-IFCM limitations, and omits non-parametric test drawbacks.
	Wahengbam et al. (2021) [54]	VOICED	3D CNN + custom group analogy	97.70	Complex implementation, Not tested against another dataset, Computationally heavy requirements

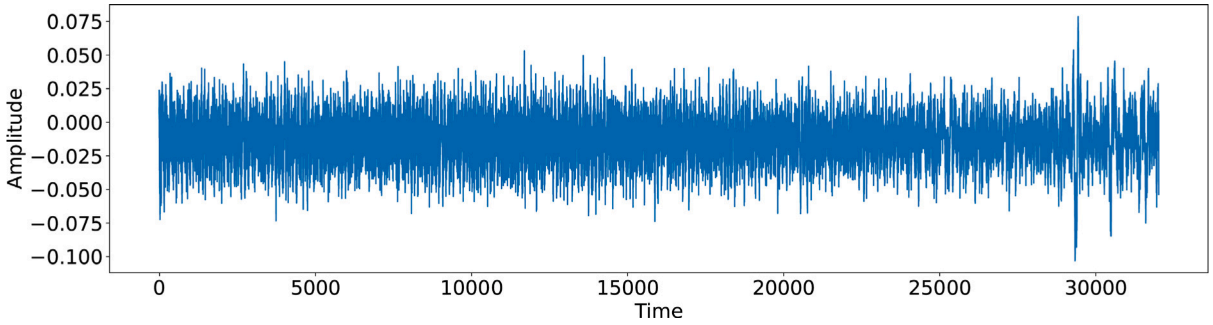


Fig. 1. Sample audio.

Table 2
Dataset distribution.

Class	Total No. of samples	Pre-processed No. of samples
Hyperkinetic Dysphonia	72	720
Hypokinetic Dysphonia	41	410
Reflux Laryngitis	38	380
Healthy	55	550
Total	206	2060

3.3. Data pre-processing and feature extraction

This subsection explains the various levels of pre-processing performed on the dataset.

3.3.1. Audio data augmentation

As we review Table 2, we observe that the dataset's class distribution is imbalanced, often resulting in excessive false positives during model evaluation. This imbalance, combined with the long duration of each audio sample (approximately 5 s or 38 720 sampling points), can lead to overfitting of the ML model. To combat this issue, we first drop the first 2500 sample points (from the existing 38 720 points) in the audio signal to minimize the bias brought about by the silence due to latency. We then divided each sample into ten equal parts, and when necessary, padding was applied to audio segments that fell short of the segment size of 3872. Fig. 2 provides a visual representation of sample audio. This procedure increased the total number of training samples from 206 to 2060, as shown in Table 2 's third column. As a result, the effects of overfitting and class imbalance in the dataset were minimized.

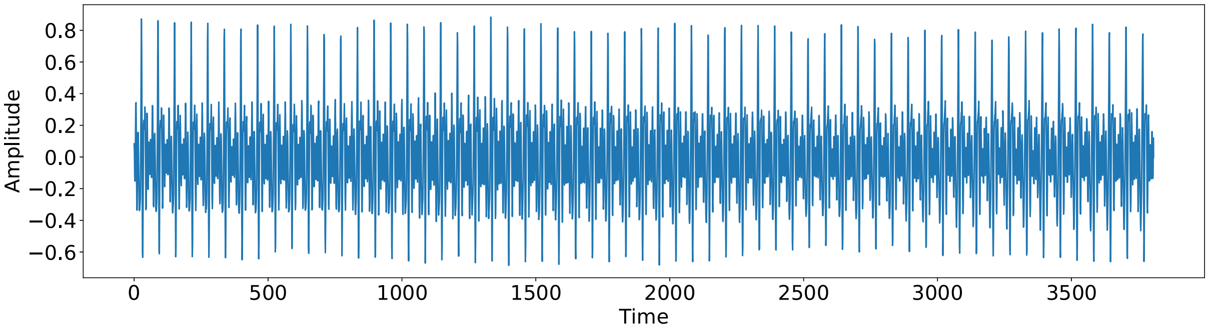


Fig. 2. Segmented sample audio.

3.3.2. Denoising

Denoising an audio sample offers numerous advantages before feeding it into an ML model [60,61]. One significant benefit is its potential as a self-supervised learning system, achieved by introducing noise to the audio file and denoising it. Training the model to restore the original clean audio from the noisy version allows it to differentiate between the noise and signal components of the audio. Through extensive training, the model becomes proficient in eliminating noise and improving signal quality from audio samples during inference, resulting in improved robustness and reliability in real-time scenarios.

During our testing, we discovered that the pathological voice has a Signal-to-Noise (SNR) ratio of −64 dB making it extremely noisy and complex by nature to handle. This severely inhibited classification results. Only the most significant signals could simplify the feature extraction process and provide an accurate feature for training. To achieve this, we overlapped an original signal (s) with a noise signal (n) to obtain the input signal i (Eq. (1)).

$$i = s + n \tag{1}$$

Next, we utilized an autoencoder-based denoising architecture trained on i as the training data and the original signal (s) as the validation data until we achieved a low and stable enough Huber loss [62] value. Once the model is trained, we feed the original signal (s) back into the autoencoder and obtain a denoised feature output.

For our specific use case, we utilized the washing machine sound obtained from the Urban Sound 8k dataset [63] (Fig. 5 given below) as the noise data (n) mixed with our pathological dataset audio (s) to obtain the input training signal (i). As for the autoencoder, after rigorous testing, we found that the 5-layer UNet++ autoencoder architecture provided the best minimization in loss and maintaining features in the signals. This architecture is a variation of the UNet autoencoder. It consists of multiple layers (L in Fig. 3), encoding (downsampling represented in Backbone in Fig. 3) and corresponding decoding (upsampling) paths through various feature maps (X at level i and j in Figure unetpp). UNet++ also utilizes a series of nested and dense skip pathways (green in Fig. 3) for finer feature extraction. Compared to the UNet[ronneberger2015u] autoencoder, the UNet++ autoencoder consists of convolution layers on its skips connections (blue in Fig. 3), which bridges the semantic gap between the encoder and decoder feature maps. Moreover, due to its support for various resolutions and skip connections, the UNet++ architecture captures high-level and low-level features with a high level of detail. This makes it an ideal autoencoder for noisy datasets that require handcrafting features and further feature extraction processes.

To train this autoencoder, we pushed the overlapped noisy signal (Fig. 6) i into the UNet++ architecture with the original audio signal s (Fig. 4) as the validation set and trained using the Adam optimizer for 150 epochs on a batch size of 2 until we achieved sufficient convergence on Huber loss value. After this, the model was frozen, and the original signal (s) was pushed through the network to obtain the denoised version of the audio signal (Fig. 7).

Table 3
Comparison between original speech signal and denoised speech signal.

Signal	SNR (dB)
Noisy speech	−64.28
Denoised speech	−25.41

Table 4
ML algorithm experiments on handcrafted features.

ML algorithm	K-fold accuracy score
ExtraTreesClassifier	0.93
LGBMClassifier	0.91
RandomForestClassifier	0.90
KNeighborsClassifier	0.89
XGBClassifier	0.89
LabelPropagation	0.84
LabelSpreading	0.84
BaggingClassifier	0.82
SGDClassifier	0.45
LogisticRegression	0.56
DecisionTreeClassifier	0.69

Although the changes appear insignificant in the visualizations, the signal-to-noise ratio (SNR) was improved from −64 dB to −25 dB, resulting in a total noise reduction by 61% as shown in Table 3. This demonstrates that a significant portion of the noise has been removed from the signal. However, since the SNR level is still negative, it suggests that the audio signal may still contain residual noise, and the signal itself may still be weak. In light of this, handcrafting features may be a viable option for processing these signals to obtain the most information from the dataset.

3.3.3. Structured data feature extraction

A set of features was created manually rather than learned through DL to train ML classifiers. The features were extracted from 3872 sample points in each segmented audio signal using the OpenSMILE feature extractor tool in Python [23]. To ensure better coverage and reproducibility, the “ComParE_2016” [55] standard acoustic feature set, consisting of 6373 various acoustic features, their derivatives, and levels, was chosen. The feature set includes multiple features such as MFCC in different quartiles of the signals, forward Fourier transform, and some of its derivatives like entropy, magnitude, and peaks.

Once the feature extraction was completed, feature selection was performed. We conducted rigorous ML experiments with K-fold cross-validation, as shown in Table 4, to determine the best features in the dataset. The Extremely Randomized Trees Classifier (Extra Trees Classifier) provided the best score, and we ran the model on different numbers of top features, ranging from 1 to 500. Finally, we found that the model provided the best results on the top 143 features, which we selected as our final feature set. This feature set will be used as an input for the deep MLP network (discussed in the training and evaluation section)

