Turkish Journal of Electrical Engineering and Computer Sciences

Volume 32 | Number 4 Article 7

7-26-2024

Multi-label voice disorder classification using raw waveforms

GÖKAY DİŞKEN

Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2024) 32: 590 – 604

doi:10.55730/1300-0632.4089

Multilabel voice disorder classification using raw waveforms

Gökay DİŞKEN*
Department of Artificial Intelligence Engineering, Faculty of Computer and Informatics,
Adana Alparslan Türkeş Science and Technology University, Adana, Turkiye

Received: 27.03.2024 • **Accepted/Published Online:** 07.06.2024 • **Final Version:** 26.07.2024

Abstract: Automated voice disorder systems that distinguish pathological voices from healthy ones have been developed with the aid of machine learning methods. Both clinicians and patients can benefit from these systems as they provide many advantages, compared to the invasive techniques. These systems can produce binary (healthy/pathological) or multiclass (healthy/selected pathologies) decisions. However, multiple disorders might exist in an individual's voice. Multilabel classification should be considered in such cases. By this time, only a single report is available on this topic, where hand-crafted features were used, and a data augmentation technique was utilized to overcome class imbalances. In this study, a similar experimental setup is followed to investigate the suitability of raw voice signals as inputs for multilabel classification. A deep learning model which consists of residual blocks and a novel gating mechanism is proposed. The gating mechanism weighs the channels of a residual block's output based on both its output and the previous layer's output. Using a SincNet filterbank that operates directly on the raw waveform as the initial layer, 0.99 accuracy and 0.98 F1 score were observed for natural /a/ vowels of Saarbruecken Voice Database with time domain augmentation to balance the class samples. On the other hand, reducing the number of augmented samples decreased the performance for both systems, indicating the need for a balanced dataset to avoid oversampling underrepresented classes. The proposed architecture performed consistently better than ResNet18 with deep connected attention, which verified the effectiveness of the proposed gating mechanism.

Key words: Convolutional neural network, deep learning, multilabel classification, voice pathology

1. Introduction

Speech is the most natural way of communication between people. It is a naturally information-rich signal, including speaker identity, sex, language, and emotion. It even includes information regarding the mental health status of the speaker [1]. Abnormal changes in the human voice generation system may lead to voice disorders. An earlier study conducted in United States reported that about 30% of the general population have experienced a voice disorder during their lifetime [2]. For teachers, the prevalence goes up to 58%, indicating that voice disorders are more common for professional groups who may overuse their voice such as teachers, singers, and actors. Despite the fact that voice disorders are very common, its diagnosis is difficult due to several factors stated in several studies [3–6]. Absence of standardized terminology, subjective evaluation, symptoms ignored by patients, difficult and expensive clinical processes are among those factors.

Automated speech pathology detection aims to solve the abovementioned difficulties with the aid of machine learning. Modern machine learning, and more specifically deep learning, algorithms have made huge impacts in many diverse areas. Voice signal analysis for health is an emerging topic where the researchers aim

*Correspondence: gdisken@atu.edu.tr

to develop systems that will objectively and accurately detect disorders and diseases. Recent surveys showed that there is an active and increasing research interest on this topic, with a particular emphasis on deep learning [3, 6–8].

Popular datasets for automated voice disorder detection are Saarbruecken Voice Database (SVD), Massachusetts Eye and Ear Infirmary (MEEI), Arabic Voice Pathology Database (AVPD), and VOice ICar fEDerico II (VOICED). Details and a metaanalysis for SVD, MEEI and AVPD sets can be found in [9] and details of the VOICED set can be found in [10]. SVD is one of the widely used datasets, and it is also utilized in this study. The SVD set provides multilabels, i.e. multiple voice disorders are present in the same individual. The majority of the reported studies in automated voice disorder detection area focused on binary classification (healthy or pathological). The pathological class can be a single disorder or can consist of a selected subset of disorders such as dysphonia, laryngitis, or it can consist of all available disorders for the dataset (for instance, 71 pathologies are available for the SVD dataset). A recent study [11] considered multiple voice disorders in the same individual, which reflects real-life scenarios, thus has a practical importance. Besides that, there is no other report regarding multilabel voice disorder detection, to the best of the author's knowledge. The primary reason for this may be the limited data. In [11], data augmentation was applied to reduce imbalanced classes. Another issue with the multilabel classification is that the system must detect all labels of a given sample to achieve a perfect score. Missing one or more labels will decrease the performance; hence it is a relatively harder task compared to single label or binary classification tasks.

In this study, the same multiclass classification with multilabel conditions used in [11] is investigated for deep learning models based on ResNet [12] style blocks. A novel gating mechanism between residual blocks proposed in this study. Proposed architecture directly operates on raw waveforms, eliminating the feature extraction processes contrary to the vast amount of the studies in the related literature. Using raw waveforms produced state-of-the-art results for various speech related studies such as speech recognition [13], speaker recognition [14]. For voice disorder detection, raw waveforms were considered in few studies [15–17]. However, these studies used speech frames as the input to the deep models. In this study, entire utterances are used as inputs, without dividing them into short segments. To verify the effectiveness of the proposed gating technique, ResNet18 with deep connected attention (DCA) model is used for comparison. DCA was proposed in [18] and applied to voice disorder detection in [19]. Also in [19], it was reported that ResNet18 generally achieved a higher performance compared to ResNet34 and ResNet50 models; hence, ResNet18 was chosen as a competitive candidate for assessing the performance of the proposed approach.

The highlights and contributions of this study can be summarized as follows:

- Raw waveforms are directly fed to deep models for multilabel multiclass voice disorder classification. To the best of the author's knowledge, this is the first study to utilize raw wave for multilabel voice disorder classification.
- A novel gating mechanism is proposed. The experimental results showed that the proposed method can achieve better results compared to deep connected attention method.
- Compared to the hand-crafted features [11], which also forms the basis of this study, better performances were observed with raw waveforms for the augmented dataset. Similar performances were observed when the number of augmented samples were reduced.

The rest of the paper is organized as follows: Section 2 reviews some of the most recent studies on automated voice disorder detection. Section 3 proposes the deep model used in this study. Section 4 gives the details of the experimental setup including the SVD dataset and the obtained results. The experimental results are discussed in Section 5 and possible future research topics are mentioned. Finally, Section 6 concludes the paper.

