IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 July 2023, accepted 22 July 2023, date of publication 1 August 2023, date of current version 4 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3300795

RESEARCH ARTICLE

A Modular Deep Learning Architecture for Voice Pathology Classification

IOANNA MILIARESI AND AGGELOS PIKRAKIS, (Member, IEEE)
Department of Informatics, University of Piraeus, 185 34 Pireas, Greece
Corresponding author: Ioanna Miliaresi (imiliaresi@unipi.gr)

ABSTRACT The development of methods that combine different sources of information for medical diagnosis is an essential challenge in the field of medical informatics. In this context, we introduce a machine-learning framework for automatic voice pathology classification and, in particular, a modular deep learning architecture that classifies voice signals stemming from four types of voice disorders. To this end, we design a multimodal deep learning architecture that fuses medical metadata with voice signals. Our classifier is a combination of fully convolutional and feed-forward sub-networks that simultaneously process low-level and mid-level features which are extracted from acoustic signals of varying duration and medical records, respectively. A key objective of our study is to develop an architecture that is capable of processing voice samples of varying duration, to enhance the system's learning and inference capabilities. Our research also focuses on overcoming performance limitations of neural networks that stem from the lack of extensive volumes of training data. We therefore, investigate problem-specific augmentation techniques based on the feature sequence segmentation and coloured noise injection and we show that the proposed method gives state-of-the-art results, achieving 64.4% classification accuracy, compared to the 63.5% classification score of the best performing method of the 2019 FEMH data challenge.

INDEX TERMS Voice pathology classification, deep multimodal neural networks, fully convolutional networks, data augmentation.

I. INTRODUCTION

The traditional approach to diagnosing voice pathology has centered on assessing laryngeal structure and mobility and examining respiratory dynamics. Laryngoscopy is the most effective technique for observing and accurately assessing the laryngeal structure, including the mobility of its tissues. Respiratory dynamics, including lung volume, airflow, pressure, and breathing patterns, are measured using spirometry and pneumotachography techniques. To reduce reliance on specialized medical equipment, implementing procedures for evaluating phonatory and respiratory dynamics through automatic speech signal analysis has proven to be a successful, cost-effective, and non-invasive alternative.

Voice pathology classification refers to the task of machine-driven decision-making of the type of pathology present in a voice recording given a set of predefined pathology classes. This classification task is traditionally approached from perspectives of pattern recognition and statistical learning and, recently, using deep learning techniques. Developing a robust voice pathology classification system based on machine learning techniques has been a prominent research assignment with a strong potential impact on public health. In recent decades increasing research has been made, primarily aiming at developing accurate feature extraction techniques with the appropriate acoustic parameters and applying classification algorithms that achieve high classification precision. To that end, collections of voice recordings of healthy subjects and patients with different types of hyper - and hypofunctional vocal fold pathologies have been gathered and arranged into databases. The recordings contain the sustained vowels /a/, /i/, /e/, and/or continuous speech. As stated in [1], the most popular voice pathology databases are the Massachusetts Eye and Ear Infirmary (MEEI) database, the Saarbruecken Voice Database (SVD) and the Arabic voice pathology database (AVPD), in

The associate editor coordinating the review of this manuscript and approving it for publication was Chuan Li.

descending order. Lately, during the COVID-19 pandemic, several COVID-19 related datasets have been assembled, including the MIT COVID-19 dataset, the University of Cambridge COVID-19 Sounds dataset, the University of Stanford Virufy dataset and the EPFL COUGHVID dataset, which mainly focus on COVID-19 cough, breathing, and speech sounds.

The development of automatic voice disorder classification systems has led researchers to experiment with different pathology types across multiple classification tasks based on available voice disorder data collections. Among these tasks, the binary classification problem of distinguishing healthy voice samples from unhealthy voice samples has been a primary focus of the investigation. According to [2], the healthy and unhealthy sustained vowels of the MEEI database turned out to be perfectly separable. This observation raises the question of whether the reported methods can generalize on unseen data, consequently increasing the need to conduct experiments on new databases.

In this line of thinking, we undertake experiments with a more challenging database of voice disorders, the Far Eastern Memorial Hospital (FEMH) database. During the 2019 ''IEEE BigData Cup,'' this database was introduced as part of the FEMH voice data Challenge and focuses on a 4-class classification problem involving voice recordings from four different categories of voice disorders: *functional dysphonia*, *phonotrauma*, *laryngeal neoplasm* and *vocal paralysis*. The availability of medical information for the voice recordings of the FEMH corpus aligns well with our research objective of integrating medical information into the classification system.

The FEMH data corpus, as it is also the case with most voice disorder databases, comprises a restricted amount of data, thus posing a significant limitation when training deep learning architectures. To the best of our best knowledge, the literature has not extensively addressed data augmentation techniques as a remedy for the lack of sufficient training data. To resolve this research gap, we investigate various data augmentation methods tailored to the voice pathology classification problem.

The observation that the voice recordings have varying duration in the dataset is another essential data-related characteristic in the context of voice disorder classification since the duration of sustained vowels correlates with the phonation capabilities of the patients. For example, vocal fold damage (as in the case of the neoplasm disorder) results in altered Maximum Phonation Time (MPT) values, and patients with incomplete glottis closure (a symptom of vocal paralysis) present lower MPT values- because air leakage is higher than in the case of healthy people [3]. Short phonation times in the case of vocal paralysis also indicate a patient's inability to pronounce sustained vowels for a long time. In addition, loudness fluctuations and silent parts in the recordings indicate patients' difficulties during vowel pronunciation.

It is worth noting that relevant research work mainly focuses on network architectures and methods that require pre-processing and resizing audio recordings to segments of predefined length, often using zero-padding techniques. As the duration of recordings correlates with the presence of pathology, segmentation to fixed length segments can result to information loss and subsequent inferior classification performance. In order to overcome the limitations of conventional classifiers that require fixed duration input, we propose a ''fully-convolutional'' architecture which is capable of processing 2-*D* representations of audio recordings of arbitrary duration. We also define a data augmentation method which splits the audio recordings into variable-length segments.

Overall, we propose a novel classifier that simultaneously processes data from two modalities (bimodal classifier), namely audio signals and medical records, based on a fully convolutional architecture [4] that analyzes voice recordings as images of varying width, thus overcoming the need for assumptions regarding recording duration. Furthermore, we enhance the training stage with custom augmentation techniques that produce variable-length training data injected by different noise types.

The rest of the paper is structured as follows: Section II presents related work and Section III describes the collection of audio recordings and medical data of the FEMH corpus and introduces the four pathologies of our study. Section IV presents the proposed method, including the feature extraction stage and the model architecture. Section V describes our experimental setup, training procedures, adopted augmentation techniques, classification results, and network performance interpretation by means of visualization techniques. Finally, Sections VI and VII discuss and summarize our research findings.

II. RELATED WORK

During the past few years, several research attempts have shown that automatic voice pathology classification systems can provide solutions to a spectrum of tasks related to voice impairment assessment by means of various feature extraction schemes and machine learning methods.

Recently, Support vector machine (SVM) classifiers [5], [6], Naïve Bayes solutions, decision trees, and ensemble classifiers [7] were employed for the detection of vowel pathology in the SVD dataset. Furthermore, [8], [9], [10], [11], [12], [13] investigated the capacity of deep neural networks to classify voice pathology in the SVD dataset.