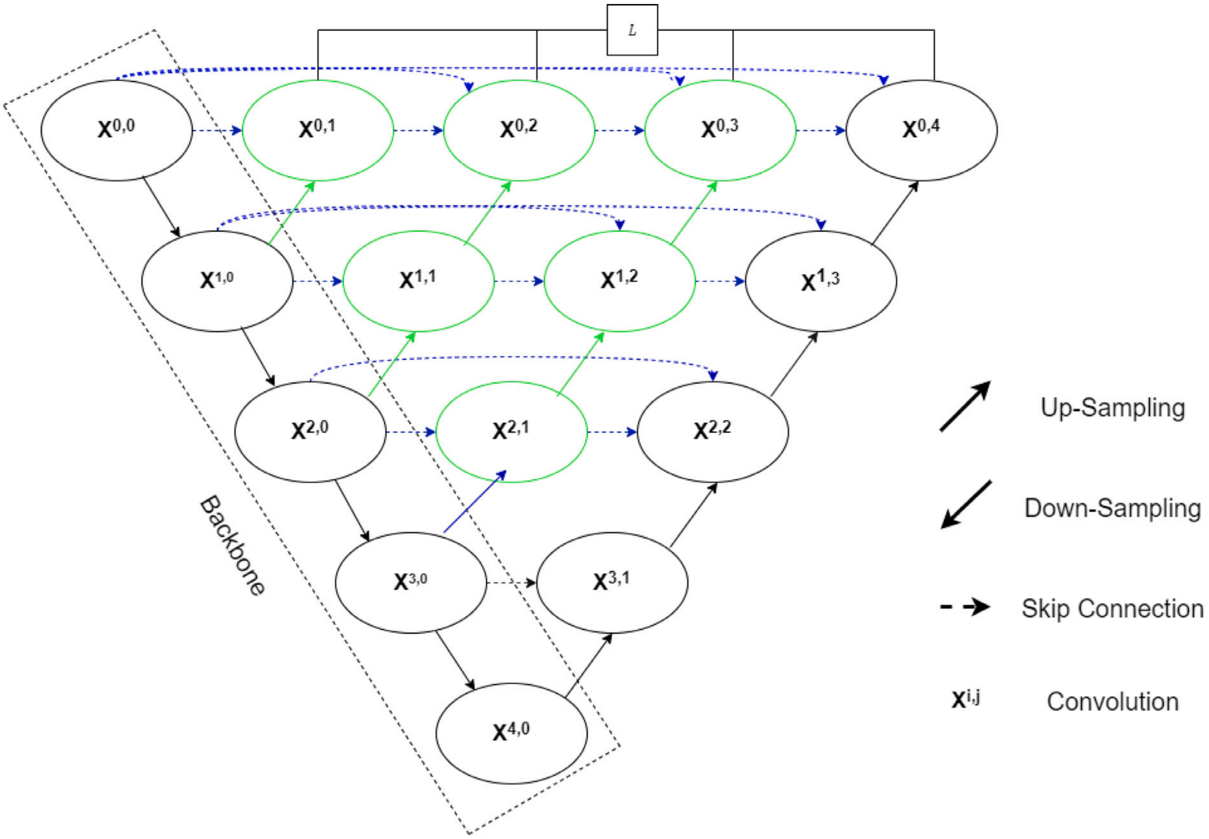


Fig. 3. UNet++ architecture.

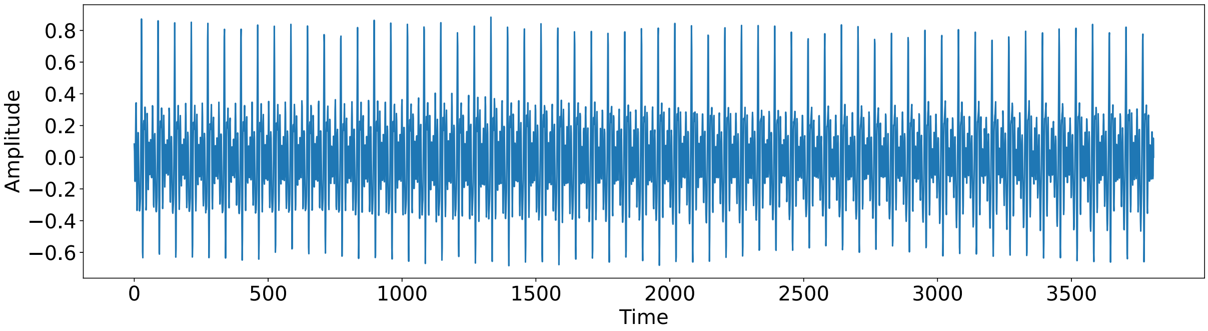


Fig. 4. Original audio segment.

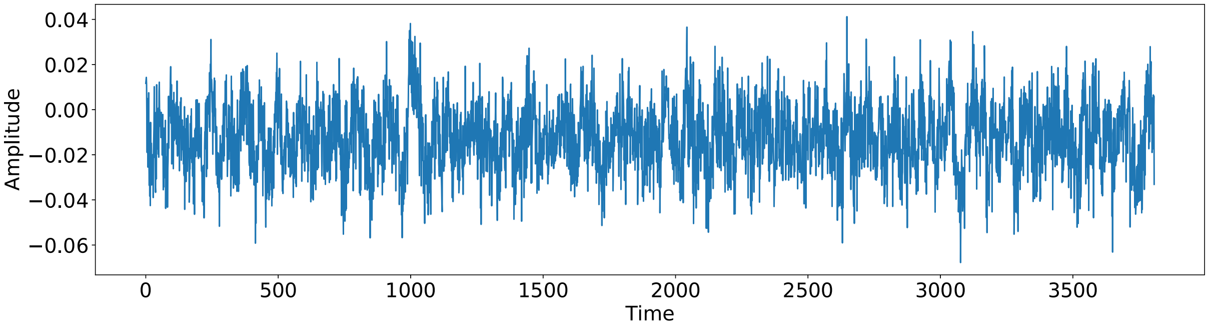


Fig. 5. Noise audio.

Further investigation into this feature set revealed that 63 features were derivatives of forward Fourier transform (FFT), and 52 were derivatives of MFCC. This indicates that 80% (115 out of 143) of the top features in the feature set heavily rely on Fourier transform and MFCC transformation of the pathology signals. Therefore, further investigation into these features is warranted.

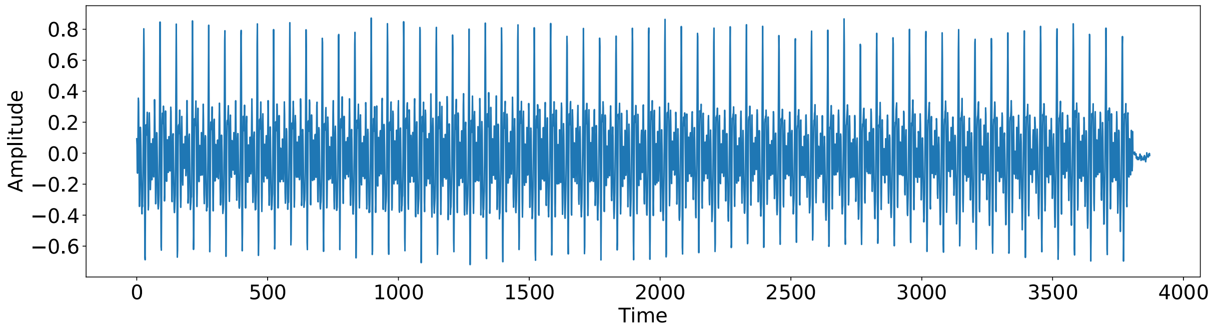


Fig. 6. Mixed noise audio.

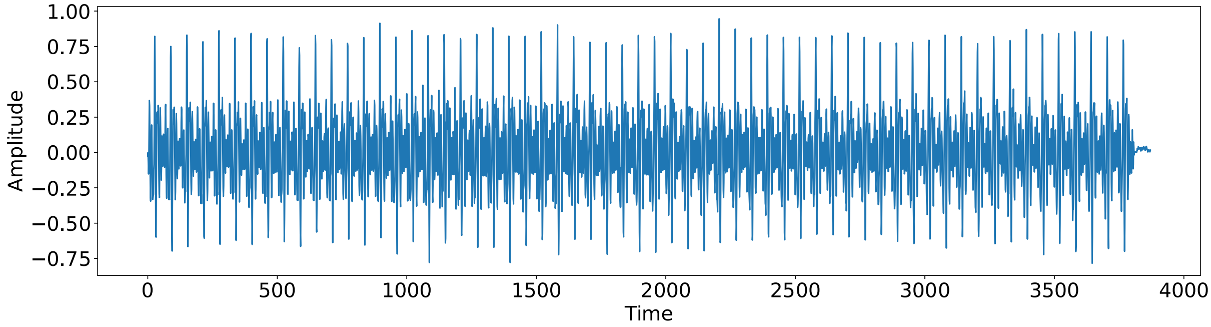


Fig. 7. Denoised audio.

3.3.4. Short-Time Fourier Transform (STFT)

Due to the significant reliance of our dataset on Forward Fourier Transform, we decided to explore Discrete Fourier Transformation (DFT) in greater depth. FFT is a form of DFT that calculates a signal's frequency throughout its entire duration by assuming that the signal is stationary, which means that its statistical properties, such as mean and variance, remain constant over time (e.g., white noise). This assumption provides us with a time–frequency representation of the signal. However, FFT has some significant limitations. It assumes the signal is stationary and provides only limited information on the signal's frequency content.

To address this issue, we utilized the STFT, which can better track the frequency content of signals over time by dividing them into overlapping segments. STFT is more effective for analyzing non-stationary signals because it provides a more localized and immediate representation of frequency changes. As a result, it provides more features that may be overlooked by FFT.

For our configuration, we implemented STFT using the "Hann" window configuration, and a hop size of 25% of the Fourier window of 2046 was used for training.

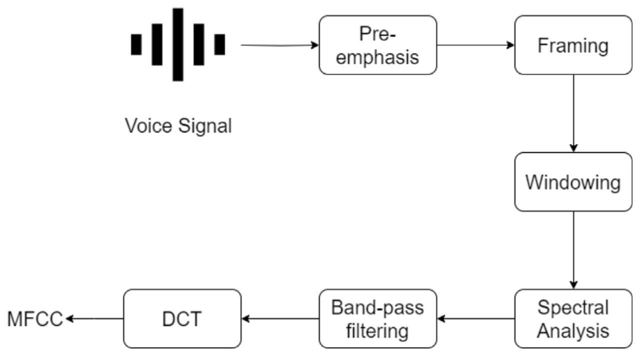


Fig. 8. MFCC workflow.

3.4. Mel-frequency cepstral coefficient

MFCC has always been highly regarded because of its high reliability in expressing essential frequency–time spectrums in a condensed manner [64].