2. Related work

Many different hand-crafted features have been proposed for voice pathology detection. Similarly, various machine learning algorithms have been used for classification. In this section, some of the most recent studies are addressed and then difficulties in this particular research topic are discussed briefly. The motivation behind this study is given at the end of this section.

In [20], convolutional neural networks (CNNs) were used for healthy-pathological classification (six pathologies from SVD dataset). Spectrograms extracted from sustained vowel /a/ at neutral pitch were used as inputs in this case. The reported values were 0.77 for accuracy and 0.78 for F1 score. Convolutional deep belief network was used in a following study [21] to pretrain weights of CNN; however, the classification accuracy decreased to 0.71. CNN and long short-term memory (LSTM) models were compared in [22] and they performed similarly, where the inputs were 27 dimensional vectors consists of various hand-crafted features. In [23–26], many different hand-crafted features were also used. VoiceLens is a system that combines deep learning based features with hand-crafted features [27] which produced 97.5% accuracy for healthy and six different pathology classes. Octave scaled spectrogram is used in [28] as input to pretrained ResNet34 model, 96.11% accuracy was reported. Pretrained VGG-19 and SVM were employed in [29] to detect voice disorders using /i/ vowels from SVD dataset. Stacked sparse autoencoder was used in [30] for VOICED dataset. CNN-LSTM model with sinusoidal rectified unit activation was proposed in [31] for multiclass classification. Bi-LSTM with constant-Q cepstral coefficients was found to be superior to MFCCs in [4]. Self-attention based LSTM was proposed in [32] where severities of pathological voice were classified. MFCCs and spectrograms were used as inputs for multibranch CNN model to detect dysphagia in [33].

Raw speech and glottal flow waveforms were used as input to CNN architectures [15], and obtained better results with glottal flow features. Similar results were observed in [34] for both SVD and MEEI datasets. Raw waveforms were also used in [35] where SincNet [36] was found to be more effective than traditional convolutional layers for FEMH Speech Disorders database. Raw waveform segments were fed into 1-D convolutional layers in [16] and severity of pathological voice was classified. In [17], raw wave segments were used as inputs to a deep model which consists of CNN, LSTM, and dense (fully-connected) layers. An overall accuracy of 68.08% was observed for binary classification considering healthy and pathological classes (consists of all 71 pathologies) based on the experiments conducted on the sustained vowel /a/ of the SVD data. Another model with raw inputs was proposed in [37] where the data was transformed into 100×100 matrix to fed 2-D convolutional layers.

Log mel-frequency spectral coefficients were used to train DCA ResNet model [19] which detects healthy and pathological (four different pathologies) classes. The DCA part is the same as squeeze-excitation block [38]. The attention modules between the adjacent ResNet blocks were connected to prevent frequent information changes between the attention modules. Multimodal transmission network [39] is another architecture that controls the information flow through the network via multimodal transfer module. Other multimodal studies include [40, 41].

The recent studies mentioned in the previous paragraphs showed that there is no standard procedure to evaluate machine learning algorithms for voice pathology detection. Varying numbers of healthy and pathological voices were used in the reported studies, which makes direct comparisons between proposed systems harder. In fact, class imbalances were analyzed in [42] for MEEI and SVD datasets, fuzzy clustering synthetic minority oversampling technique (FC-SMOTE) was used to increase the balance between classes, which led to clear performance improvements. Different data augmentation strategies were investigated for voice pathology detection in [43]. Random Gaussian noise was used in [44] and time stretching was used in [32] for data augmentation purposes. Hence, the data itself can be considered as a limiting factor for this area. On the other hand, creating a balanced, labelled, real-life dataset is very hard as stressed in [11]. Further, in [44], three previously published systems were reimplemented in order to compare their performances on the same data and performance metric (performance metrics varies between studies too, although accuracy and F1 score values are preferred in general).

Another issue related to this subject is the chosen classes. Most of the studies dealt with binary case (i.e. healthy-pathological). However, the number of selected pathologies varies as some studies considered all available pathologies, some focused on a small number of pathologies. A limited number of researchers considered multiclass case, for instance [11, 27, 39, 42, 45]. Multiple disorders in the same individual, i.e. multilabel for a single sample, have not been investigated except a recent study [11]. Therefore, more effort should be put through developing systems which can identify multiple disorders in an individual's voice. It is also more appropriate in terms of practical applications. With this regard, deep models operate on raw voice signals are developed in this study to fill the mentioned gap.

3. Proposed system

The authors of [46] argued that as a network gets deeper it performs better at semantic understanding tasks, but deeper networks may not be necessary for tasks that do not require high level semantic information. They developed a relatively shallow ResNet variant, namely Res-TSSDNet, for audio spoof detection. Following the same argument, the same residual blocks are used in this study since voice disorder detection does not require any semantic information. As the first convolutional layer that operates on the raw voice signal, both the traditional convolution as in the original Res-TSSDNet and sinc filters are considered. The purpose of SincNet filterbank [36], or sinc filters, is to extract more interpretable features compared to the traditional convolutional layers. Time-domain convolution is performed on the input waveform with sinc filters. The lowest and highest frequencies (cut-in and cut-off) are learned through training. Sinc filters were proven to be effective for voice disorder detection [35]. The fixed scale was found to be more effective than learnable scales for audio spoof detection [47], hence also preferred in this study. Both 1-D and 2-D configurations are considered. For 1-D, all convolutional layers in the model are 1-D convolutions. On the other hand, 2-D models consist of 2-D convolutions except the initial convolution/sinc filter layer. Once the initial layer is applied to the raw signal, a channel dimension is added to the obtained feature maps, hence transformed into a 2-D time-frequency representation [48].

Many modern CNN architectures exploit channel weights. Assigning different weights to each channel of convolutional block can be viewed as feature recalibration, attention, or gating mechanism. The purpose of this approach is to suppress less informative channels, hence better flow of information through the layers. For the considered task, the information can be understood as artifacts, corruptions, etc. found in the convolutional maps that will aid to the discrimination of healthy/pathological samples. One of the most popular methods

for this operation is the squeeze-and-excitation (SE) block [38], which uses global average pooling to "squeeze" global spatial information. Another method is convolutional block attention module (CBAM) [49] where channel attention is followed by spatial attention. Efficient channel attention (ECA) [50] utilizes 1-D convolution to create channel weights in a fast manner without dimensionality reduction, contrary to SE and CBAM. DCA adds connection between attention blocks so the layers can exploit both the extracted features and attention information of previous block. These types of gating mechanisms were also applied to speech related tasks. Channel-wise gating was implemented within Res2Net blocks [51, 52] for audio spoof detection and DCA was used for voice pathology detection [19].