As the four-class classification problem which we are dealing with is relatively new, we also include results reported in the literature for related voice pathology classification problems, starting with a feature extraction stage survey. More specifically, the work in [14] has shown that Mel-frequency-coefficients (MFCCs) combined with pitch frequency measurements yields a 99.44% classification accuracy for the binary problem of discriminating normal from pathological

speech for the case of the sustained vowel /a/. Furthermore, in [15], perturbation methods (including jitter and shimmer), in combination with signal-to-noise ratio, and nonlinear dynamic methods (correlation dimension and second-order entropy) were tested for the analysis of sustained and continuous vowels with laryngeal pathologies. In paper [16], complexity measures of noise parameters and MFCC coefficients were analyzed, while in [17], the wavelet packet transform was employed to analyze dysphonic voices. For the special case of dysphonia detection in recordings of the sustained vowel /a/, biologically inspired AM Analysis features [18], modulation spectral features combined with MFCCs [19] and modulation spectral features [20] have shown to exhibit good performance. Recently, bio-inspired algorithms with innovative graphical representations of audio signals and heuristic methods were introduced in [21]. In [22], features extracted from pitch contours, MFCCs, gramatone cepstral coefficients (GTCC), Gabor wavelets and auditory spectrograms, were processed to address the problem of impaired voice classification. Furthermore, research reports have stated that mel-frequency coefficients and their derivatives can serve as discriminating features for voice pathology types [12], [23], [24]. For the more specific case of laryngeal carcinomas, MFCCs were proposed as a feature [25].

Relevant studies can be found in the list of submitted methods at the FEMH 2018 challenge which introduced a three-class pathology detection problem (neoplasm, phonotrauma and vocal palsy). It can be observed that MFCCs, delta MFCCs and Mel-scaled spectrograms were proposed by most of the participants as the most discriminative features for detection and classification tasks [26], [27], [28], [29], [30], [31], [32], [33]. The work in [34] has shown that the pertrubation features of Normalized Noise Entropy, Cepstral Harmonics-to-Noise Ratio, Glottal-to-Noise Excitation Ratio, Smoothed Cepstral Peak Prominence and Low-to-High Frequency Spectral Energy Ratio can serve to measure the presence of noise resulting from incomplete glottal closure of the vocal folds, as well as the presence of modulation noise due to irregularities in vocal fold movement.

Interesting results have also been reported on the integration of electroglottographic signals (EGG) as a source of supplementary information and related studies have used the Saarbrücken Voice Database. Specifically, in [35] and [36], convolutional neural networks are used to classify healthy and pathological voice signals and the integration of EGG signals has shown to increase classification accuracy. Similarly, in [37], a pre-trained convolutional neural network in combination with a Long short-term memory network analyzes EGG data to obtain a better classification result. The inclusion of EGG data via a convolutional neural network has also increased classification performance in [38] and [39], a two-level classifier based on a combined CNN-RNN architecture has given good results for the problem of detecting voice pathology. Recently, in [40], spectrograms of the EGG signals are used for detecting the presence of post-COVID-19 disease.

From a classification perspective, the detection of functional *dysphonia*, *phonotrauma*, *laryngeal neoplasm* and *vocal paralysis* is a 4-class problem that was introduced by the 2019 ''IEEE BigData Cup'' challenge, which only reported performance results of the participating methods without disclosing algorithmic details. The authors of this paper also submitted a method that introduced a deep learning architecture that fused mid-term fixed-length segments of acoustic features and medical descriptors into convolutional and feed forward neural networks [41]. Based on published results, the method achieved a 57% classification accuracy on the test set of the challenge, with the best-performing method reaching a 63.5% classification score.

The competition was an extension of the 2018 challenge, where the objective was the simpler task of distinguishing recordings of healthy subjects from pathological cases belonging to three voice pathologies (neoplasm, phonotrauma, and vocal palsy). Based on the respective technical reports, the methods of the 2018 challenge showed a preference for mainstream machine learning approaches. More specifically, several approaches employed SVMs [26], [27], [30], [32], [42], [43], Gaussian mixture models [32], [44], Bayesian networks and Random forest classifiers [28] while neural network architectures were less popular [28], [32], [33]. For the classification task of healthy against pathological recordings, the best-performing method with respect to classification accuracy was [27], which was based on Gaussian mixture models to achieve a classification accuracy of 96.9%. In the FEMH 2018 three-class voice pathology classification problem the unweighted average recall was employed as a performance metric. The best result (60.67%) was achieved by the method presented in [30] that used Gaussian mixture models. The incorporation of demographic data was initially investigated in [32], [33], and [42]. Compared to our approach, which is an end-to-end one without intermediate, hand-crafted features, the method in [33] uses deep neural networks to combine acoustic and medical data but only considers conventional feed-forward architectures, thus enforcing the use of Gaussian Mixture Model (GMM) as a pre-processing step to represent an audio recording statistically before feeding it to the neural network architecture.

Finally, for the sake of completeness, we provide references to the neighbouring task of Covid-19 infection detection, which was treated by voice pathology classifiers operating on speech recordings. Specifically, SVMs were used to detect the presence of COVID-19 cough in [45] and [46], along with linear regression models in [47]. Convolutional neural networks (CNN) were employed in [48] and [49] and bi-directional long short-term memory (BiLSTM) networks were used in [50] and [51]. Furthermore, several studies leveraged transfer learning techniques, as in [52], where three pre-trained ResNet50 networks were employed to process cough recordings. Similarly, in [53]

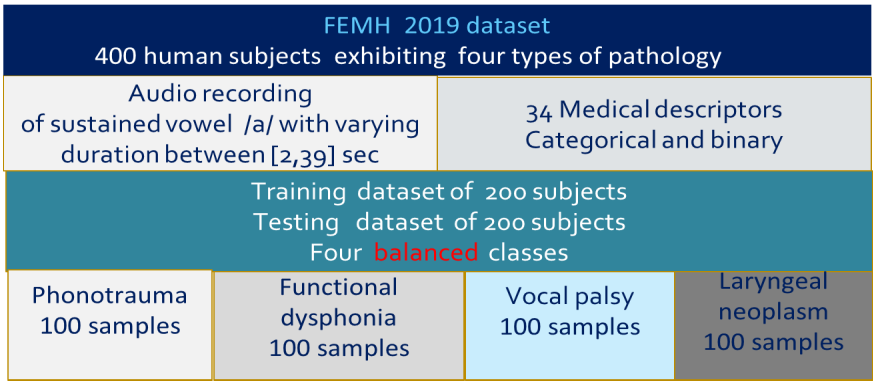


FIGURE 1. Four balanced classes training-testing datasets.

and [54], pre-trained deep neural networks involving CNN, LSTM, and Resnet50 architectures were adopted to process recordings of coughing, breathing, and speech sounds.

All the previously mentioned studies propose a wealth of methods that experiment with feature extraction mechanisms and network architectures, but they do not investigate database augmentation techniques. As the performance of deep learning models depends strongly on the quantity of training data, data augmentation is crucial for developing a robust voice pathology classifier. Our work, therefore, aims to overcome this limitation by developing augmentation techniques tailored to the task of voice pathology detection.

III. DATASET DESCRIPTION

The 2019 FEMH dataset consists of voice signals and medical records of patients with four types of voice disorders, namely hyperfunctional dysphonia, phonotrauma, vocal palsy, and neoplasm but there are no recordings of healthy speech. The voice recordings and the related medical records were obtained from a voice clinic in the Far Eastern Memorial Hospital (FEMH). The dataset formed the basis of the 2019-FEMH Challenge of the 2019 IEEE Big Data Cup. Each medical record contains 34 demographic questions, some of which are categorical and others binary, including age, gender, job, habits, and symptoms at the time of voice quality degradation, how it happened, whether internal surgery took place or not, how severe the gastroesophageal reflux was and so on. The training dataset consists of fifty voice recordings per disease, in which the sustained vowel /a/ is pronounced by pathological speakers of different ages and sex. Notable characteristics of the dataset are the varying recording duration in the range of [2-39] seconds and the fact that different sampling frequencies were employed during the recording procedure. The challenge rules defined that classification accuracy would be the metric for ranking submissions, computed over a testing set of 200 recordings covering the four types of pathology. Fig. 1 summarizes most of the FEMH 2019 dataset characteristics.