The MFCC process involves seven steps, as illustrated in Fig. 8. The steps from the input signal to the output MFCC coefficients are:

1. Pre-Emphasis: The signal is first passed through a pre-emphasis filter, which amplifies higher frequencies and compensates for the attenuation of lower frequencies.
2. Framing: The signal is divided into N segments of equal size, with an overlap between any two adjacent segments.
3. Windowing: A "Hamming window" is applied to each segment to reduce discontinuity between adjacent segments. The Hamming window is given by Eq. (2)

$$W(n) = 0.54 - 0.46cos\frac{2\pi n}{N-1}, 0 \le n \le N-1 \tag{2}$$

4. Spectral Analysis: The Fast Fourier Transform (FFT) is applied to each segment to obtain its spectral form.
5. Band-Pass Filtering: The log amplitude of the spectrograms is filtered using a Mel-Scale filter bank. The conversion of frequency f (in Hz) to Mel scale is achieved by the formula Eq. (3)

$$Mel(f) = 1125 * ln(1 + \frac{f}{700}) \tag{3}$$

6. Discrete Cosine Transform (DCT): The DCT is applied to the resulting mel log amplitude from the previous step.
7. MFCC: The inverse FFT is applied to the mel coefficients to obtain the MFCC coefficients, as given by Eq. (4):

$$c_n = \sum_{k=0}^{n-1} log(S_k)cos[n(k - \frac{1}{2})\frac{\pi}{k}], n = 1, 2, \dots ., K \tag{4}$$

where S_k is the output power spectrum of filters and K is chosen to be 26 in our specific use case.

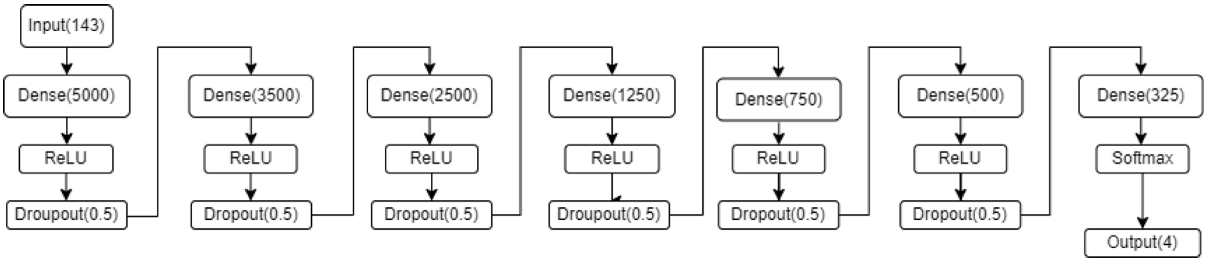


Fig. 9. Multi layer perceptron model to train handcrafted feature set.

4. Model architecture and data classification

4.1. Hardware and software

The experiments conducted in this research were performed on a Windows 10 computer equipped with an Intel Core i5 processor, 16 GB of RAM, and a dedicated graphics card (GTX 1060 Max-Q) with 6 GB of VRAM. Jupyter Notebook was the development environment, with Python 3 being the primary programming language. The dataset analysis was performed using the Matplotlib library. For feature extraction, training, and evaluation of the machine learning models, the Scikit-learn and Pandas libraries were used. In contrast, deep learning experiments were conducted on the GPU-accelerated Tensorflow 2.2.0 architecture.

4.2. Deep learning models and classification

In our research, we recognized that maintaining the same neural network architecture for all three input features would result in early flattening and effectively converting all the networks into an MLP model. This would limit the potential benefits and features derived from utilizing 1D and 2D convolutions designed explicitly for the STFT and MFCC input types. Therefore, we tried to maximize each neural network's performance individually, ensuring they can provide reliable saliency activation and extract valuable features from their respective input types. By doing so, we aimed to optimize the classification accuracy and capture the unique characteristics offered by each input feature.

After conducting tests on various MLP architectures, we discovered that the dataset is susceptible to overfitting when the MLP's depth is high, whereas increasing the model's width leads to a performance boost. As a result of thorough testing of multiple configurations, we concluded that the MLP model architecture illustrated in Fig. 9, having total parameters of around 35 million, produced the most accurate results for the given dataset. We employed the selected 143 features and utilized them as input to the model. We adopted specific data separation and training techniques to provide a robust evaluation of model's performance. The dataset was randomly shuffled before splitting into training and testing sets to mitigate any potential bias in the data distribution. We used a train-to-test split ratio of 8:2, meaning 80% of the data was used for training, and the remaining 20% for testing the model's generalization ability. The model underwent 5000 training epochs with a batch size of 8192 and utilized the Adam optimizer and categorical cross-entropy as its loss function. Additionally, we used a callback function to monitor the model's performance and record the best version of the model at each epoch.

Similarly, the DL architecture for the STFT format of the input signal also showed tendencies to overfit when the model depth was high but an increase in performance when we increased the width by increasing the kernel and filter size of the convolutional layers. Keeping this in mind, the model in Fig. 10 of size 3.2 million parameters provided us with the best accuracy after training on the 2060 STFT transformed data signal samples with the same train-to-test ratio of 8:2. The model was trained on a batch size of 128 and 500 epochs and utilized the same callback function, optimizer, and loss function as the MLP model.

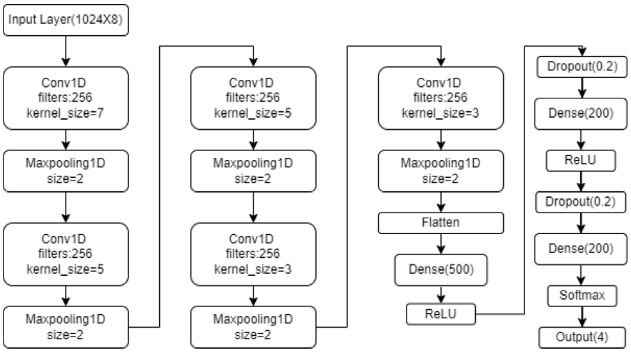


Fig. 10. 1D model to train STFT transformed audio data.

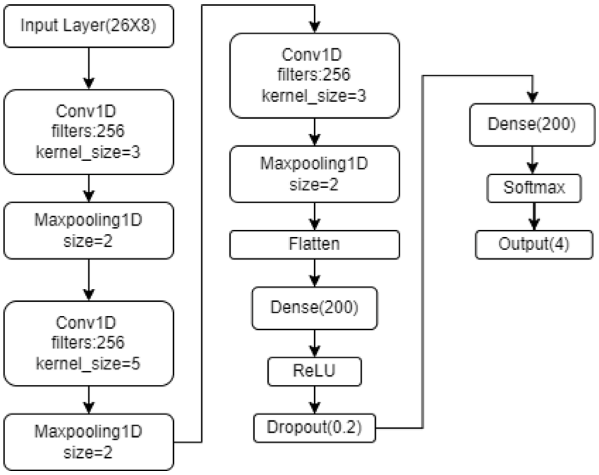


Fig. 11. 1D model to train MFCC transformed audio data.

Lastly, the MFCC model (Fig. 11), which has 500,000 parameters, is the smallest architecture among the three models compared to the parameter count. This is 6 times lower than the parameter count of the STFT model. Both models were trained using the same hyperparameter configuration, training split, and batch size.

5. Results

5.1. Model evaluation

To evaluate model performance, we use Accuracy, Sensitivity, Specificity, and F1 Score with their formulas shown in Eqs. (5), (6), (7), (8) respectively. TP, TN, FP, FN stands for true positive, true negative, false positive, and false negative, respectively. To check the model's

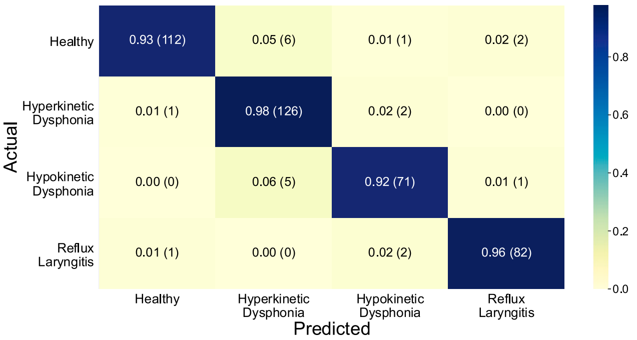


Fig. 12. Confusion matrix of the MLP model.

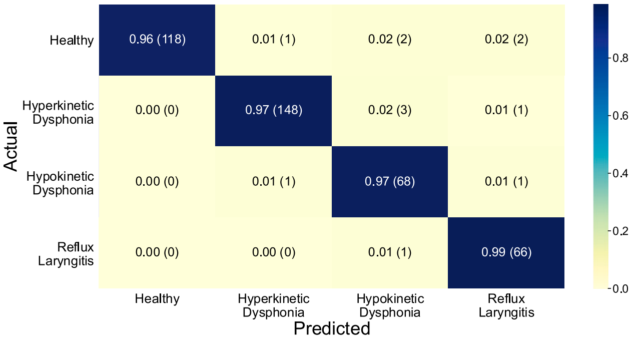


Fig. 13. Confusion matrix of the STFT model.

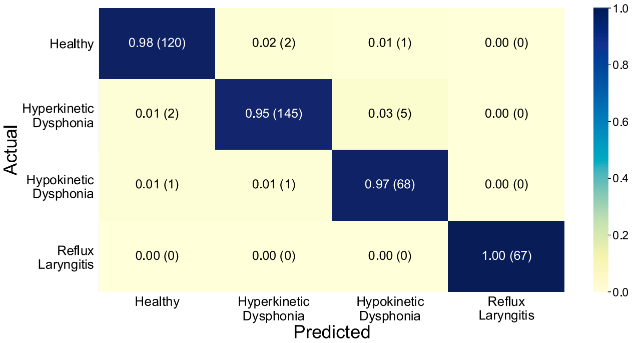


Fig. 14. Confusion matrix of the MFCC model.

reliability and robustness to produce similar results, we also employ Standard deviation (SD).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$Specificity = \frac{TN}{TP + TN + FP + FN} \tag{7}$$

$$F1score = \frac{2TP}{2TP + FP + FN} \tag{8}$$

After training the MLP model, it achieved an accuracy of 97.1% with a sensitivity of 98.3%, specificity of 99.9%, and an F1 score of 95%. To gain insight into the accuracy of each class, we utilized a confusion matrix (as shown in Fig. 12). The STFT model, which was also trained and evaluated, achieved an accuracy rate of 97.1%, equivalent to that of the MLP model. However, the STFT model exhibited a significant increase in sensitivity (99.8%), specificity (99.2%), and F1 score (97%). Additionally, based on Fig. 13, we achieved a 99% accuracy rate in the Reflux Laryngitis category.