The proposed model includes a novel gating mechanism which utilizes the multigroup channel-wise gate of [51] with inverted inputs and applies it between the residual blocks. Therefore, the deeper blocks can exploit the information obtained at the previous blocks. In [51], previous feature maps were gated based on the current feature maps within a Res2Net block. This approach lets integration of the previous information that is analogous to the current information. Contrary to this, in the proposed approach, the previous block's feature maps are used to gate the current block's feature map, hence the inverted inputs. The goal here is to create an information flow mechanism similar to the DCA approach while benefiting from multigroup attention. Contrary to [51], proposed approach will encourage the model to extract information analogous to the previous layers.

The proposed model is illustrated in Figure 1. The details of the model that are not shown in the figure for a compact illustration are as follows. After the initial convolution/sinc filters, batch normalization (BN), ReLU, and max pooling are applied. A total of four residual blocks are used with three gate blocks (the output of the first residual block is not gated). Note that 1×1 convolutions are applied to match the channel dimensions between residual blocks if necessary. After obtaining the final gated output, a global average pooling layer is applied. Hence, the following classification layer's input dimension is equal to the number of channels in the last residual block and the output dimension is equal to the number of classes. For a single labeled sample, the output for the corresponding class will be one, while the others are zero. For a multilabeled sample, all true labels will be one, while the others are zero. The inner architectures of the residual block and gate block are also shown in Figure 1. The skip connections joined to the main path via element-wise addition and element-wise multiplication for residual and gate blocks, respectively.

The mathematical expressions for the gating mechanism are given below for completeness. The first step is to apply average pooling to both inputs. It should be reminded that both 1-D and 2-D versions of the proposed model are considered. Therefore, average value is computed over the time dimension for 1-D case, and over the frequency and time dimensions for 2-D case (assuming time-frequency representation after the initial layer). Following [51], the average value for 2-D case can be computed as Eq. 1, where $\mathrm{F}_{\mathrm{avg}}$ is the average value, $R_i \in \mathbb{R}^{C \times S \times T}$ is the output of ith residual block with C channels. Eq. 2 shows the average value for the output of the previous block's max pooling layer.

$$F_{avg}(R_i) = \frac{1}{S \times T} \sum_{s=1}^{S} \sum_{t=1}^{T} R_i(:, s, t) \tag{1}$$

$$F_{avg}(M_{i-1}) = \frac{1}{S \times T} \sum_{s=1}^{S} \sum_{t=1}^{T} M_{i-1}(:, s, t) \tag{2}$$

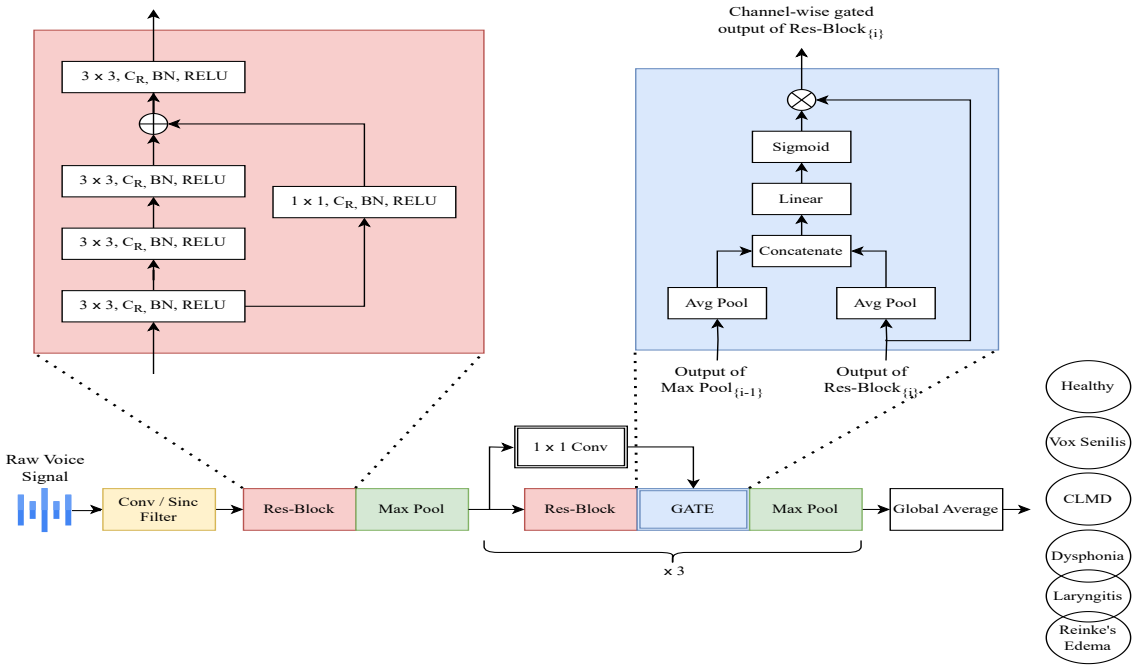

Figure 1. Proposed deep model with channel-wise gating between blocks. The model outputs represent single labels (healthy and five pathologies) and multilabels (dysphonia-laryngitis or laryngitis-Reinke's edema).