The aforementioned four types of pathology can be described in brief from a medical perspective as follows:

- *Hyperfunctional dysphonia* is an excessive involuntary muscle contraction, as a consequence of improper phonation [55]. It results from the overuse of the laryngeal muscles and occasionally, the use of the false vocal folds (the upper two vocal folds that are not involved in vocalization). Its typical symptoms include breathy, hoarse, or rough voice, voice instability, and voice fatigue, which are common to many voice disorders.
- *Phonotrauma*, is defined as "trauma to the laryngeal mechanism (vocal folds) as the result of vocal behavior that includes yelling, screaming and throat-clearing" [56]. It refers to the formation of common vocal-fold lesions (e.g. vocal fold nodules) that affect how the folds vibrate, and its symptoms are similar to dysphonia.
- *Vocal fold paresis/paralysis (palsy)* refers to the situation where one (unilateral) or both (bilateral) vocal folds are paralyzed [57]. The airway and breathing are thus severely compromised, and this results in voice changes like hoarseness, breathy voice, extra effort while speaking, need for excessive air pressure during the usual conversational voice style and diplophonia (voice sounds like a gargle).
- Finally, the term *neoplasm*, refers to various types of cancer, including laryngeal, voice box, or vocal cords tumors. Common symptoms are hoarseness, painful swallowing, and fatigue. [58].

IV. METHOD DESCRIPTION

A. FEATURE EXTRACTION

To extract the input feature vector, we follow the approach presented in the baseline implementation of our classifier [41]. Specifically, we design input vectors from audio recordings and medical data. The audio feature vector is defined to capture the spectral shape of the speech signal and its evolution over short periods. It includes the first 13 MFCCs coefficients, augmented with their first-order derivatives and the logarithm of the mel-filterbank outputs. This input modality is fed to a fully convolutional network branch. The perturbation-medical feature vector fuses the mid-term

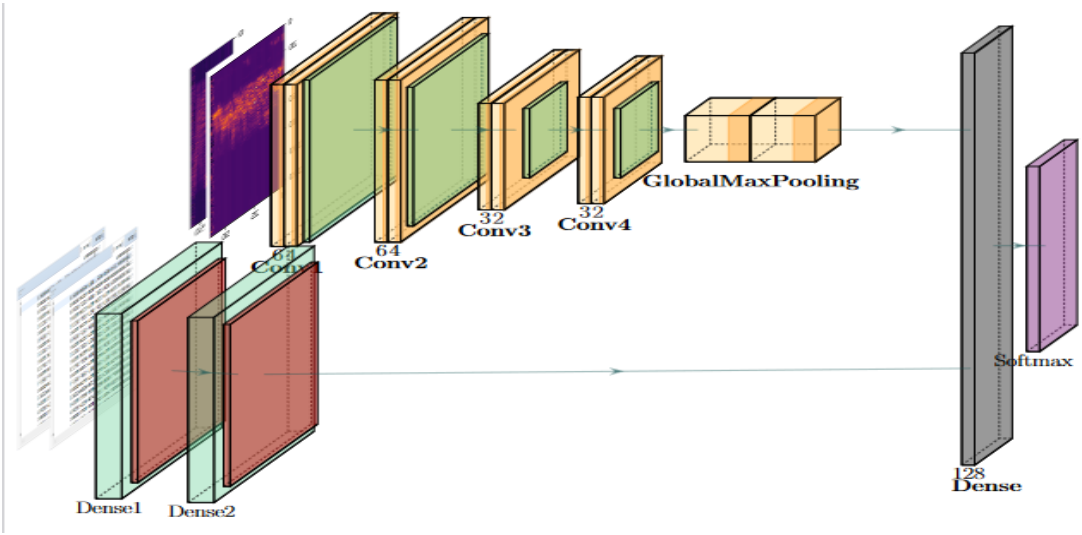


FIGURE 2. Overview of the structure of proposed network architecture. The audio features branch implements a fully convolutional network with four convolutional layers (Conv1, Conv2, Conv3, Conv4) followed by a GlobalMaxPooling layer. The perturbation and medical features processing branch consists of two fully connected layers (Dense1, Dense2). The pooling output of the GlobalMaxpooling layer and the output of the ReLu of the second fully connected layer (Dense2) merge into the final fully connected layer (Dense). The final branch uses a softmax layer to output classification results.

features of the fundamental frequency, jitter, and harmonic-to-noise ratio with the metadata descriptors from the medical records. The second input vector is fed into a fully connected network branch.

In more detail, each audio recording is first resampled to 44.100Hz and its amplitude is normalized in the interval $[-1, +1]$. Then, the signal is parsed with a 40ms long moving window with a hop size of 20ms. During each frame, the Discrete Fourier Transform (DFT) is computed, and the resulting DFT coefficients are then fed as input to a mel-filter bank. Each mel-filter in the bank performs a weighted sum of the magnitudes of the DFT coefficients that lie within its specific frequency range. After obtaining the filterbank output, the logarithm of each output is calculated, and subsequently, the discrete cosine transform of the logarithms is computed.

The literature suggests that only the first 13 MFCCs contain usable information and that the first coefficient must be discarded. Against common practice, we follow a different approach and include the first MFCC in our feature vector. The first MFCC corresponds to the mean value of signal intensity. We experimentally prove that the signal intensity and its fluctuations correlate with certain types of pathology.

In addition, to capture the dynamics of the signal, we compute the first-order derivative of the MFCCs vector over time, and append it to the vector of the MFCCs. Then we augment the feature vector with the logarithm of 26 mel-filterbank outputs, thus resulting in a total of 52 feature values per frame and a varying-length sequence of 64 feature vectors per audio recording. The resulting feature sequence is formed as a 2-D image representation with dimensionality $N \times 64$, where N ranges in [124, 1462], defined by the samples' duration, and gets processed by the convolutional module.

To compose the second *perturbation-medical input vector*, we stack the 34 medical measurements described in the database metadata and the 3 mid-term segment features, namely fundamental frequency, jitter, and harmonic to noise ratio measurements.

The mean fundamental frequency (F0) is computed according to the probabilistic YIN (pYIN) algorithm [59]. Since F0 does not yield a sufficiently detailed representation of vocal fold vibratory patterns, we include a more sensitive acoustic measurement of vocal function, jitter. Jitter measures the cycle-to-cycle variations of the fundamental frequency and it is computed as the average absolute difference between consecutive periods. The Harmonic-to-Noise Ratio (HNR) reflects cycle-to-cycle variability both in frequency and amplitude, as well as additive noise generated at the glottis. It serves as a descriptor of the breathiness and hoarseness of a voice. HNR is defined as the ratio of the energy of a periodic signal to the energy of the noise in the signal expressed in decibels. This ratio is estimated and included in the feature vector.

The medical records contain features related to gender and age, along with patients' answers regarding the symptoms they experience, like fatigue, breathiness, etc. The corresponding categories are incorporated into the perturbation-medical feature vector, preserving their numerical values as numbers. As a result, for each human subject, a 37×1 dimensional data vector is formed, with each element normalized within the interval of $[-1, +1]$. Finally, the resulting vector is processed through the feed-forward input branch for further processing.