Table 5
Model class metrics.

Models	Classes	Accuracy score	Sensitivity	Specificity	F1 score
MLP	Healthy	93%	93%	99%	0.95
	Hyperkinetic Dysphonia	98%	98%	96%	0.95
	Hypokinetic Dysphonia	92%	92%	98%	0.93
	Reflux Laryngitis	96%	96%	99%	0.96
STFT	Healthy	97%	96%	100%	0.98
	Hyperkinetic Dysphonia	97%	97%	99%	0.98
	Hypokinetic Dysphonia	96%	97%	98%	0.94
	Reflux Laryngitis	99%	99%	98%	0.96
MFCC	Healthy	98%	98%	98%	0.98
	Hyperkinetic Dysphonia	95%	95%	98%	0.97
	Hypokinetic Dysphonia	97%	97%	98%	0.94
	Reflux Laryngitis	100%	100%	100%	1.00

The MFCC model, which has significantly fewer parameters than the STFT and MLP models, achieved an accuracy of 97.1%, a sensitivity of 98.1%, a specificity of 99%, and an F1 score of 97.1%. Remarkably, the MFCC model attained 100% accuracy on the Reflux Laryngitis class (as shown in Fig. 14). Notably, the MFCC model is smaller and more efficient than the STFT model. The performance metrics of each model and its respective class are summarized in Table 5

Table 6 gives a summary of the evaluation metrics for each model. Among them, the MLP model stands out due to its large number of parameters, resulting in a high Standard Deviation (SD) for both accuracy and sensitivity. The high SD implies that this model is less efficient and robust than others. On the other hand, the STFT model shows exceptional performance across all metrics, although it exhibits intermediate SD values. This suggests some fluctuation in the model's reliability. The MFCC model performs similarly to the STFT model but has a significantly smaller size, only 6 times smaller. Additionally, it demonstrates substantially lower SD, indicating stable training and making it the most reliable model in our evaluation.

Fig. 15 illustrates the training trends of the various models. Upon observing Fig. 15(A), a relatively consistent training trend becomes evident. However, the model's reliability diminishes due to the substantial number of epochs (5000 epochs) required to attain maximum accuracy, resulting in an accumulation of standard deviation. This phenomenon is further observable in the amplified fluctuation of the model's loss values as depicted in Fig. 15(D). In contrast, the STFT model and the MFCC model achieve their peak accuracy within fewer than 500 epochs, showcasing minimal fluctuations in accuracy trends, as seen in Fig. 15(B) and (C). Notably, the STFT model demonstrates comparatively more variation in comparison to the MFCC model. Akin fluctuations are discernible in the loss values during training, as displayed in Fig. 15(E) and (F).

The Receiver Operating Characteristic curve (ROC), a graphical probability curve, and the Area Under Curve (AUC), which quantifies separability, provide insight into the models' capability to differentiate among the four classes. Figs. 16, 17, and 18 showcase these metrics. The ROC curves exhibit distinct inflection points, indicating a low False Positive (FP) rate. This signifies the model's high confidence in predicting the negative class, thereby reducing false positives and enhancing the model's ability to make accurate positive predictions. Such precision can prove critical, particularly in medical applications.

5.2. Saliency analysis

This section delves deeper into the saliency activation of each model, focusing first on the MLP model. Specifically, we explore the top five activated features within this model. As shown in the Fig. 19, each class has its specific set of top five activated features. Aside from FFT, it is worth noting that most of the feature set consists of 'sma' and its variations. Spectral mean amplitude (SMA) is a feature derived from STFT, and it measures the average amplitude of the signal across

Table 6
Performance metrics.

Model	Accuracy	Accuracy SD	Sensitivity	Sensitivity SD	Specificity	F1-score	Model parameters
MLP	97.10%	0.156	98.30%	0.047	99.10%	95.30%	3,50,19,665
STFT	97.10%	0.203	99.80%	0.096	99.20%	97.00%	30,26,560
MFCC	97.10%	0.019	98.30%	0.013	99.00%	97.10%	5,83,404

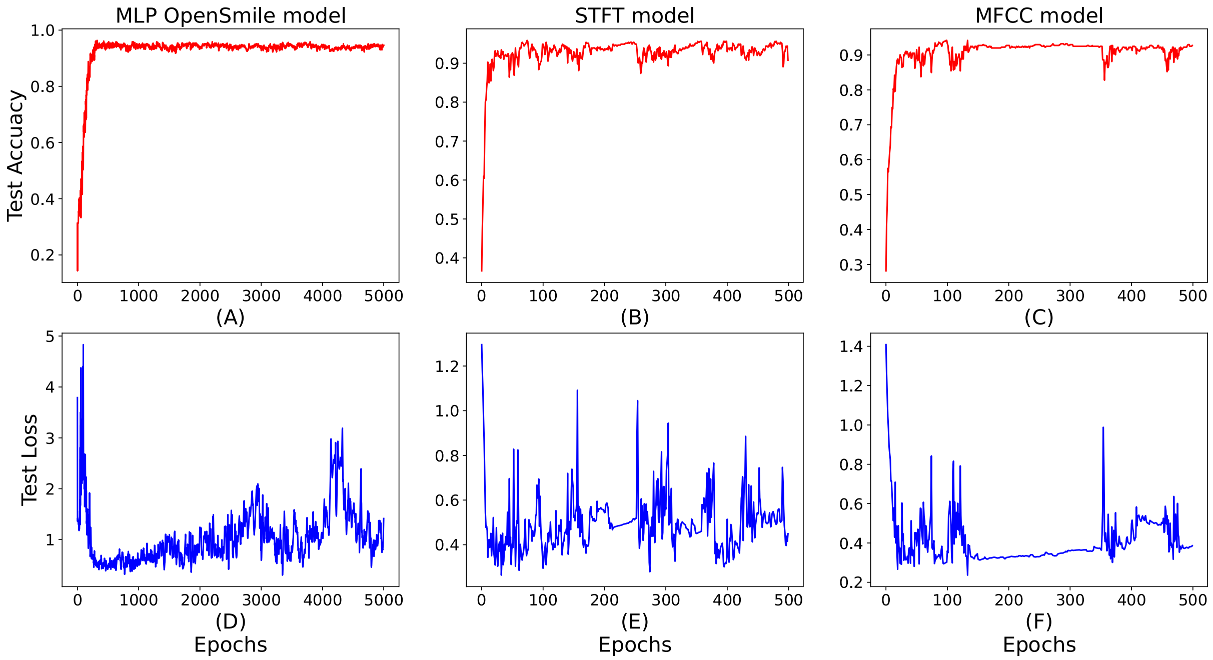


Fig. 15. Training trends of models.

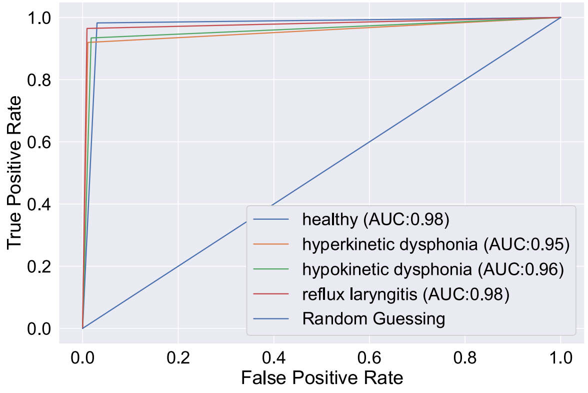


Fig. 16. ROC curve of MLP model.

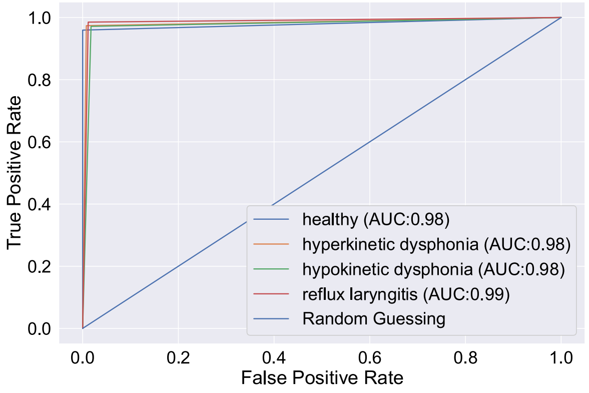


Fig. 17. ROC curve of STFT model.

all frequencies. This suggests that the model heavily relies on frequency and amplitude characteristics. To better understand the specific ranges of frequencies and amplitudes that the model activates, we will utilize the STFT and MFCC models.

Fig. 20 depicts four columns specific to each class and four rows for comparative analysis. The first row showcases the input spectrograms into the STFT model, and the second row displays its saliency map. While no conclusive pattern can be derived from the frequency and time activations, it is evident that the model activates in regions of low amplitude (as denoted by the color range in the spectrogram). To enhance visualization, an orthogonal view of the model is plotted in the third row, which confirms that the model frequently exhibits activations in regions of low amplitude. Furthermore, the saliency map is normalized, scaled on top of the input, and converted back to an audio signal using inverse STFT. However, the feature audio signals were inaudible due to their low amplitudes, as shown in red in the final row of Fig. 20. It should be noted that the feature amplitude highlighted in red in this row has been magnified tenfold compared to its actual features; otherwise, the amplitudes are too small to be visible in the visualization.