In the case of different channel dimensions, the max pooling output first goes through a 1×1 convolutional layer before being fed into the gate block. Once the average values are obtained, they can be concatenated and fed to a linear layer which transforms 2C channel dimensions back to C channels via a linear matrix $W \in \mathbb{R}^{2C \times C}$. By applying a sigmoid function (σ), channel weights are obtained. Eq. 3 gives the expression for channel weights (a_i),

$$a_i = \sigma\left\{W^T[F_{avg}(R_i)||F_{avg}(M_{i-1})]\right\} \tag{3}$$

where || represents concatenation. Note that for the 1-D case, the frequency dimension (S) is absent in Eqs. 1 and 2. Once the channel weights are computed, they are element-wise multiplied with R_i, creating the gated output. Table 1 shows the details such as channel dimensions, kernel sizes, and output sizes for the proposed deep model when the initial layer is the sinc filter. For the traditional convolution as the initial layer, kernel size is chosen as 7 and the channel dimension as 16, all other layers of the model are the same and the output sizes vary accordingly. An important difference between the 1-D and 2-D models is the kernel size of max pooling layers. For 1-D, all kernels are chosen as 4. On the other hand, 2-D models use (2, 5) and (1, 5) kernels. The reason for this choice is the dimension difference between frequency and time (70 and 23,750, respectively, for sinc filter). When using the same numbers for each axis, the frequency dimension gets smaller much faster than the time dimension. Hence, time dimension is reduced more aggressively than the frequency dimension by using the kernel sizes given in Table 1.

Table 1. Layer parameters of the proposed deep model for 1-D/2-D cases. The first layer is common for both cases, and the input is 1 × 24000 raw voice signal.

Layer	Kernel	Channel	Output size
Sinc Filters	1 × 251	70	70 × 23,750/1 × 70 × 23,750
Max Pool	4/(2, 5)	-	70 × 5937/1 × 35 × 4750
Res-Block 1	1 × 3/3 × 3	32/16	32 × 5937/16 × 35 × 4750
Max Pool 1	4/(1, 5)		32 × 1484/16 × 35 × 950
Res-Block 2	1 × 3/3 × 3	64/32	64 × 1484/32 × 35 × 950
Max Pool 2	4/(2, 5)		64 × 371/32 × 17 × 190
Res-Block 3	1 × 3/3 × 3	128/64	128 × 371/64 × 17 × 190
Max Pool 3	4/(1, 5)		128×92/64 × 17 × 38
Res-Block 4	1 × 3/3 × 3	128/128	128×92/128 × 17 × 38
Global Max Pool	-	-	128
Classification	-	-	6

4. Experiments

The experimental setup follows [11] as it is the only one to consider multilabel multiclass voice disorder detection for SVD dataset and will serve as a baseline. Organization of the data and performance metrics are chosen similarly. Although a direct comparison may not be meaningful due to the different setups, results from some of the recent deep learning models are given at the end of this section.

4.1. Dataset

SVD dataset[1] contains more than 2000 individuals' voice records. It consists of /a/, /i/, and /u/ vowels with different pitch options and a continuous sentence in German. /a/ vowels in normal pitch have been used in many studies and it reportedly achieved good detection performances. They were also used in this study. The dataset contains 71 different pathologies, but a subset of pathologies was used. For single label pathologies, Dysphonia (DYS), Reinke's edema (RDE), Vox Senilis (VSE), Laryngitis (LAR), Central Laryngeal Motion Disorder (CLMD) were chosen. For multilabel pathologies (i.e. multiple disorder in a voice signal) Dysphonia-Laryngitis (DYS-LAR) and Laryngitis-Reinke's edema (LAR-RDE) were chosen. Multilabel means that the voice record was labeled with "both" of the disorders. Therefore, it belongs to each of the labels. As an example, if a sample was labeled as DYS, then detecting dysphonia from the sample is necessary. If a sample was labeled as DYS-LAR, then detecting both disorders is necessary, missing any of them will lead to a reduced performance. Besides the chosen five pathological classes, healthy (HEA) class was also considered. Therefore, a total of six classes were available.

Table 2 shows the distribution of voice files for the chosen classes. The data is highly imbalanced. To overcome this situation, data augmentation methods in the time domain were applied to the original signals as they were found to be effective previously [32, 43]. Audiomentations[2] library was used to apply Gaussian noise, shift, time stretch, and polarity inversion transforms. Gaussian noise was added with high signal-to-noise ratios to avoid corrupting the original record. Shift means shifting the samples of the record forwards or backwards. Time stretch corresponds to changing the speed without affecting the pitch. Polarity inversion multiplies the

[1]Saarbruecken Voice Database [online]. Website https://www.stimmdatenbank.coli.uni-saarland.de [Accessed 28 November 2023].

[2]Audiomentations Python Library [online]. Website https://github.com/iver56/audiomentations [Accessed 14 January 2024].

waveform by –1, and thus, the sound remains the same, but the sample signs are inverted. All transformations applied randomly with random parameters (within a sensible range) to create as many different data as possible. Therefore, a created sample may be the result of more than one transformation.

The HEA class has 687 samples. The "Augmented" dataset has 687 samples for each class to match the number of HEA samples. Note that some classes have a very limited number of samples. Despite the randomly applied transformations as augmentation method, this fact will inevitably affect the results as in [11]. For this reason, another set ("Aug_{140}") with a smaller number of artificial data is created. The "Aug_{140}" dataset has 140 samples for each pathological class to analyze the effect of sample size on the detection performance. These augmented versions were akin to the augmented sets of [11]. As a final note, the sampling rate of SVD dataset was 50,000 but the raw voices were down sampled to 16,000. Also, the mean duration of the chosen utterances was found to be around 1.2 s. Therefore, all files were padded or truncated to give 24,000-dimensional vector per raw voice file. Padding was realized by replicating the required number of samples from the beginning of the file. For truncation, the samples after the 24000th sample were dropped.

Table 2. Number of samples per class for the original and augmented sets. Aug_{140} represents the case where pathology classes were augmented to 140 samples to imitate 20% balancing rate of [11].

	HEA	**DYS**	**LAR**	**RDE**	**VSE**	**CLMD**	**DYS-LAR**	**LAR-RDE**
Original	687	70	82	34	23	11	5	10
Augmented	687	687	687	687	687	687	687	687
Aug_140	687	140	140	140	140	140	140	140
[11] (95%)	686	621	648	627	638	650	648	648
[11] (20%)	686	137	137	132	132	130	136	135

4.2. Performance metrics

Widely used metrics such as precision (4), recall (5), F1 score (6), and accuracy (7) were used to assess the performance of the proposed system.

$$precision = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \cap Z_i|}{|Z_i|} \tag{4}$$

$$recall = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \cap Z_i|}{|Y_i|} \tag{5}$$

$$F1score = \frac{1}{m}\sum_{i=1}^{m}\frac{2|Y_i \cap Z_i|}{|Y_i| \cup |Z_i|} \tag{6}$$

$$accuracy = 1 - \frac{1}{m}\sum_{i=1}^{m}\frac{|Z_i \Delta Y_i|}{|L|} \tag{7}$$

In the given equations, Y_i represents the ith sample of the true set of labels, Z_i represents the ith instance of the predicted set of labels, Δ is the symmetric difference, and L is the total number of samples. The accuracy is calculated as "1-Hamming loss" since it is suitable for multilabel classification [11]. All labels must be correctly detected for the multilabeled samples in order to achieve a high performance with the considered metrics.