B. SYSTEM ARCHITECTURE

We propose a modular deep learning architecture that processes data stemming from two modalities and performs the tasks of feature representation learning, information fusion, and class prediction. This classifier processes two information sources via two sub-networks, the outputs of which are concatenated and fed into a learning module that produces the final classification decision. Fig. 2 presents the architecture of our model.

The framework contains two main parallel branches that process the audio feature vectors and the perturbation-medical data respectively. The two branches merge into the last module, where the final prediction is made. In more detail:

- The first branch processes feature sequences extracted from the audio recordings. It is built by adopting the principles of a fully convolutional neural network [4]. This type of neural network architecture operates on the input of arbitrary dimensions and produces an output vector of fixed dimensionality.
  Standard convolutional neural network classifiers usually repeat a pattern of a convolutional layer followed by an average pooling layer with a fully connected layer at the end of the processing pipeline. The feature maps are then flattened and fed to a cascade of fully connected layers to yield the final prediction. The last fully connected layer has a fixed number of inputs, which enforces the requirement for input images of fixed size. To overcome this constraint, ''fully convolutional networks'' are adopted. To this end, the final dense layers of a hybrid convolutional-dense classifier are omitted and replaced by a block of a convolutional layer with kernel size 1×1, and stride 1, followed by a global max pooling layer. The output of this block has constant dimensionality, defined by the number of filters, n, i.e., $1 \times 1 \times n$, independently of the size of the input image. In our implementation, we transform each audio recording into a one-channel, two-dimensional ''image'' of size $h \times w$, where h and w are spatial dimensions. The image dimensions depend on the duration of each recording. As described, the feature vector is an $N \times 64$ matrix, where N depends on sample duration and lies in the range $[124 - 1462]$. The resulting feature vector is fed to the first fully convolutional network branch. As a consequence, the input shape of the first convolutional layer does not have fixed dimensions. As we adopted a batch size equal to one, the batch shape is eventually $1 \times N \times 64 \times 1$.
  The respective branch consists of four consecutive convolutional - max pooling - batch normalization blocks and a final global max pooling layer. The first three convolutional layers contain 64, 64, and 32 convolutional masks respectively, with ReLu activation functions and each one has a kernel of size 3×3 with stride equal to 1 (without zero padding). The last convolutional layer performs the 1×1 convolution through 32 masks with stride equal to 1 (again without zero padding). The last global max pooling layer subsamples the output.
- The second branch, designed to process the remaining features, is a feed forward neural network with two hidden layers consisting of 128 and 128 units, respectively, with Rectified linear unit activation functions. It processes each 37×1 input vector of medical metadata and perturbation features, as described in the subsection feature extraction.
- Subsequently, the outputs of the aforementioned ''sub-networks'' are concatenated and fed to a dense layer with 128 neurons and a ReLu activation function.
  The architecture of the network with a detailed description of the dimensionality of the layers and hyper-parameters values are listed in Table 1.

V. EXPERIMENTS AND RESULTS

The proposed architecture is formulated trough extensive experimentation over possible model configurations and data augmentation techniques. Specifically, we experiment with early-stage, intermediate, and late-stage fusion strategies. The results demonstrate that the network's performance improves when the intermediate features learned from different modalities are concatenated to inform the classification decision.

A. TRAINING

An ablation study analyzed the impact of the growth in size and complexity of the network architecture. We conducted an extensive set of experiments in which various components of the network architecture were removed or replaced to assess their impact on the system's performance. We validated the network's complexity by experimenting with the number of layers of the two sub-networks and the number of filters of the convolutional layers.

Reducing the number of convolutional layers to 3 resulted in a decrease in classification accuracy. However, when we experimented with a configuration using three convolutional layers and the number of filters set to 64, 64, 64, classification accuracy increased to 58.5%, with the number of filters specified to 64, 64, 128 equals 60.5%, while setting the number of filters to 128, 128, 64 leads to a value of 60.6%. Adjusting the number of filters to 64, 64, 128 yielded a higher accuracy of 60.5%, while setting the number of filters to 128, 128, 64 led to a slight improvement at 60.6%. The highest accuracy of 62.99% was achieved when using 128, 128, 128 filters.

On the other hand, increasing the number of layers and filters led to a drop in classification accuracy. With four convolutional layers and filter sizes set to 64, 64, 128, 128, the accuracy reduced to 60%. Similarly, using 64, 64, 64, 64 filters resulted in a classification accuracy of 63.1%. Adding an extra layer with 64, 64, 64, 64, 64, filters further decreased the accuracy to 58%.

Additionally, we conducted experiments to explore different sizes for the fully connected layers. An additional layer with 512 nodes significantly reduced the accuracy

TABLE 1. Network configuration description. The output of every layer is fed as input to the next layer. ReLu and dropout layers always follow convolutional and fully connected layers. Parameter N lies in the range of [124 – 1462] depending on the audio recording duration.

Layer	Output shape	Filters	Kernel, stride
Input	$(1 \times N \times 64 \times 1)$	-	- , -
Conv1	$(1 \times (N-2) \times 62 \times 64)$	64	3×3 , 1×1
Max pooling	$(1 \times ((N-2)/2) \times 31 \times 64)$	-	2
Conv2	$(1 \times (((N-2)/2)-2) \times 29 \times 64)$	64	3×3 , 1×1
Maxpooling	$(1 \times (((((N-2)/2)-2)/2) \times 14 \times 64)$	-	2
Conv3	$(1 \times ((((N-2)/2)-2)/2)-2) \times 12 \times 32)$	32	3×3 , 1×1
Maxpooling	$(1 \times ((((((N-2)/2)-2)/2)-2)/2) \times 6 \times 32)$	-	2
Conv4	$(1 \times (((((((N-2)/2)-2)/2)-2)/2)-2) \times 4 \times 32)$	32	1×1 , 1×1
GlobalMaxpooling	$(1 \times 1 \times 32)$	-	-
Dense1	(1×64)	-	-
Dense2	(1×64)	-	-
Dense	(1×128)	-	-
Classifier	(1×4)	-	-

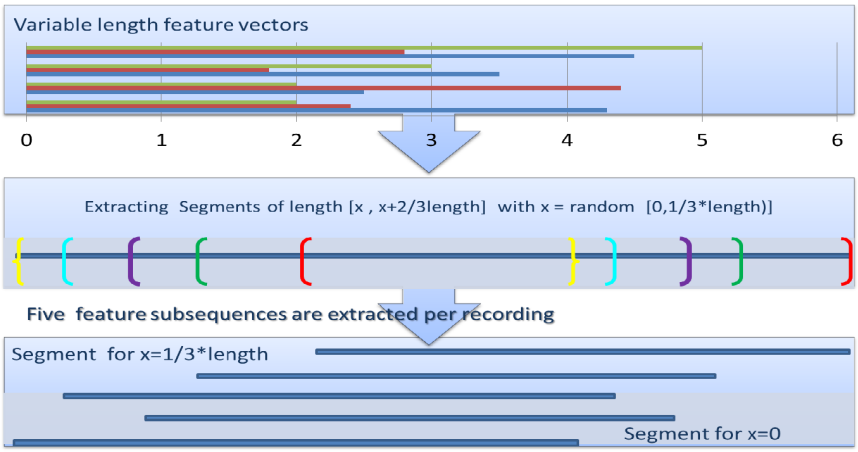


FIGURE 3. Feature sequence segmentation.

to 51%. Moreover, when we attempted to decrease the number of neurons in the fully connected layer from 1024 to 512, the performance dropped, resulting in an accuracy of 49%.