Similarly, the input spectrograms of the MFCC model and its respective saliency map, as illustrated in Fig. 21, indicate that the model also activates in low amplitude regions. Additionally, it is noteworthy that the model does not exhibit activation in lower frequency ranges.

To analyze frequency patterns across the dataset, we first map all activation values for each sample in the test set onto spectrograms for both the STFT and MFCC models. This process is illustrated in Figs. 22 and 23, respectively. The resulting feature maps show an activation

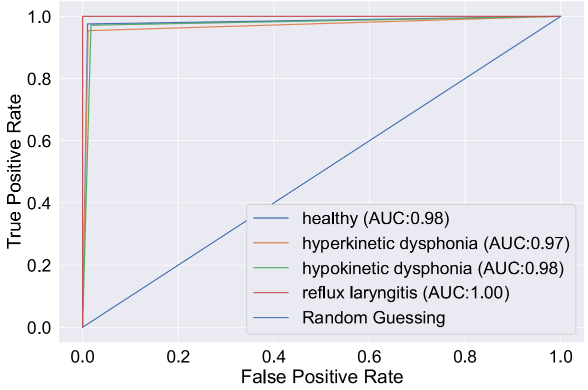


Fig. 18. ROC curve of MFCC model.

map for all test samples, with a plot point registered on the map when the activation is greater than 80% of the model’s activation. The feature map plots this information at the specified frequency and time location. When a specific spot has more activations from other test samples, the color intensity of that plot increases. However, it is important to note that lighter regions should not be ignored as they represent the top 80% activation of a particular test set. From these feature maps, we can conclude that deep learning models rely heavily on high ranges of frequencies, with minimal activation in the first lower half of the frequency range.

5.3. Comparative analysis

A comparison chart presented in Table 7 provides an overview of recent state-of-the-art approaches conducted by Chui et al. [36] and Wahengbam et al. [54] to address multi-class classification on the VOICED database. Chui et al. [36] focused on addressing the class imbalance in the database using techniques such as SMOTE, IFCM, CGANS, Random Forests, and Support Vector Machines. Although their evaluation metrics consisted solely of sensitivity and specificity, these metrics are crucial as sensitivity represents the number of individuals with voice pathology and specificity represents the number of healthy individuals. Among the experiments conducted by Chui et al. CGAN achieved the best results, with 90.1% sensitivity and 89.4% specificity. This approach outperformed the 2D scalogram approach by Wahengbam et al. [54], utilizing MobileNet [65] and GoogleLeNet [66], in terms of sensitivity but exhibited lower specificity.

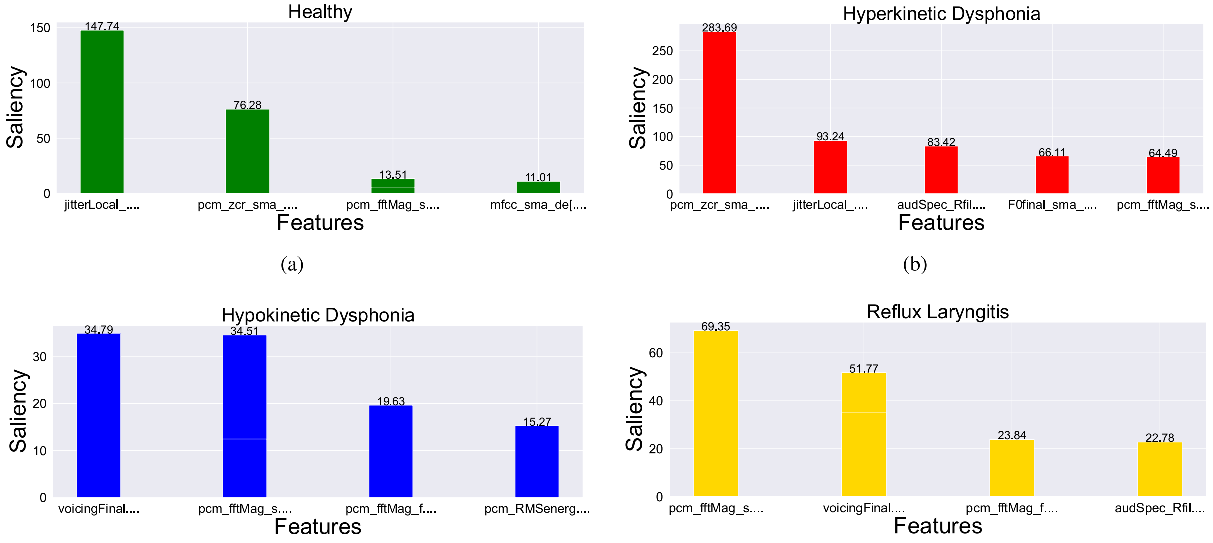
On the other hand, Wahengbam et al. proposed the highest-performing model, utilizing a heavy 3D CNN model alongside a custom group Decision Analogy using Amor, Bump, and Morse scalograms combined. This approach achieved the highest accuracy of 97.7%, sensitivity of 97.7%, and specificity of 99.23%.

In comparison, our proposed system achieves better results in sensitivity. It demonstrates comparable performance in terms of accuracy and specificity when compared to the 3D approach, utilizing less complex MLP, 1D, and 2D neural network architectures.

To ensure a reliable comparison of the saliency analysis conducted in our proposed models, we utilized the SVD database [57]. This database was chosen because it contains classes relatively similar to those in the VOICED database. While the SVD database does not explicitly provide hyperkinetic dysphonia or hypokinetic dysphonia, it does include the voice pathology “Laryngitis”, which is comparable to the ”Reflux Laryngitis” class in the VOICED dataset. Additionally, the SVD database offers the “Dysphonie” class, which exhibits similarities to the aforementioned voice pathologies. Although the SVD database consists of various other pathological voices, we selected only relevant classes to ensure high performance and reliability of the SVD classifier in generating saliency activation maps. By limiting the number of classes, we maintained the accuracy of the SVD classifier at 75.27%, with a sensitivity of 87.10% and a specificity of 94.89%. The saliency activations obtained from the analysis revealed similar activation patterns in regions of high frequencies and low amplitude. This demonstrates common underlying features in pathological voices, which can be leveraged to effectively classify voice pathologies. A visualization of the saliency activation maps for the MFCC input feature is provided in Fig. 24.

6. Discussion

The field of voice pathology has seen recent advancements driven by the non-invasive and efficient nature of the procedure. This research focusses on the multi-class classification of three pathological and healthy voices, aiming to identify the important features of each voice type. While binary classification (healthy vs. pathological) has



(a)

(b)

(c)

(d)

Fig. 19. Saliency activation on features of class (a) Healthy (b) Hyperkinetic Dysphonia (c) Hypokinetic Dysphonia (d) Reflux Laryngitis (The features are mentioned in the Appendix for better readability).

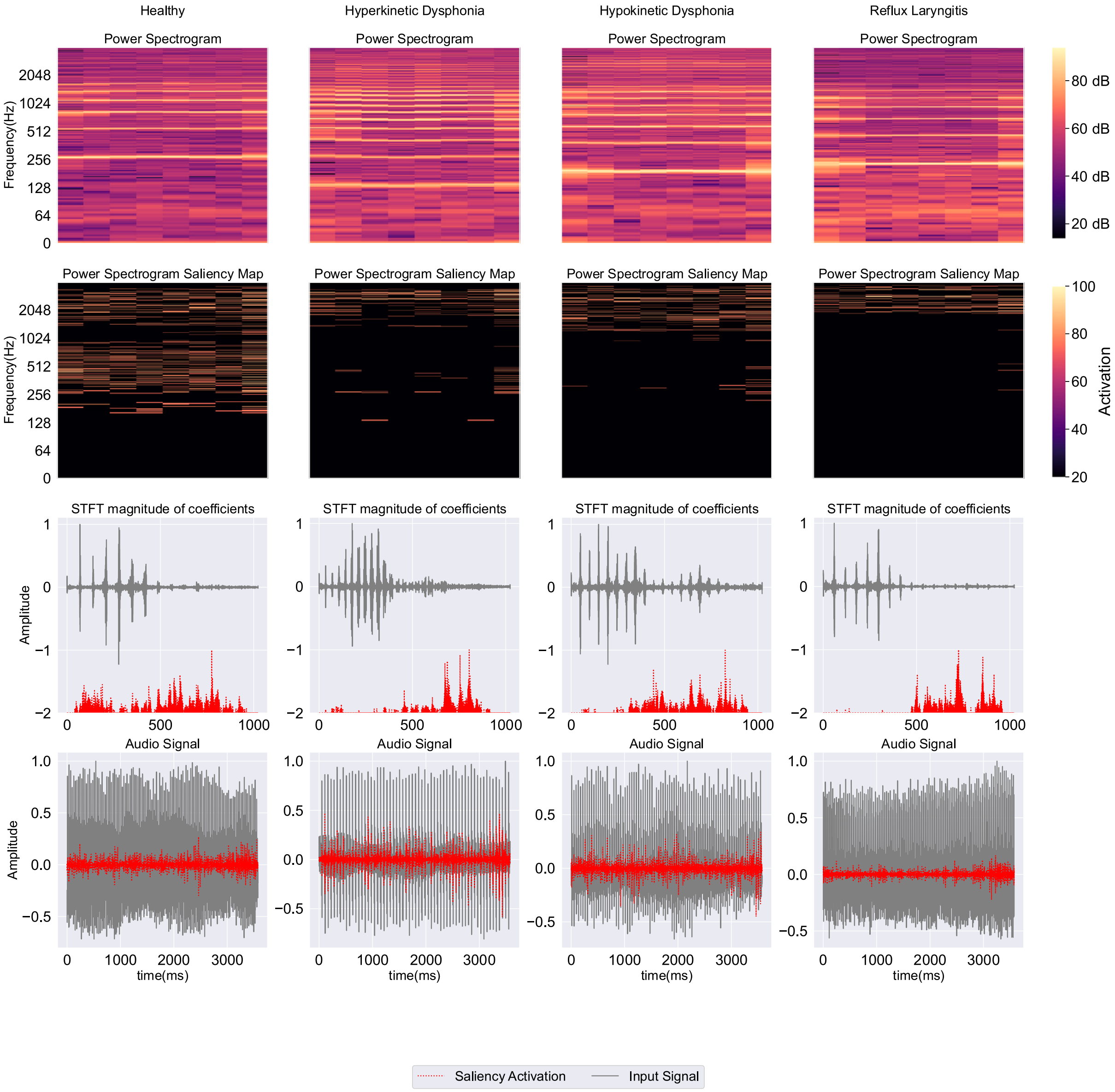


Fig. 20. Saliency activation map on the STFT model.