4.3. Training parameters

All deep learning processes were realized via PyTorch toolkit on an Nvidia RTX 3090 GPU. A 10-fold stratified crossvalidation strategy was followed, as it was common in the automated voice disorder detection literature. Scikit-learn library was used for crossvalidation with random seed set to 1 to avoid dataset variation between models. All models were trained from scratch with a fixed learning rate of 0.001, ADAM optimizer, and binary cross entropy loss function. Training was stopped when the training loss did not decrease for 15 epochs. In general, the models converged around 100 epochs.

4.4. Experimental results

The original version of Res-TSSDNet [46] (ResNet-1D in short), its sinc filter (ResNetSinc-1D) and gated versions with sinc filter were considered (GatResNetSinc-1D) as 1-D models. ResNetSinc-2D and GatResNetSinc-2D are their corresponding 2-D versions. Also, to verify the effectiveness of sinc filters, a gated model was used where the initial convolution layer had 70 filters to match the number of sinc filters (ResNet70-2D). To compare the proposed model's performance, DCA-ResNet18 [19] model was chosen. To make a fair comparison, the initial convolutional layer was changed to sinc filters in DCA-ResNet18 model too. All models were trained from scratch as explained previously. Figure 2 shows the results for the augmented SVD dataset. The results showed that sinc filters were more suitable for raw voice than the traditional convolution. ResNet70-2D performed worse than 1-D sinc filter models. On the other hand, 2-D sinc filter models performed better than 1-D sinc filter ones. The proposed gating mechanism boosted the performance of both 1-D and 2-D models. GatResNetSinc-2D model outperformed the others at all metrics. This outcome indicated that the gating/attention mechanisms were useful for voice disorder detection. Further, the proposed gating blocks have effectively boosted the performance of the ResNetSinc model.

F1 scores for each class are given in Figure 3 for the best performing models, i.e. ResNetSinc-2D, GatResNetSinc-2D, and DCA-ResNet18. An interesting observation is that the F1 scores for healthy class were less than the pathological ones. Furthermore, when Aug_{140} set was used, the performances rapidly decreased. This situation emphasizes the importance of a balanced dataset. A similar observation can be found in [11] for the selected classifiers. Also, other studies such as [42] reported superior performance with the balanced set compared to the imbalanced set for various classifiers. To summarize the results, the proposed gating approach achieved the best overall results with relatively low computational load increment. ResNetSinc-2D model had about 511 K learnable parameters, while GatResNetSinc-2D had 565 K parameters.

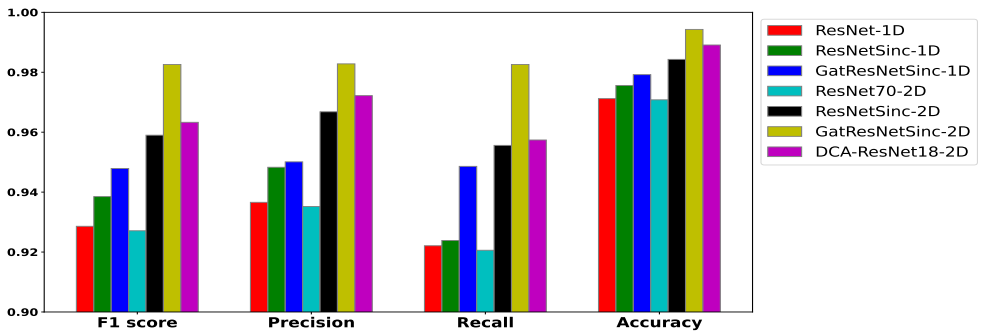

Figure 2. Performances of the deep models for the augmented dataset.

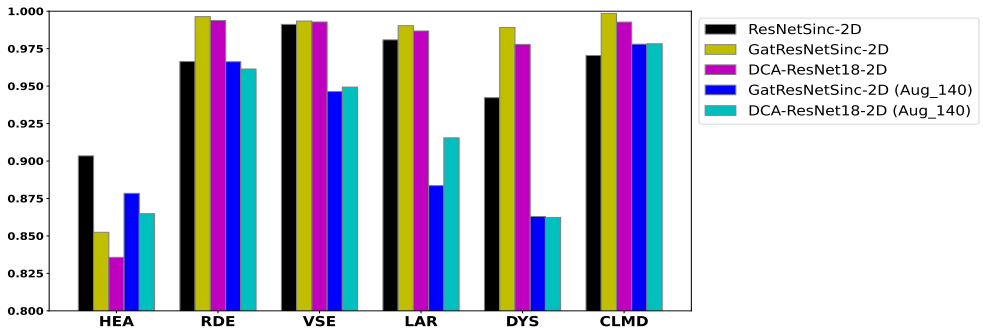

Figure 3. F1 scores for each class obtained via the best performing models.

4.5. Comparison with other systems

In this subsection, proposed model's performance is compared to other models from the recent studies. Number of healthy and pathological samples used in the respective studies are given in the table. However, train/validation and test distributions are not necessarily the same. Most of the studies given in the table have used /a/ wovel at neutral pitch and 71 pathologies. Using less number of pathologies can increase the accuracy, as seen for LSTM in [27]. The upper part of the table shows the multilabel classification results as discussed previuosly. The lower part of the table shows the results for binary classification. Note that in this case, data augmentation was not applied to the proposed model since all available data were used. A 10-fold validation was utilized and the average results are shown in the table for the proposed GatREsNetSinc-2D and for the DCA-ResNet18 models. As seen in the lower part of Table 3, the proposed approach achieved on par performance compared to several other models. SMMFNet and MFCC-SVM models had the hightest accuracies and F1 scores, where the others had accuracies varying between 0.681 and 0.696. The results showed that the proposed approach is effective for both multilabel classification and binary classification.