Through our hyper-parameter optimization, we observed that when the Stochastic Gradient Descent optimizer is assigned, the performance of the classifier decreases, with testing accuracy reported to be 63%. Our ablation studies also involved the investigation of alternative loss functions; when the sigmoid was tested, classification accuracy dropped to 53.5%, and with tanh a further decrease was observed down to 51.5%. At the final configuration, Rectified linear unit was used. To validate the importance of learning rate we trained the model with three alternative learning rates, 0.1, 0.01, and 0.001 with testing accuracy scores of 48%, 52%, and 59% respectively.

All network configurations are trained using a 4-fold cross-validation scheme. In other words, at each run, three folds are used for training and one for validation (for a total of four runs), and the final classifier accuracy is computed over a separate testing set. The FEMH training dataset contains 200 audio recordings, while the testing dataset contains 200 audio recordings. Since we adopted a 4-fold cross-validation training scheme, 150 samples are used for training and 50 for validation. Augmentation methods are only applied on the training subset and produce a balanced set of 750 audio clips with 175 samples per pathology class. The models are trained for a maximum of 300 epochs using the Adam gradient descent algorithm with a learning rate of 0.0001 while observing the validation error. An early-stopping criterion of 5 epochs is used to restrict training times. The categorical cross-entropy loss is used to compute the error signal, and validation accuracy is used as an auxiliary metric. To prevent over-fitting, a dropout regularization scheme is adopted with the dropout value set to 0.5 for the convolutional and dense layers. From an implementation perspective, to address the training requirements of the fully convolutional branch, a data generator is used to create batches of one image (recording) at a time.

In addition, as it is common practice, we include normalization layers to reduce the so-called internal covariance shift, defined as changes in the distribution of network activations due to changes in network parameters. To evaluate the performance of batch normalization in such a setup and investigate

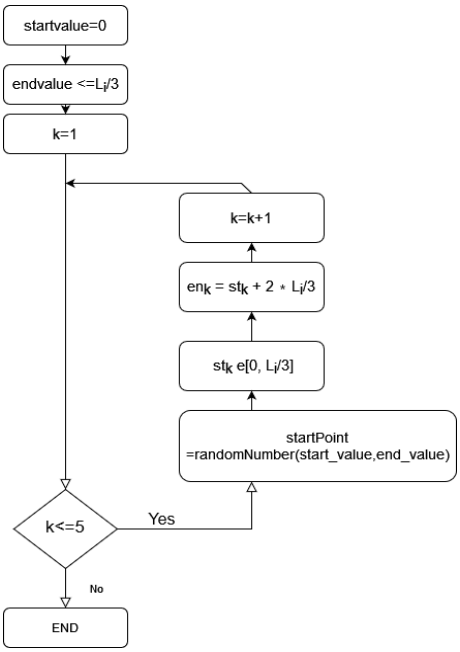


FIGURE 4. Segmentation flowchart.

the effect of mini-batch size, we performed a grid search on normalization techniques.

We experiment with four alternative normalization approaches: batch normalization [60], layer normalization [61], weight normalization [62] and instance normalization [63]. We also evaluate the contribution of non-linear activation functions (hyperbolic tangent, sigmoid, and ReLu). At first, we construct a network architecture without any normalization layers and with sigmoid activation functions. This network configuration gives a testing accuracy of 49%. An alternative configuration without normalization layers and tanh as the activation function yields an improved testing accuracy of 58.9%. The use of a weight normalization layer decreased classification accuracy to 57%. Instance and batch normalization achieved an almost identical testing accuracy of 63%. As a final system configuration, we adopt a scheme with batch normalization and Rectified linear units.

B. AUGMENTATION METHODS

1) SEGMENTATION

The FEMH database contains a small amount of training data, which is an important limitation when training deep learning architectures. More specifically, the training set consists of 200 recordings for the four types of pathology, and recording duration varies in the range of $[2-39]$s. Despite this requirement of analyzing recordings of varying duration, current methods for voice pathology classification favor fixed length inputs, hence the need for fixed size segmentation and zero padding procedures. When it comes to voice pathology classification tasks, this is only a simplified approach, because recording duration is often correlated with some patients' disability to pronounce certain vowels.

Therefore, in order to deal with the limitations imposed by the small amount of training data and overcome the side effects of fixed-length segmentation and zero padding techniques, we propose an augmentation method that extracts multiple segments per recording as it is shown in Fig.3. The length of all segments stemming from a recording has been set equal to two-thirds the length of the recording but with different endpoints. As recording length varies, the set of extracted segments will inevitably contain segments of varying duration. The flowchart of the method is given in Fig. 4 and a respective pseudocode description is presented below:

- *Step 1*: Define a random starting point, st_k, between zero and one-third of the recording length measured in frames, i.e.,

$$st_k \in [0, \frac{L_i}{3}] \tag{1}$$

  where L_i is the number of frames of the i-th feature sequence ($L_i \in [124, 1462]$ in our experiments). We refer to frame numbers because it is assumed that a moving window technique is applied on the recording during a subsequent feature extraction stage, yielding a feature sequence per recording.
- *Step 2*: Define an endpoint, en_k, as:

$$en_k = st_k + \frac{2 \times L_i}{3} \tag{2}$$

- *Step 3*: Repeat the previous two steps five times, i.e., $k = 1, \ldots, 5$, yielding five segments starting in random positions of the recording, while ensuring that segment length is equal to two-thirds of the length of the recording (measured in number of frames).

Algorithm 1 Segmentation Algorithm

```
startvalue ← 0
endvalue ← random in [0, ≤ Li/3]
k ← 1
while k ≤ 5 do
    startpoint ← random in [startvalue, endvalue]
    stk ← startpoint
    enk ← stk + 2*Li/3
    k ← k + 1
end while
```

The aforementioned values for endpoint and segment duration selection are the result of a grid search performed in conjunction with the measurement of the classification performance of the classifier.

2) DATA AUGMENTATION WITH NOISE INJECTION

In addition to segment-based augmentation, we experiment with noise injection techniques. Of course, we note the existence of alternative popular time-domain methods for dataset augmentation, including time warping, pitch shifting, and

dynamic range compression [64], [65]. For example, time-warping methods stretch or compress the duration of a given audio signal without significantly altering basic signal properties. Pitch-shifting, on the other hand, lowers or increases the pitch of the audio recording without altering its length and dynamic range compression reduces the dynamic amplitude range of the audio signal. However, due to the nature of the audio signals under study, applying transformations that affect signal duration, pitch, or amplitude could alter sound properties that are necessary for discriminating among different types of voice pathology. Therefore, we excluded these data augmentation methods and instead focused on noise injection techniques, which are known to help mitigate over-fitting and improve the generalization capabilities of a model [66], [67]. We apply noise injection at the input signal during the training phase and not on the neural network parameters (layer activation outputs, weights, and gradients).

We experiment with the application of Gaussian noise, as in [67] and [68] and then proceed with coloured noise, which exhibits a different power spectrum profile. More specifically, we adopt established techniques for generating white, pink, and brown noise signals. In the augmented training dataset, each recording is randomly corrupted by one noise type. A pseudocode description follows:

For each audio recording in the training set:
Generate a white noise corrupted signal w.
Generate a pink noise corrupted signal p.
Generate a brown noise corrupted signal b.
Insert w, p and b in the training.

Noise injection is applied before the segmentation technique and, during the training stage, batches are formed via random selection over the final augmented training set. The classification impact of the data augmentation techniques under study is presented in the next section.