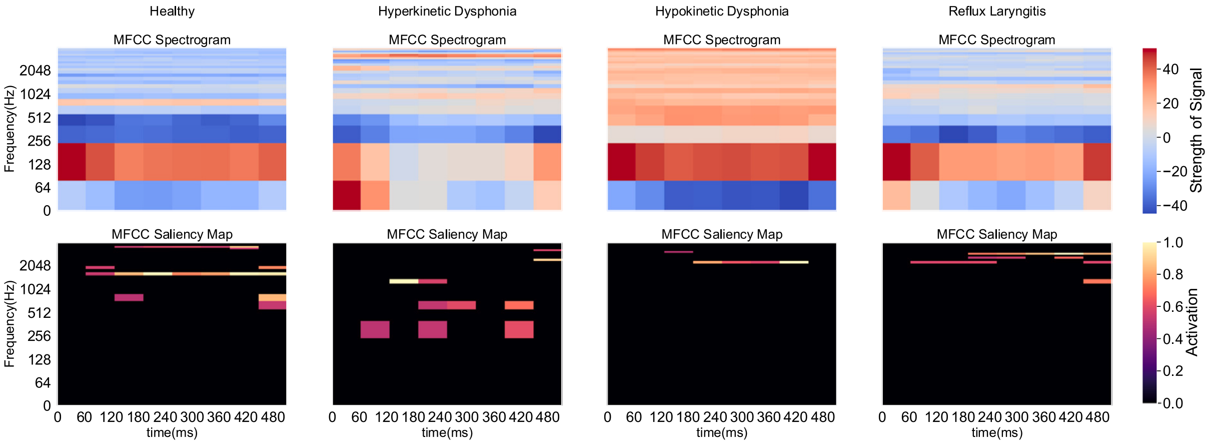


Fig. 21. Saliency activation map on the MFCC model.

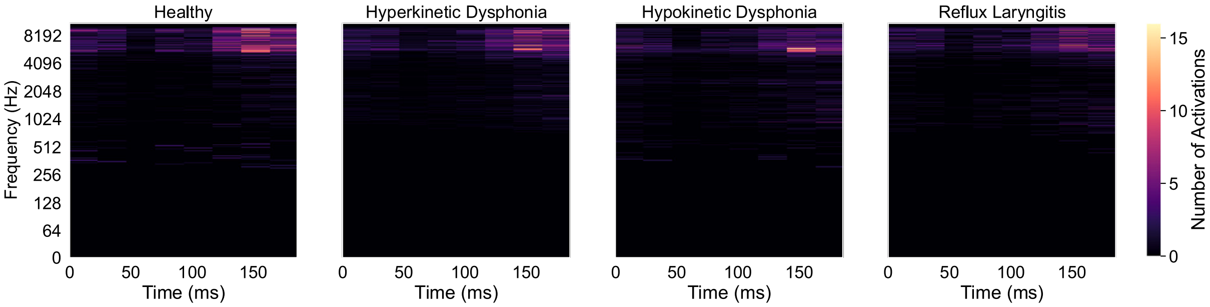


Fig. 22. Most common point of saliency activation on the STFT model.

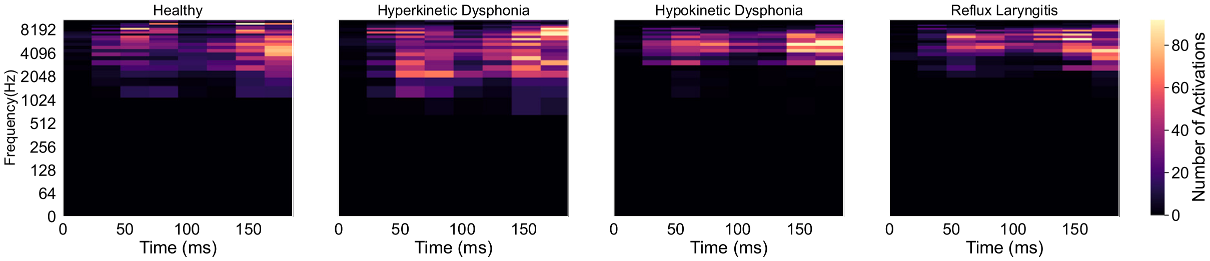


Fig. 23. Most common point of saliency activation on the MFCC model.

Table 7
Comparison on related works which deal with same dataset and task.

Reference	Methodology	Input feature	Accuracy (%)	Sensitivity (%)	Specificity (%)
[36]	IFCM	1D Audio signal	–	79.20	80.70
[36]	CGAN-FCM	1D Audio signal	–	81.60	83.30
[36]	CGAN-IFCM	1D Audio signal	–	90.10	89.40
[36]	SMOTE-IFCM	1D Audio signal	–	86.50	86.20
[36]	CSL-IFCM	1D Audio signal	–	85.60	85.00
[36]	Random Forests	1D Audio signal	–	77.10	78.50
[36]	Support Vector Machines	1D Audio signal	–	75.40	76.90
[54]	GoogleLeNet [66]	2D Amor Scalogram	80.59	80.59	93.53
[54]	GoogleLeNet [66]	2D Bump Scalogram	81.58	81.58	93.86
[54]	GoogleLeNet [66]	2D Morse Scalogram	87.83	87.83	95.94
[54]	MobileNet [65]	2D Amor Scalogram	86.18	86.18	95.39
[54]	MobileNet [65]	2D Bump Scalogram	82.24	82.24	94.08
[54]	MobileNet [65]	2D Morse Scalogram	86.84	86.84	95.61
[54]	Custom 3D-CNN	3D Group Decision Analogy	97.70	97.70	99.23
Proposed	***Custom MLP***	***OpenSmile***	***97.10***	***98.30***	***99.10***
Proposed	***Custom 1D CNN***	***STFT***	***97.10***	***99.80***	***99.20***
Proposed	***Custom 2D CNN***	***MFCC***	***97.10***	***98.30***	***99.00***

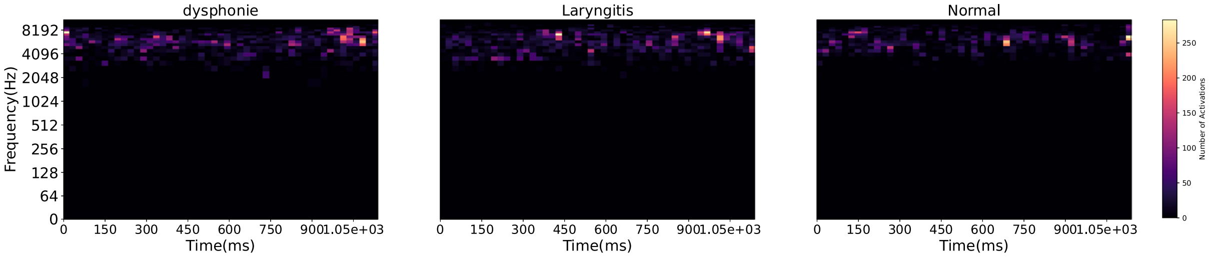


Fig. 24. MFCC saliency activation on SVD database voice samples.

been extensively studied, multi-class classification has received limited attention in the literature. By investigating multi-class classification, we aim to identify more complex features and their relationship with their target class, leading to a more comprehensive analysis of the feature set.

The only known work on the multi-class classification of the VOICED dataset is by Wahengbam et al. (2021) [54]. The noisy nature of the VOICED dataset makes multi-class classification challenging. However, we have addressed this by utilizing a heavier UNet++ autoencoder, demonstrating significant improvement in classification accuracy compared to the lighter autoencoder network used by Wahengbam et al. [54]. Moreover, both research works have considered the small size of the VOICED dataset, with only 206 unique audio samples, by applying necessary pre-processing techniques, resulting in robust classification results.

In this research, we propose three deep learning models (MLP, STFT, and MFCC) to classify voice disorders. While all the models achieved a high 97.1% accuracy rate, the STFT model showed a significant increase in sensitivity, specificity, and F1 score, achieving a 99% accuracy rate in the Reflux Laryngitis category. The MFCC model had significantly fewer parameters than the STFT and MLP models and attained 100% accuracy on the Reflux Laryngitis class. Thus the MFCC model demonstrated a much simpler and significantly lighter 1D CNN model, compared to the work by Wahengbam et al. which utilized a 3D CNN and a custom group analogy architecture and achieved an accuracy of 97.7%, which is only 0.6% higher than the MFCC model. This highlights the potential of the MFCC data format to extract crucial features in a condensed space and utilize a lighter DL model for pathological multi-class classification.

We conducted a saliency activation analysis of each model. The top five activated features for each class were explored, revealing that most of the feature set consists of Spectral mean amplitude and its variations, indicating a heavy reliance on amplitude and frequency components of the signal. The STFT and MFCC models activated in the inaudible regions of low amplitude and high frequencies, corresponding to high-pitch sounds. Training the models on this time–frequency wavelength can improve performance, and tools such as microphones with high-frequency response, contact microphones, and ultrasonic detectors can be utilized for data collection to enhance the accuracy of the DL models further.