Table 3. Accuracy and F1 scores of the proposed model and other deep learning models. (ML: multilabel, MC: multiclass, BC: binary class.)

Model	Acc.	F1	Condition	Number of healthy/ pathological	Number of pathologies
GatResNetSinc-2D	0.994	0.982	ML – MC	687/4809	6
GatResNetSinc-2D (Aug_140)	0.962	0.919	ML – MC	687/980	6
DCA-ResNet18	0.989	0.963	ML – MC	687/4809	6
DCA-ResNet18 (Aug_140)	0.963	0.922	ML – MC	687/980	6
DNN (r = 95%) [11]	-	0.972	ML – MC	686/4480	6
DNN (r = 20%) [11]	-	0.916	ML – MC	687/939	6
GatResNetSinc-2D	0.685	0.683	BC	687/1356	71
DCA-ResNet18	0.674	0.673	BC	687/1356	71
SMMFNet [41]	0.781	0.844	BC	687/1354	71
2-D CNN [43]	0.696	0.69	BC	357/357	-
CNN – SVM [53]	0.690	-	BC	687/1354	71
MFCC – SVM [53]	0.764	-	BC	687/1354	71
LSTM [27]	0.975	0.981	BC	-	6
DNN [57]	0.681	-	BC	687/1356	71

A more detailed comparison with [11] is beneficial at this point. Albeit not given here, [11] also analyzed different problem transformation methods. The proposed GatResNetSinc-2D achieved a higher accuracy for the augmented data compared to all those problem transformation methods. Another important issue is the selected features. The proposed models operated directly on raw waveforms. On the other hand, signal energy [54], zero-crossing rate [55], and signal entropy [56] were considered as feature vectors in [11]. Hence, this study showed that deep models with raw waveform inputs can achieve a similar performance compared to the hand-crafted. The results observed with the DCA-ResNet18 model support this claim.

5. Discussion

The experimental results have provided important insights for the handled task. The most important conclusion may be the influence of the artificially created data. Due to the very limited amount of data, a meaningful result could not be derived without data augmentation. On the other hand, creating new samples from a limited number of original samples is not ideal. Experimental results observed in this study and in [11] verifies this situation. Although not a simple task, creating a more balanced multilabeled data should be prioritized by the researchers, which will inherently boost the number of research conducted on this topic.

Another important outcome of the experiments is the effectiveness of the proposed gating block. It was experimentally proven to be effective for both 1-D and 2-D models. For the chosen metrics, GatResNetSinc-2D achieved the highest values among the other ResNet variant models, including the DCA-ResNet18. Although in this study a relatively shallow network was used, the proposed gate block can be exploited by different CNN architectures. It can be also used in different research areas since it is not specific to voice disorder detection.

Majority of the automated voice pathology detection studies employed hand-crafted features. Experimental results showed that models operating on the raw voice files can compete with the hand-crafted features for multilabel multiclass classification. Superior performance of sinc filters compared to the traditional convolution filters was verified. 2-D models with the sinc filter delivered better performances than their 1-D counterparts. This situation may indicate the usefulness of time-frequency representations of the voice signals. Hence, when operating directly on raw waveform, transforming it into a 2-D representation after the initial layer can be a good starting point before applying any other layer. Future studies may focus on finding more optimal sinc filter parameters or designing a novel filter set in order to extract more discriminative information that suits voice disorder detection task. Due to the consistent performance of the 2-D models over the 1-D models, an attention mechanism that put more emphasis on the frequency information may be another topic worth investing.

6. Conclusion

This study proposed a deep learning architecture which consists of residual blocks and gating mechanisms. Gate blocks were placed between residual blocks to control the flow of information through the model. Hence, discriminative information captured in the earlier layers may be used by later layers in the forward propagation. The proposed model was applied to multilabel multiclass voice disorder detection task, which is a neglected task primarily due to the absence of a balanced dataset. However, a patient may suffer from more than one disorder and detecting them automatically from voice samples has enormous practical advantages. In order to achieve this, raw waveforms were used as inputs to deep models. Using sinc filters to obtain 2-D representations from the raw waveforms, slightly better results compared to a DNN model with handcrafted features were achieved. The proposed gate blocks were also found to be more efficient than deep connected attention. Although the experimental results were promising, a more balanced dataset is necessary to avoid effects of oversampling the less presented classes.

Conflict of interest

The author has no conflict of interest.

References

[1] Kim HB, Park YS, Lee JE, Han KD, Park YH. Study on relationship between self-recognition of voice disorder and mental health status: Korea National Health and Nutrition Examination Survey. Journal of Affective Disorders 2023; 338: 482-486. https://doi.org/10.1016/j.jad.2023.05.082

[2] Roy N, Merrill RM, Thibeault S, Parsa RA, Gray SD et al. Prevalence of voice disorders in teachers and the general population. Journal of Speech, Language, and Hearing Research 2004; 47 (2): 281-293. https://doi.org/10.1044/1092-4388(2004/023)

[3] Constantini AC, Ribeiro VV, Behlau M. Voice disorder classifications: a scoping review – Part A. Journal of Vioce (Article in Press) 2022. https://doi.org/10.1016/j.jvoice.2022.11.016

[4] Vuppala AK, Reddy RVS. Classification of voice pathology using different features and Bi-LSTM. In: 2023 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES); Tumaruku, India; 2023. pp. 1-4. https://doi.org/10.1109/ICSSES58299.2023.10200529

[5] Ribas D, Pastor MA, Miguel A, Martinez D, Ortega A et al. Automatic voice disorder detection using self-supervised representations. IEEE Access 2023; 11: 14915-14927. https://doi.org/10.1109/ACCESS.2023.3243986

[6] Hegde S, Shetty S, Rai S, Dodderi T. A survey on machine learning approaches for automatic detection of voice disorders. Journal of Voice 2019; 33 (6): 11-33. https://doi.org/10.1016/j.jvoice.2018.07.014

[7] Cummins N, Baird A, Schuller BW. Speech analysis for health: current state-of-the-art and the increasing impact of deep learning. Methods 2018; 151: 41-54. https://doi.org/10.1016/j.ymeth.2018.07.007