C. ABLATION STUDY OF MODALITIES

We first explore the contribution of each modality to the classification accuracy. Therefore, we start with testing a unimodal classifier that processes the extracted two-dimensional representations of audio signals through a convolutional branch. To understand why it is important to analyze signals at their original duration, we compare two versions of this classifier while ignoring data augmentation options. The first version employs a standard segmentation procedure with zero padding to extract fixed-length segments (1.28 seconds long) from each recording and create the training set. The classifier used in this version is a standard convolutional architecture consisting of a cascade of convolutional-ReLU layers followed by a dense layer comprising 128 neurons. On the other hand, we also evaluate an alternative fully convolutional architecture, where a global max-pooling operation replaces the dense layer. This modified version enables the analysis of each recording at its original duration, thus eliminating the need for a prior segmentation stage.

Table 2 displays a notable performance difference between the fully convolutional architecture and the conventional one. The fully convolutional model achieves a classification accuracy of 48.5% on the testing set, outperforming the conventional method's score of 36.5%. This result can be perceived as experimental evidence that processing audio recordings at their initial duration is a preferable approach compared to architectures that demand fixed-length segments at the input.

Furthermore, to assess the contribution of the medical parameters and perturbation features to the overall system performance, we experiment with two alternative configurations. The first configuration relies entirely on a fully connected branch that processes medical information and perturbation features. The second configuration implements the complete system shown in Fig. 2, where all input modalities are present. As it can be seen in Table 2, the fusion of medical and audio modalities yields a significantly improved classifier that exhibits a classification accuracy of 54.5%, i.e., an accuracy that is almost 8% higher than the performance of the medical/perturbation processing branch (47%).

TABLE 2. Classification accuracy for unimodal and bimodal classifiers.

Input	Model	Accuracy
Audio features	Convolutional	36.5%
Audio features	Fully convolutional	48.5%
Medical data, perturbation features	Fully connected	47.0%
Audio and medical data	Overall system	54.5%

D. DATA AUGMENTATION IMPACT

All previous models were trained without the application of data augmentation techniques at the training stage. To investigate the impact of augmentation methods on classification performance we conducted experiments and their results are shown in Table 3. Table 3 demonstrates the effect of segmentation-based augmentation on testing accuracy where it is observed that the application of segmentation-based augmentation raises the testing accuracy to 57.8%. This small increase is obtained by extracting five audio segments per recording using the algorithm described in Section V-B.

We further experiment with noise injection and investigate different power spectrum distributions of the added noise signals. We first add a standard Gaussian layer at the input of the fully convolutional branch of the combined model, to generate additive zero-mean Gaussian noise, with a standard deviation equal to 0.1. We use the relevant Keras implementation for this layer, which is only active during the training phase of the model. We observe that the presence of additive Gaussian noise contributes a further small improvement to the testing accuracy (accuracy is now increased to 58.5%).

We then study the impact of coloured noise injection on the model's performance. As shown in Table 3, when white noise, pink noise, and brown noise are individually injected, the classification accuracy becomes 59.5%, 60.0%, and 62.0%, respectively. Brown noise mainly targets low frequencies and therefore affects the first 13 MFCC coefficients of the

feature vector. This targeted effect on low frequencies is advantageous and enhances the classification robustness in that range compared to other types of noise. Therefore, the improved results obtained with brown noise injection can be attributed to its ability to specifically address low-frequency components.

Finally, we create an augmented training set by randomly corrupting each recording with one of the three noise types. This setup yields the best results, with the classification accuracy reaching 64.4%, a score that outperforms the best classifier of the 2019 FEMH voice data challenge (63%).

Table 3 summarizes our findings and shows that the best classification performance is achieved when sequence segmentation and injection of three coloured noise types are simultaneously applied during the training stage.

TABLE 3. Classification accuracy with respect to different augmentation techniques.

Segmentation	Noise injection	Testing accuracy
No	None	54.5%
Yes	None	57.8%
Yes	Gaussian	58.5%
Yes	Pink	60.0%
Yes	White	59.5%
Yes	Brown	62.0%
Yes	All colours	64.4%

To gain a deeper understanding of the performance of the best classifier across the four types of pathology, we also compute the confusion matrix (Table 4), with element (i, j) representing the number of testing samples with true label in the i-th class and predicted label in the j-class. It can be easily observed that in the case of hyperfunctional dysphonia (first row of the matrix), class recall is high and the distribution of errors is practically uniform across the other classes. A slightly different behavior can be observed for the case of phonotrauma (second row of the matrix), where the majority of false predictions are assigned to the class of vocal palsy and no neoplasm misclassification is observed. The third row indicates that the class of neoplasm disorder suffers from low recall (only eighteen neoplasm samples were correctly classified), with errors distributed almost uniformly over the remaining classes. Finally, from a classification perspective, vocal palsy behaves similarly to dysphonia regarding class precision and recall. Overall, it can be stated that classification performance is imbalanced over the four types of pathology, with the class of dysphonia attracting most of the errors (low-class precision) and the class of neoplasm pathology exhibiting the lowest recall.

E. NETWORK INTERPRETATION

Model explainability has been widely acknowledged [69] to be important in the field of healthcare and we thus provide insight into the functionality of the intermediate feature layers of the proposed network architecture and the learned patterns during the training stage.

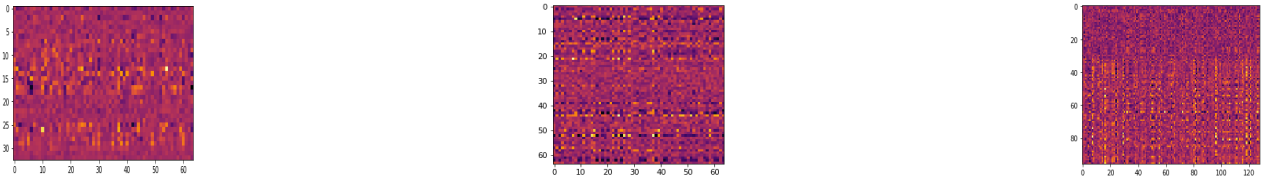
TABLE 4. Confusion matrix of the best-performing classifier.

Class	Dysphonia	Phonotr.	Neoplasm	Voc. Palsy
Dysphonia	38	3	4	5
Phonotrauma	4	37	0	9
Neoplasm	15	7	18	10
Vocal Palsy	6	5	3	36

We first focus on the functionality of the convolutional layers. To this end, we visualize the 2-D representation at the input of the network along with the feature activation maps for the four convolutional layers. It can be observed that the input images of different pathology classes exhibit similarities regarding the presence of common patterns and their variations, which, in turn, explains to a certain degree why the classification task under study is a hard one. Representative images are shown in Fig. 6, 7, 8, and 9, for vocal palsy, hyperfunctional dysphonia, phonotrauma, and neoplasm respectively. In addition, in Fig. 6b, 6c, 6d, 7b, 7c, 7d, 8b, 8c, 8d, 9b, 9c and 9d we show an indicative subset of the 32 activation maps of the fourth convolutional layer. These maps exhibit maximum activation output for the respective convolutional mask.

To further study the performance of the fully connected layers, we use the post-hoc interpretability method of feature analysis as in [70], where the interpretation of the functionality of a deep model can be extracted from each layer directly by the activation values of all neurons. Therefore, for each of the three fully connected layers, we depict the values of the activation weights of all neurons of the layer (x-axis) in relation to the layer's thirty-seven medical-perturbation input features sequence (y-axis). The process results in the three images in Fig. 5. More specifically, Figure 5b illustrates the values of the weights of the 64 neurons of Dense layer 1 with respect to the 37 input features. Horizontal linear patterns help us gain insight into which neurons are dominantly activated and, as a consequence, which corresponding features have the highest contribution. More specifically, we observe that features with indices 14, 15, 16 and 25, 26, and 27 have the greatest share in the model's prediction outcome. According to our mapping formula for transforming the medical descriptors to numbers, the specific indices represent parameters that describe a patient's symptoms (i.e., if they feel dysphonia, dryness, lumping, and if they have occupational vocal demands, hypertension, and head and neck cancer). On the other hand, we can observe that features with assigned indices into regions $[0, 4]$ and $[20, 25]$, trigger the lowest activation values. Therefore, we can conclude that these medical features have a smaller contribution to the network's inference capabilities. These features correspond to patients' demographic information, i.e., sex, age, onset, tiredness and night meal, chocking, eye dryness, smoking, and drinking.