7. Conclusion and future works

This study aims to identify critical features of voice abnormalities by conducting saliency analysis on the multi-class classification of three voice pathologies and a healthy signal. The data was pre-processed and denoised using the UNet++ architecture, demonstrating robust performance on the machine learning classifiers. The classifiers included a Multi-Layer Perceptron to identify crucial handcrafted features and two deep learning models based on Short-Time Fourier Transform (STFT) and Mel-frequency cepstral coefficients (MFCC) transformed input of the audio signal, which provided information on the nature of the detected signal. The three models achieved an accuracy of 97.1%. Among them, the STFT model exhibited the highest sensitivity, reaching 99.8%. On the other hand, the MFCC model demonstrated comparable performance, with 100% accuracy on the reflux laryngitis class. Moreover, the MFCC model was six times smaller than that of the STFT model. The saliency analysis of the data revealed that detecting voice abnormalities requires the identification of regions of inaudible high-pitched sounds with low amplitude and high frequency. These findings have potential applications in clinical trials, improving existing deep learning systems and the development of specialized tools for audio capturing, such as contact microphones and ultrasonic detectors. Future studies could explore these characteristics further to enhance the performance of classifying voice abnormalities by capturing or filtering the audio signal only in the specific feature space.

The work has contributed to the voice pathology field by focusing on the multi-class classification of pathological and healthy voices. One of the future scopes is to expand the dataset used in this research. The VOICED dataset, although valuable, is relatively small. Including a larger and more diverse dataset would enhance the generalizability of our findings and further validate the performance of the proposed models. Another future direction is to explore additional deep learning architectures and techniques. While our research utilized MLP, STFT, and MFCC models, various other architectures such as recurrent neural networks (RNNs) or transformer-based models can be investigated. Exploring these alternatives may provide further insights and potentially improve classification accuracy.

Furthermore, extending the research to have a broader range of voice pathologies would be beneficial. This would involve incorporating additional classes of voice disorders to develop a more comprehensive and clinically relevant classification system. Expanding the scope of voice pathologies would enable better detection and diagnosis of a broader range of voice disorders. Additionally, it is essential to consider the limitations of our research. Although the UNet++ autoencoder proved effective in handling the dataset's noisy nature, other noise reduction techniques could yield even better results. Exploring advanced noise reduction algorithms and incorporating them into the classification pipeline could further enhance the accuracy and robustness of the models. Lastly, our research focused on analyzing audio signals and their spectral features. Future studies could explore the integration of additional data modalities, such as textual or visual information, to improve classification performance. Combining multiple modalities may provide a more comprehensive understanding of voice disorders and enable more accurate classification.

Voice is increasingly used in future healthcare systems: voice biomarkers remotely monitor essential health parameters. They are used to comprehensively phenotype patients or design innovative experiments, paving the way to precision medicine [67]. Hence, it is imperative to integrate novel technologies (including the proposed research) into clinical practice. Moreover, clinical trials are needed to validate the technology for practical use. For the field to mature, we need to move from a technology-based approach to a more health-focused one, creating research and valuable data sets demonstrating such an approach's benefits. In conclusion, this study provides valuable insights into identifying voice abnormalities through saliency analysis. It demonstrates the potential for further research to improve the accuracy of classification systems and tools for audio capturing.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

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.

Availability of data and materials/code

Data and code will be made available on request

Appendix

Features in Saliency activation of each class as shown in Fig. 19 from left to right are given as from top to bottom

- **Healthy:**
  ‘jitterLocal_sma_quartile2’,
  ‘pcm_zcr_sma_amean’,
  ‘pcm_fftMag_spectralEntropy_sma_lpgain’,
  ‘mfcc_sma_de[14]_stddev’,
  ‘pcm_fftMag_spectralSlope_sma_quartile3’
- **Hyperkinetic Dysphonia:**
  ‘pcm_zcr_sma_amean’,
  ‘jitterLocal_sma_quartile2’,
  ‘audSpec_Rfilt_sma_de[5]_posamean’,
  ‘F0final_sma_ff0_maxSegLen’,
  ‘pcm_fftMag_spectralFlux_sma_quartile2’
- **Hypokinetic Dysphonia:**
  ‘voicingFinalUnclipped_sma_lpgain’,
  ‘pcm_fftMag_spectralSlope_sma_de_iqr1-3’,
  ‘pcm_fftMag_fband1000-4000_sma_de_peakMeanMeanDist’,
  ‘pcm_RMSenergy_sma_percentile99.0’,
  ‘pcm_fftMag_spectralHarmonicity_sma_de_rqmean’

- **Reflux Laryngitis:** *'pcm_fftMag_spectralSlope_sma_de_iqr1-3', 'voicingFinalUnclipped_sma_lpgain', 'voicingFinalUnclipped_sma_percentile99.0', 'pcm_fftMag_fband1000-4000_sma_de_peakMeanMeanDist', 'audSpec_Rfilt_sma_de[5]_posamean'*

References

[1] G. Muhammad, M. Melhem, Pathological voice detection and binary classification using MPEG-7 audio features, Biomed. Signal Process. Control 11 (2014) 1–9.

[2] M.E. Powell, D.D. Deliyski, S.M. Zeitels, J.A. Burns, R.E. Hillman, T.T. Gerlach, D.D. Mehta, Efficacy of videostroboscopy and high-speed videoendoscopy to obtain functional outcomes from perioperative ratings in patients with vocal fold mass lesions, J. Voice 34 (5) (2020) 769–782.

[3] S. Hegde, S. Shetty, S. Rai, T. Dodderi, A survey on machine learning approaches for automatic detection of voice disorders, J. Voice 33 (6) (2019) 947–e11.

[4] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, K.H. Malki, T.A. Mesallam, M.F. Ibrahim, Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions, Ieee Access 6 (2017) 6961–6974.

[5] F.T. Al-Dhief, M.M. Baki, N.M.A. Latiff, N.N.N.A. Malik, N.S. Salim, M.A.A. Albader, N.M. Mahyuddin, M.A. Mohammed, Voice pathology detection and classification by adopting online sequential extreme learning machine, IEEE Access 9 (2021) 77293–77306.

[6] N. Steffen, V.P. Vieira, R.K. Yazaki, P. Pontes, Modifications of vestibular fold shape from respiration to phonation in unilateral vocal fold paralysis, J. Voice 25 (1) (2011) 111–113.

[7] V. Uloza, A. Verikas, M. Bacauskiene, A. Gelzinis, R. Pribuisiene, M. Kaseta, V. Saferis, Categorizing normal and pathological voices: automated and perceptual categorization, J. Voice 25 (6) (2011) 700–708.

[8] A. Yamauchi, H. Yokonishi, H. Imagawa, K.-I. Sakakibara, T. Nito, N. Tayama, T. Yamasoba, Quantitative analysis of digital videokymography: a preliminary study on age-and gender-related difference of vocal fold vibration in normal speakers, J. Voice 29 (1) (2015) 109–119.

[9] S. Jothilakshmi, Automatic system to detect the type of voice pathology, Appl. Soft Comput. 21 (2014) 244–249.

[10] K. Muraleedharan, K.B. Kumar, S. John, R.S. Kumar, Combined use of nonlinear measures for analyzing pathological voices, Int. J. Image Graph. (2023) 2450035.

[11] L. Chen, J. Chen, Deep neural network for automatic classification of pathological voice signals, J. Voice 36 (2) (2022) 288–e15.

[12] A. Tegene, Q. Liu, Y. Gan, T. Dai, H. Leka, M. Ayenew, Deep learning and embedding based latent factor model for collaborative recommender systems, Appl. Sci. 13 (2) (2023) 726.

[13] J. Crabbé, Z. Qian, F. Imrie, M. van der Schaar, Explaining latent representations with a corpus of examples, Adv. Neural Inf. Process. Syst. 34 (2021) 12154–12166.

[14] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning (still) requires rethinking generalization, Commun. ACM 64 (3) (2021) 107–115.

[15] X. Liu, R. Zhang, Z. Meng, R. Hong, G. Liu, On fusing the latent deep CNN feature for image classification, World Wide Web 22 (2019) 423–436.

[16] P. Antoniadis, Latent space in deep learning, 2023, URL https://www.baeldung.com/cs/dl-latent-space.

[17] A. Sellami, S. Tabbone, Deep neural networks-based relevant latent representation learning for hyperspectral image classification, Pattern Recognit. 121 (2022) 108224.

[18] G. Mumović, M. Veselinović, T. Arbutina, R. Škrbić, Vocal therapy of hyperkinetic dysphonia, Srpski Arhiv Za Celokupno Lekarstvo 142 (11–12) (2014) 656–662.

[19] A. Nacci, B. Fattori, V. Mancini, E. Panicucci, J. Matteucci, F. Ursino, S. Berrettini, Posturographic analysis in patients with dysfunctional dysphonia before and after speech therapy/rehabilitation treatment, Acta Otorhinolaryngol. Ital. 32 (2) (2012) 115.

[20] A.M. Campagnolo, J. Priston, R.H. Thoen, T. Medeiros, A.R. Assunção, Laryngopharyngeal reflux: diagnosis, treatment, and latest research, Int. Arch. Otorhinolaryngol. 18 (2014) 184–191.

[21] U. Cesari, G. De Pietro, E. Marciano, C. Niri, G. Sannino, L. Verde, A new database of healthy and pathological voices, Comput. Electr. Eng. 68 (2018) 310–321.

[22] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, Springer, 2018, pp. 3–11.

[23] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM International Conference on Multimedia, 2010, pp. 1459–1462.

[24] N. Narendra, P. Alku, Dysarthric speech classification using glottal features computed from non-words, words and sentences, in: Speech Commun. Assoc. (INTERSPEECH), 2018, pp. 3403–3407, http://dx.doi.org/10.21437/Interspeech.2018-1059.