[8] Gupta R, Kumari S, Senapati A, Ambasta RK, Kumar P. New era of artificial intelligence and machine learning-based detection, diagnosis, and therapeutics in Parkinson's disease. Ageing Research Reviews 2023; 90. https://doi.org/10.1016/j.arr.2023.102013

[9] Abid Syed S, Rashid M, Hussain S. Meta-analysis of voice disorders databases and applied machine learning techniques. Mathematical Biosciences and Engineering 2020; 17 (6): 7958-7979. https://doi.org/10.3934/mbe.2020404

[10] Cesari U, De Pietro G, Marciano E, Niri C, Sannino G et al. A new database of healthy and pathological voices. Computers and Electrical Engineering 2018; 68: 310-321. https://doi.org/10.1016/j.compeleceng.2018.04.008

[11] Barbon S, Guido RC, Aguiar GJ, Santana EJ, Proença ML et al. Multiple voice disorders in the same individual: investigating handcrafted features, multi-label classification algorithms, and base-learners. Speech Communication 2023; 152: 1-14. https://doi.org/10.1016/j.specom.2023.102952

[12] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Las Vegas, NV, USA; 2016. pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[13] Baevski A, Zhou H, Mohamed A, Auli M. wav2vec 2.0: a framework for self-supervised learning of speech representations. In: 34th Conference on Neural Information Processing Systems (NeurIPS 2020); Vancouver, BC, Canada; 2020. pp. 1-12. https://doi.org/10.48550/arXiv.2006.11477

[14] Jung JW, Kim Y, Heo HS, Lee BJ, Kwon Y, Chung JS. Pushing the limits of raw waveform speaker recognition. In: Interspeech 2022; Incheon, Korea 2022. pp. 2228-2232. https://doi.org/10.21437/Interspeech.2022-126

[15] Narendra NP, Alku N. Glottal source information for pathological voice detection. IEEE Access 2020; 8: 67745-67755. https://doi.org/10.1109/ACCESS.2020.2986171

[16] Fujimura S, Kojima T, Okanoue Y, Shoji K, Inoue M et al. Classification of voice disorders using a one-dimensional convolutional neural network. Journal of Voice 2022; 36 (1): 15-20. https://doi.org/10.1016/j.jvoice.2020.02.009

[17] Harar P, Alonso-Hernandezy JB, Mekyska J, Galaz Z, Burget R et al. Voice pathology detection using deep learning: a preliminary study. In: International Conference and Workshop on Bioinspired Intelligence (IWOBI); Funchal, Portugal; 2017. pp. 1-4. https://doi.org/10.1109/IWOBI.2017.7985525

[18] Ma X, Guo J, Tang S, Qiao Z, Chen Q et al. Learning connected attentions for convolutional neural networks. In: IEEE International Conference on Multimedia and Expo (ICME); Shenzen, China; 2021. pp. 1-6. https://doi.org/10.1109/ICME51207.2021.9428397

[19] Ding H, Gu Z, Dai P, Zhou Z, Wang L et al. Deep connected attention (DCA) ResNet for robust voice pathology detection and classification. Biomedical Signal Processing and Control 2021; 70: 1-9. https://doi.org/10.1016/j.bspc.2021.102973

[20] Wu H, Soraghan J, Lowit A, Di Caterina G. Convolutional neural networks for pathological voice detection. In: 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); Honolulu, HI, USA 2018. pp. 1-4. https://doi.org/10.1109/EMBC.2018.8513222

[21] Wu H, Soraghan J, Lowit A, Di Caterina G. A deep learning method for pathological voice detection using convolutional deep belief networks. In: Interspeech 2018; Hyberabad, India 2018. pp. 446-450. https://doi.org/10.21437/Interspeech.2018-1351

[22] Syed SA, Rashid M, Hussain S, Zahid H. Comparative analysis of CNN and RNN for voice pathology detection. BioMed Research International 2021; 2021: 1-8. https://doi.org/10.1155/2021/6635964

[23] Verde L, De Pietro G, Sannino G. Voice disorder identification by using machine learning techniques. IEEE Access 2018; 6: 16246-16255. https://doi.org/10.1109/ACCESS.2018.2816338

[24] Verde L, De Pietro G, Alrashoud M, Ghoneim A, Al-Mutib KN et al. Leveraging artificial intelligence to improve voice disorder identification through the use of a reliable mobile app. IEEE Access 2019; 7: 124048-124054. https://doi.org/10.1109/ACCESS.2019.2938265

[25] Borovikova DV, Makukha VK, Shevchenko TA. Comparative analysis of acoustic parameters of the Saarbruecken Database's voice records. In: 19th International Conference of Young Specialists on Micro/Nanotechnologies and Electron Devices (EDM); Erlagol, Russia; 2018. pp. 6403-6406. https://doi.org/10.1109/EDM.2018.8435044

[26] Barche P, Gurugubelli K, Vuppala AK. Towards automatic assessment of voice disorders: a clinical approach. In: Interspeech 2020; Shanghai, China; 2020. pp. 2537-2541. https://doi.org/10.21437/Interspeech.2020-2160

[27] Bhattacharjee S, Xu W. VoiceLens: a multi-view multi-class disease classification model through daily-life speech data. Smart Health 2022; 23: 1-14. https://doi.org/10.1016/j.smhl.2021.100233

[28] Mohammed MA, Abdulkareem KH, Mostafa SA, Ghani MKA, Maashi MS et al. Voice pathology detection and classification using convolutional neural network model. Applied Sciences 2020; 10 (11): 1-13. https://doi.org/10.3390/app10113723

[29] Reid J, Parmar P, Lund T, Aalto DK, Jeffery CC. Development of a machine-learning based voice disorder screening tool. American Journal of Otolaryngology 2022; 43 (2): 1-5. https://doi.org/10.1016/j.amjoto.2021.103327

[30] Chen L, Chen J. Deep neural network for automatic classification of pathological voice signals. Journal of Voice 2022; 36 (2): 15-24. https://doi.org/10.1016/j.jvoice.2020.05.029

[31] Chaiani M, Selouani SA, Boudraa M, Yakoub MS. Voice disorder classification using speech enhancement and deep learning models. Biocybernetics and Biomedical Engineering 2022; 42 (2): 463-480. https://doi.org/10.1016/j.bbe.2022.03.002