Figure 5c shows the last fully connected layer, (Dense 3). At the input of this layer, audio and medical-perturbation embeddings are concatenated, with 32 features originating from the fully convolutional module and the remaining



(a) (b) (c)

FIGURE 5. Visualization of all the weights of the three fully connected layers.

(a) (b) (c) (d)

FIGURE 6. Vocal palsy: input "image" along with three feature activation maps of the final convolutional layer.

(a) (b) (c) (d)

FIGURE 7. Hyperfunctional dysphonia: input "image" along with three feature activation maps of the final convolutional layer.

(a) (b) (c) (d)

FIGURE 8. Phonotrauma: input "image" along with three feature activation maps of the final convolutional layer.

(a) (b) (c) (d)

FIGURE 9. Neoplasm: input "image" along with three feature activation maps of the final convolutional layer.

FIGURE 10. Perceptually uniform sequential colour map inferno.

64 ones from the fully connected branch. The image depicts the weight values of the 128 neurons of the Dense layer. An analysis of the graph reveals that the upper region (which refers to feature indices $0-30$) exhibits lower weight activation values, compared to the lower part that shows higher activation values. The lower region of the image refers to the feature embeddings created by the fully connected branch. This observation justifies that the second module of our model, which learns features stemming from medical information, is a vital part of the fused model. Therefore, medical

features improve considerably the classification capabilities of the model.

VI. DISCUSSION

A key feature of the proposed modular deep learning architecture is that it is trained in the presence of limited training data. Furthermore, different modalities are merged into a single model via separate processing branches, intermediate feature representations are consequently learned and the final classification decision is taken after concatenating learned features from the individual modalities. Our experimental results verify that the fusion of data from different modalities increases classification performance. In particular, comparative results among unimodal and bimodal architectures reveal that the inclusion of medical information contributes to a 6% performance improvement and that mid-term descriptors have a non-negligible contribution to the final classification decision.

Our study has also shown that recording duration and voice pathology are correlated. Therefore, our classifier, which is designed to process recordings of arbitrary duration via a fully convolutional stack, exhibits improved classification performance compared to fixed duration schemes. To that end, ignoring data augmentation techniques, the proposed architecture exhibits an improved classification accuracy of 48.5%, compared to the 36.5% accuracy of a conventional convolutional configuration.

Moreover, our study has revealed that an augmentation scheme that extracts segments of varying duration from the audio recordings can lead to improved generalization capabilities. In addition, noise injection with pink, white and brown coloured noise contributes to a further significant performance improvement, yielding a final classification accuracy of 64.4%, which outperforms the winner of the FEMH 2019 challenge by approximately 1%.

VII. CONCLUSION

We propose a multimodal classification framework for the four types of voice pathology of the FEMH dataset (hyperfunctional dysphonia, phonotrauma, laryngeal neoplasm, and unilateral vocal paralysis). Our experiments on the fusion of audio-based features and medical descriptors verify that medical parameters can serve as a supplementary information source for voice pathology classification. Furthermore, the proposed segmentation-based data augmentation and coloured noise injection techniques have shown to be effective data augmentation techniques for the task at hand. Our exploratory study has also justified our claim that in a voice pathology classification task, models should be designed to process audio recordings at their original, possibly varying duration. This result reinforces the conclusion that sustained utterance duration and intensity are frequently affected by the disorder and, therefore, such information should not be discarded, as it has so far been the case with conventional fixed-size segmentation schemes and zero padding procedures. Future work will focus on problems involving a larger number of phonemes and transfer learning among datasets and tasks, to provide a solution that will be able to deal with diverse information sources regarding structure and data volume.

REFERENCES

[1] S. A. Syed, M. Rashid, and S. Hussain, "Meta-analysis of voice disorders databases and applied machine learning techniques," *Math. Biosci. Eng.*, vol. 17, no. 6, pp. 7958–7979, 2020.

[2] K. Daoudi and B. Bertrac, "On classification between normal and pathological voices using the MEEI-KayPENTAX database: Issues and consequences," in *Proc. INTERSPEECH*, 2014, pp. 1–6.

[3] S.-J. Choi, H.-S. Choi, J.-O. Kim, and Y.-L. Choi, "Comparison of maximum phonation time associated with the changes in vocal intensity in patients with unilateral vocal fold palsy and sulcus vocalis," *Phonetics Speech Sci.*, vol. 4, no. 1, pp. 125–131, Mar. 2012.

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[5] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, K. H. Malki, T. A. Mesallam, and M. Farahat Ibrahim, "Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions," *IEEE Access*, vol. 6, pp. 6961–6974, 2018.

[6] A. Fethi and F. Mohamed, "Voice pathologies classification using GMM and SVM classifiers," *Int. J. Math. Comput. Simul.*, vol. 15, pp. 110–114, Nov. 2021.

[7] S. A. Syed, M. Rashid, S. Hussain, A. Imtiaz, H. Abid, and H. Zahid, "Inter classifier comparison to detect voice pathologies," *Math. Biosci. Eng.*, vol. 18, no. 3, pp. 2258–2273, 2021.

[8] N. Souissi and A. Cherif, "Speech recognition system based on short-term cepstral parameters, feature reduction method and artificial neural networks," in *Proc. 2nd Int. Conf. Adv. Technol. Signal Image Process. (ATSIP)*, 2016, pp. 667–671.

[9] P. Harar, J. B. Alonso-Hernandezy, J. Mekyska, Z. Galaz, R. Burget, and Z. Smekal, "Voice pathology detection using deep learning: A preliminary study," in *Proc. IWOBI*, 2017, pp. 1–4.

[10] J. Moon and S. Kim, "An approach on a combination of higher-order statistics and higher-order differential energy operator for detecting pathological voice with machine learning," in *Proc. ICTC*, 2018, pp. 46–51.

[11] V. Guedes, F. Teixeira, A. Oliveira, J. Fernandes, L. Silva, A. Junior, and J. P. Teixeira, "Transfer learning with AudioSet to voice pathologies identification in continuous speech," *Proc. Comput. Sci.*, vol. 164, pp. 662–669, Jan. 2019.

[12] S. A. Syed, M. Rashid, S. Hussain, and H. Zahid, "Comparative analysis of CNN and RNN for voice pathology detection," *BioMed Res. Int.*, vol. 2021, pp. 1–8, Apr. 2021.

[13] A. Al-nasheri, G. Muhammad, M. Alsulaiman, and Z. Ali, "Investigation of voice pathology detection and classification on different frequency regions using correlation functions," *J. Voice*, vol. 31, no. 1, pp. 3–15, Jan. 2017.

[14] A. A. Dibazar, S. Narayanan, and T. W. Berger, "Feature analysis for automatic detection of pathological speech," in *Proc. 2nd Joint 24th Annu. Conf. Annu. Fall Meeting Biomed. Eng. Soc., Eng. Med. Biol.*, 2002, pp. 182–183.

[15] Y. Zhang and J. J. Jiang, "Acoustic analyses of sustained and running voices from patients with laryngeal pathologies," *J. Voice*, vol. 22, no. 1, pp. 1–9, Jan. 2008.