[25] S. Hadjitodorov, B. Boyanov, B. Teston, Laryngeal pathology detection by means of class-specific neural maps, IEEE Trans. Inf. Technol. Biomed. 4 (1) (2000) 68–73, http://dx.doi.org/10.1109/4233.826861.

[26] M. Akay, Time frequency and wavelets in biomedical signal processing, Biomed. Eng. (1998).

[27] B. Boyanov, S. Hadjitodorov, Acoustic analysis of pathological voices. A voice analysis system for the screening of laryngeal diseases, IEEE Eng. Med. Biol. Mag. 16 (4) (1997) 74–82, http://dx.doi.org/10.1109/51.603651.

[28] D. Michaelis, T. Gramss, H. Strube, Glottal-to-noise excitation ratio-A new measure for describing pathological voices, Acta Acust. United Acust. 83 (4) (1997) 700–706.

[29] H. Kasuya, S. Ogawa, K. Mashima, S. Ebihara, Normalized noise energy as an acoustic measure to evaluate pathologic voice, J. Acoust. Soc. Am. 80 (5) (1986) 1329–1334, http://dx.doi.org/10.1121/1.394384.

[30] L. Gavidia-Ceballos, J. Hansen, Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection, IEEE Trans. Biomed. Eng. 43 (4) (1996) 373–383, http://dx.doi.org/10.1109/10.486257.

[31] J. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, P. Gómez-Vilda, M. Blanco-Velasco, F. Cruz-Roldán, The effectiveness of the glottal to noise excitation ratio for the screening of voice disorders, J. Voice 24 (1) (2010) 47–56, http://dx.doi.org/10.1016/j.jvoice.2008.04.006.

[32] X. Xie, H. Cai, C. Li, F. Ding, A voice disease detection method based on MFCCs and shallow CNN, 2023, arXiv preprint arXiv:2304.08708.

[33] J. Godino-Llorente, P. Gomez-Vilda, Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors, IEEE Trans. Biomed. Eng. 51 (2) (2004) 380–384, http://dx.doi.org/10.1109/tbme.2003.820386.

[34] Y. Jiao, M. Tu, V. Berisha, J. Liss, Simulating dysarthric speech for training data augmentation in clinical speech applications, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 6009–6013, http://dx.doi.org/10.1109/icassp.2018.8462290.

[35] B. Vachhani, C. Bhat, S. Kopparapu, Data augmentation using healthy speech for dysarthric speech recognition, in: Interspeech 2018, ISCA, 2018, pp. 471–475, http://dx.doi.org/10.21437/interspeech.2018-1751.

[36] K. Chui, M. Lytras, P. Vasant, Combined generative adversarial network and fuzzy C-means clustering for multi-class voice disorder detection with an imbalanced dataset, Appl. Sci. 10 (13) (2020) 4571, http://dx.doi.org/10.3390/app10134571.

[37] E. Fonseca, R. Guido, P. Scalassara, C. Maciel, J. Pereira, Wavelet time-frequency analysis and least squares support vector machines for the identification of voice disorders, Comput. Biol. Med. 37 (4) (2007) 571–578, http://dx.doi.org/10.1016/j.compbiomed.2006.08.008.

[38] G. Muhammad, T. Mesallam, K. Malki, M. Farahat, A. Mahmood, M. Alsulaiman, Multidirectional regression (MDR)-based features for automatic voice disorder detection, J. Voice 26 (6) (2012) 817.e19–817.e27, http://dx.doi.org/10.1016/j.jvoice.2012.05.002.

[39] T. Mesallam, M. Farahat, K. Malki, M. Alsulaiman, Z. Ali, A. Al-Nasheri, G. Muhammad, Development of the arabic voice pathology database and its evaluation by using speech features and machine learning algorithms, J. Healthc. Eng. 2017 (8783751) (2017) 1–13, http://dx.doi.org/10.1155/2017/8783751.

[40] J. Godino-Llorente, R. Fraile, N. Sáenz-Lechón, V. Osma-Ruiz, P. Gómez-Vilda, Automatic detection of voice impairments from text-dependent running speech, Biomed. Signal Process. Control 4 (3) (2009) 176–182, http://dx.doi.org/10.1016/j.bspc.2009.01.007.

[41] E. Weber, The massachusetts eye and ear infirmary illustrated manual of ophthalmology, 3rd edition, J. Neuro-Ophthalmol. 30 (1) (2010) 106, http://dx.doi.org/10.1097/01.wno.0000369166.94555.db.

[42] L. Verde, G. De Pietro, M. Alrashoud, A. Ghoneim, K. Al-Mutib, G. Sannino, Leveraging artificial intelligence to improve voice disorder identification through the use of a reliable mobile app, IEEE Access 7 (2019) 124048–124054, http://dx.doi.org/10.1109/access.2019.2938265.

[43] L. Chen, C. Wang, J. Chen, Z. Xiang, X. Hu, Voice disorder identification by using Hilbert-huang transform (HHT) and K nearest neighbor (KNN), J. Voice 35 (6) (2021) 932.e1–932.e11, http://dx.doi.org/10.1016/j.jvoice.2020.03.009.

[44] F. Amara, M. Fezari, H. Bourouba, An improved GMM-SVM system based on distance metric for voice pathology detection, Appl. Math. 10 (3) (2016) 1061–1070.

[45] P. Harar, J. Alonso-Hernandezy, J. Mekyska, Z. Galaz, R. Burget, Z. Smekal, Voice pathology detection using deep learning: a preliminary study, in: 2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI), IEEE, 2017, pp. 1–4, http://dx.doi.org/10.1109/iwobi.2017.7985525.

[46] S. Kadiri, P. Alku, Analysis and detection of pathological voice using glottal source features, IEEE J. Sel. Top. Sign. Proces. 14 (2) (2020) 367–379, http://dx.doi.org/10.1109/jstsp.2019.2957988.

[47] J. Wang, H. Xu, X. Peng, J. Liu, C. He, Pathological voice classification based on multi-domain features and deep hierarchical extreme learning machine, J. Acoust. Soc. Am. 153 (1) (2023) 423–435, http://dx.doi.org/10.1121/10.0016869.

[48] M. Altayeb, A. Al-Ghraibah, Classification of three pathological voices based on specific features groups using support vector machine, Int. J. Electr. Comput. Eng. (IJECE) 12 (1) (2022) 946, http://dx.doi.org/10.11591/ijece.v12i1.pp946-956.

[49] S.P. Kumar, N. Narayanan, J. Ramachandran, B. Thangavel, Convolutional neural network for voice disorders classification using kymograms, Biomed. Signal Process. Control 86 (2023) 105159.

[50] H. Kim, H.-Y. Park, D. Park, S. Im, S. Lee, Non-invasive way to diagnose dysphagia by training deep learning model with voice spectrograms, Biomed. Signal Process. Control 86 (2023) 105259.

[51] M. Huckvale, Z. Liu, C. Buciuleac, Automated voice pathology discrimination from audio recordings benefits from phonetic analysis of continuous speech, Biomed. Signal Process. Control 86 (2023) 105201.

[52] J.-Y. Han, C.-J. Hsiao, W.-Z. Zheng, K.-C. Weng, G.-M. Ho, C.-Y. Chang, C.-T. Wang, S.-H. Fang, Y.-H. Lai, Enhancing the performance of pathological voice quality assessment system through the attention-mechanism based neural network, J. Voice (2023).

[53] B. Dianat, P. La Torraca, A. Manfredi, G. Cassone, C. Vacchi, M. Sebastiani, F. Pancaldi, Classification of pulmonary sounds through deep learning for the diagnosis of interstitial lung diseases secondary to connective tissue diseases, Comput. Biol. Med. 160 (2023) 106928.

[54] K. Wahengbam, M.P. Singh, K. Nongmeikapam, A.D. Singh, A group decision optimization analogy-based deep learning architecture for multiclass pathology classification in a voice signal, IEEE Sens. J. 21 (6) (2021) 8100–8116.

[55] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J.K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, K. Evanini, The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language, in: 17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), Vols 1-5, Vol. 8, ISCA, 2016, pp. 2001–2005.

[56] S.S. Nayak, A.D. Darji, P.K. Shah, Machine learning approach for detecting Covid-19 from speech signal using mel frequency magnitude coefficient, Signal, Image Video Process. (2023) 1–8.

[57] B. Woldert-Jokisz, Saarbruecken voice database, 2007.

[58] L.M. Jesus, I. Belo, J. Machado, A. Hall, The advanced voice function assessment databases (AVFAD): Tools for voice clinicians and speech research, in: Advances in Speech-Language Pathology, IntechOpen, 2017.

[59] A.L. Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000) e215–e220.

[60] Y.M. Jung, A review on denoising, J. Korean Soc. Ind. Appl. Math. 18 (2) (2014) 143–156.

[61] M. Michelashvili, L. Wolf, Audio denoising with deep network priors, 2019, arXiv preprint arXiv:1904.07612.

[62] P.J. Huber, Robust estimation of a location parameter, in: Breakthroughs in Statistics: Methodology and Distribution, Springer, 1992, pp. 492–518.

[63] J. Salamon, C. Jacoby, J.P. Bello, A dataset and taxonomy for urban sound research, in: Proceedings of the 22nd ACM International Conference on Multimedia, 2014, pp. 1041–1044.

[64] A.H. Al-Noori, P. Duncan, Robust speaker recognition in noisy conditions by means of online training with noise profiles, J. Audio Eng. Soc. 67 (4) (2019) 174–189.

[65] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017, arXiv preprint arXiv:1704.04861.

[66] X. Zhang, Y. Zou, W. Shi, Dilated convolution neural network with LeakyReLU for environmental sound classification, in: 2017 22nd International Conference on Digital Signal Processing (DSP), Ieee, 2017, pp. 1–5.

[67] G. Fagherazzi, A. Fischer, M. Ismael, V. Despotovic, Voice for health: the use of vocal biomarkers from research to clinical practice, Digit. Biomark. 5 (1) (2021) 78–88.