[32] Han JY, Hsiao CJ, Zheng WZ, Weng KC, Ho GM et al. Enhancing the performance of pathological voice quality assessment system through the attention-mechanism based neural network. Journal of Voice (Article in Press) 2023. https://doi.org/10.1016/j.jvoice.2022.12.026

[33] Kim H, Park HY, Park DG, Im S, Lee S. Non-invasive way to diagnose dysphagia by training deep learning model with voice spectrograms. Biomedical Signal Processing and Control 2023; 86: 1-11. https://doi.org/10.1016/j.bspc.2023.105259

[34] Wu Y, Zhou C, Fan Z, Wu D, Zhang X et al. Investigation and evaluation of glottal flow waveform for voice pathology detection. IEEE Access 2021; 9: 30-44. https://doi.org/10.1109/ACCESS.2020.3046767

[35] Hung CH, Wang SS, Wang CT, Fang SH. Using SincNet for learning pathological voice disorders. Sensors 2022; 22 (17): 1-18. https://doi.org/10.3390/s22176634

[36] Ravanelli M, Bengio Y. Speaker recognition from raw waveform with SincNet. In: IEEE Spoken Language Technology Workshop (SLT); Athens, Greece; 2018. pp. 1021-1028. https://doi.org/10.1109/SLT.2018.8639585

[37] Islam R, Abdel-Raheem E, Tarique M. Voice pathology detection using convolutional neural networks with electroglottographic (EGG) and speech signals. Computer Methods and Programs in Biomedicine Update 2022; 2: 1-13. https://doi.org/10.1016/j.cmpbup.2022.100074

[38] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; Salt Lake City, UT, US; 2018. pp. 7132-714. https://doi.org/10.1109/CVPR.2018.00745

[39] Geng L, Liang Y, Shan H, Xiao Z, Wang W et al. Pathological voice detection and classification based on multimodal transmission network. Journal of Voice (Article in Press) 2022. https://doi.org/10.1016/j.jvoice.2022.11.018

[40] Muhammad G, Alhussein M. Convergence of artificial intelligence and Internet of Things in smart healthcare: a case study of voice pathology detection. IEEE Access 2021; 9: 89198-89209. https://doi.org/10.1109/ACCESS.2021.3090317

[41] Mohammed HMA, Omeroglu AN, Oral EA. MMHFNet: multi-modal and multi-layer hybrid fusion network for voice pathology detection. Expert Systems with Applications 2023; 223: 1-13. https://doi.org/10.1016/j.eswa.2023.119790

[42] Fan Z, Wu Y, Zhou C, Zhang X, Tao Z. Class-imbalanced voice pathology detection and classification using fuzzy cluster oversampling method. Applied Sciences 2021; 11 (8): 1-21. https://doi.org/10.3390/app11083450

[43] Javanmardi F, Kadiri SR, Alku P. A comparison of data augmentation methods in voice pathology detection. Computer Speech and Language 2024; 83: 101552. https://doi.org/10.1016/j.csl.2023.101552

[44] Huckvale M, Buciuleac C. Automated detection of voice disorder in the Saarbrücken Voice Database: effects of pathology subset and audio materials. In: Interspeech 2021; Brno, Czechia; 2021. pp. 1399-1403. https://doi.org/10.21437/Interspeech.2021-1507

[45] Chui KT, Lytras MD, Vasant P. Combined generative adversarial network and fuzzy c-means clustering for multi-class voice disorder detection with an imbalanced dataset. Applied Sciences 2020; 10 (13): 4571. https://doi.org/10.3390/app10134571

[46] Hua G, Teoh ABJ, Zhang H. Towards end-to-end synthetic speech detection. IEEE Signal Processing Letters 2021; 28: 1265-1269. https://doi.org/10.1109/LSP.2021.3089437

[47] Ge W, Patino J, Todisco M, Evans N. Raw differentiable architecture search for speech deepfake and spoofing detection. In: 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge; Brno, Czechia; 2021. pp. 22-28. https://doi.org/10.21437/ASVSPOOF.2021-4

[48] Tak H, Jung JW, Patino J, Kamble M, Todisco M et al. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. In: 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge; Brno, Czechia; 2021. pp. 1-8. https://doi.org/10.21437/ASVSPOOF.2021-1

[49] Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional Block Attention Module. In: 15th European Conference on Computer Vision; Munich, Germany; 2018. pp. 3-19.

[50] Wang Q, Wu B, Zhu P, Li P, Zuo W et al. ECA-Net: efficient channel attention for deep convolutional neural networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Seattle, WA, USA; 2020. pp. 11531-11539. https://doi.org/10.1109/CVPR42600.2020.01155

[51] Li X, Wu X, Lu H, Liu X, Meng H. Channel-wise gated Res2Net: towards robust detection of synthetic speech attacks. In: Interspeech 2021; Brno, Czechia; 2021. pp. 4314-4318. https://doi.org/10.21437/Interspeech.2021-2125

[52] Fan C, Xue J, Tao J, Yi J, Wang C et al. Spatial reconstructed local attention Res2Net with F0 subband for fake speech detection. arXiv preprint 2023; 2308.09944v1 [cs.SD]. http://arxiv.org/abs/2308.09944

[53] Omeroglu AN, Mohammed HMA, Oral EA. Multi-modal voice pathology detection architecture based on deep and handcrafted feature fusion. Engineering Science and Technology, an International Journal 2022; 36: 1-11. https://doi.org/10.1016/j.jestch.2022.101148

[54] Guido RC. A tutorial on signal energy and its applications. Neurocomputing 2016; 179: 264-282. https://doi.org/10.1016/j.neucom.2015.12.012

[55] Guido RC. ZCR-aided neurocomputing: a study with applications. Knowledge-Based Systems 2016; 105: 248-269. https://doi.org/10.1016/j.knosys.2016.05.011

[56] Guido RC. A tutorial review on entropy-based handcrafted feature extraction for information fusion. Information Fusion 2018; 41: 161-175. https://doi.org/10.1016/j.inffus.2017.09.006

[57] Harar P, Alonso-Hernandezy JB, Mekyska J, Galaz Z, Burget R et al. Voice pathology detection using deep learning: a preliminary study. In: International Conference and Workshop on Bioinspired Intelligence; Funchal, Portugal; 2017. pp. 1-4. https://doi.org/10.1109/IWOBI.2017.7985525