[16] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 2, pp. 370–379, Feb. 2011.

[17] C. D. P. Crovato and A. Schuck, "The use of wavelet packet transform and artificial neural networks in analysis and classification of dysphonic voices," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 10, pp. 1898–1900, Oct. 2007.

[18] N. Malyska, T. F. Quatieri, and D. Sturim, "Automatic dysphonia recognition using biologically-inspired amplitude-modulation features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2005, pp. 1873–1876.

[19] M. Markaki, Y. Stylianou, J. D. Arias-Londono, and J. I. Godino-Llorente, "Dysphonia detection based on modulation spectral features and cepstral coefficients," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2010, pp. 5162–5165.

[20] M. Markaki and Y. Stylianou, "Voice pathology detection and discrimination based on modulation spectral features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 1938–1948, Sep. 2011.

[21] D. Połap, M. Woźniak, R. Damaševičius, and R. Maskeliūnas, "Bio-inspired voice evaluation mechanism," *Appl. Soft Comput.*, vol. 80, pp. 342–357, Jul. 2019.

[22] A. Lauraitis, R. Maskeliūnas, R. Damaševičius, and T. Krilavičius, "Detection of speech impairments using cepstrum, auditory spectrogram and wavelet time scattering domain features," *IEEE Access*, vol. 8, pp. 96162–96172, 2020.

[23] R. Hamdi, S. Hajji, and A. Cherif. (2020). *Recognition of Pathological Voices by Human Factor Cepstral Coefficients (HFCC)*. [Online]. Available: https://www.researchsquare.com/article/rs-23108/v1

[24] N. Souissi and A. Cherif, "Artificial neural networks and support vector machine for voice disorders identification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 5, pp. 339–344, 2016.

[25] R. Maskeliūnas, A. Kulikajevas, R. Damaševičius, K. Pribuišis, N. Ulozaitė-Stanienė, and V. Uloza, "Lightweight deep learning model for assessment of substitution voicing and speech after laryngeal carcinoma surgery," *Cancers*, vol. 14, no. 10, p. 2366, May 2022.

[26] M. Pham, J. Lin, and Y. Zhang, "Diagnosing voice disorder with machine learning," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 5263–5266.

[27] K. Degila, R. Errattahi, and A. E. Hannani, "The UCD system for the 2018 FEMH voice data challenge," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 5242–5246.

[28] C. Bhat and S. K. Kopparapu, "FEMH voice data challenge: Voice disorder detection and classification using acoustic descriptors," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 5233–5237.

[29] M. Ju, Z. Jiang, Y. Chen, and S. Ray, "A multi-representation ensemble approach to classifying vocal diseases," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 5258–5262.

[30] T. Grzywalski, A. Maciaszek, A. Biniakowski, J. Orwat, S. Drgas, M. Piecuch, R. Belluzzo, K. Joachimiak, D. Niemiec, J. Ptaszynski, and K. Szarzynski, "Parameterization of sequence of MFCCs for DNN-based voice disorder detection," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 5247–5251.

[31] M. Pishgar, F. Karim, S. Majumdar, and H. Darabi, "Pathological voice classification using mel-cepstrum vectors and support vector machine," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 5267–5271.

[32] S.-H. Fang, Y. Tsao, M.-J. Hsiao, J.-Y. Chen, Y.-H. Lai, F.-C. Lin, and C.-T. Wang, "Detection of pathological voice using cepstrum vectors: A deep learning approach," *J. Voice*, vol. 33, no. 5, pp. 634–641, Sep. 2019.

[33] S.-H. Fang, C.-T. Wang, J.-Y. Chen, Y. Tsao, and F.-C. Lin, "Combining acoustic signals and medical records to improve pathological voice classification," *APSIPA Trans. Signal Inf. Process.*, vol. 8, no. 1, p. e14, 2019.

[34] Z.-Y. Chuang, X.-T. Yu, J.-Y. Chen, Y.-T. Hsu, Z.-Z. Xu, C.-T. Wang, F.-C. Lin, and S.-H. Fang, "DNN-based approach to detect and classify pathological voice," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 5238–5241.

[35] R. Islam, E. Abdel-Raheem, and M. Tarique, "Voice pathology detection using convolutional neural networks with electroglottographic (EGG) and speech signals," *Comput. Methods Programs Biomed. Update*, vol. 2, 2022, Art. no. 100074.

[36] R. Islam, E. Abdel-Raheem, and M. Tarique, "Deep learning based pathological voice detection algorithm using speech and electroglottographic (EGG) signals," in *Proc. Int. Conf. Electr. Comput. Technol. Appl. (ICECTA)*, Nov. 2022, pp. 127–131.

[37] L. Geng, H. Shan, Z. Xiao, W. Wang, and M. Wei, "Voice pathology detection and classification from speech signals and EGG signals based on a multimodal fusion method," *Biomed. Eng., Biomedizinische Technik*, vol. 66, no. 6, pp. 613–625, Dec. 2021.

[38] I. Miliaresi, A. Pikrakis, and K. Poutos, "A deep multimodal voice pathology classifier with electroglottographic signal processing capabilities," in *Proc. 7th Int. Conf. Frontiers Signal Process. (ICFSP)*, Sep. 2022, pp. 109–113.

[39] A. Ksibi, N. A. Hakami, N. Alturki, M. M. Asiri, M. Zakariah, and M. Ayadi, "Voice pathology detection using a two-level classifier based on combined CNN–RNN architecture," *Sustainability*, vol. 15, no. 4, p. 3204, Feb. 2023.

[40] F. Alshehri and G. Muhammad, "A comprehensive survey of the Internet of Things (IoT) and AI-based smart healthcare," *IEEE Access*, vol. 9, pp. 3660–3678, 2021.

[41] I. Miliaresi, K. Poutos, and A. Pikrakis, "Combining acoustic features and medical data in deep learning networks for voice pathology classification," in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2021, pp. 1190–1194.

[42] S.-Y. Tsui, Y. Tsao, C.-W. Lin, S.-H. Fang, F.-C. Lin, and C.-T. Wang, "Demographic and symptomatic features of voice disorders and their potential application in classification using machine learning algorithms," *Folia Phoniatrica Logopaedica*, vol. 70, nos. 3–4, pp. 174–182, 2018.

[43] K. A. Islam, D. Perez, and J. Li, "A transfer learning approach for the 2018 FEMH voice data challenge," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 5252–5257.

[44] J. D. Arias-Londoño, J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, "ByoVoz automatic voice condition analysis system for the 2018 FEMH challenge," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 5228–5232.

[45] N. M. Manshouri, "Identifying COVID-19 by using spectral analysis of cough recordings: A distinctive classification study," *Cognit. Neurodynamics*, vol. 16, no. 1, pp. 239–253, Feb. 2022.

[46] L. Verde, G. De Pietro, A. Ghoneim, M. Alrashoud, K. N. Al-Mutib, and G. Sannino, "Exploring the use of artificial intelligence techniques to detect the presence of coronavirus Covid-19 through speech and voice analysis," *IEEE Access*, vol. 9, pp. 65750–65757, 2021.

[47] Z. Ren, Y. Chang, K. D. Bartl-Pokorny, F. B. Pokorny, and B. W. Schuller, "The acoustic dissection of cough: Diving into machine listening-based COVID-19 analysis and detection," *J. Voice*, Jun. 2022, doi: 10.1016/j.jvoice.2022.06.011.

[48] A. Tena, F. Clariá, and F. Solsona, "Automated detection of COVID-19 cough," *Biomed. Signal Process. Control*, vol. 71, Jan. 2022, Art. no. 103175.

[49] A. Gokcen, B. Karadag, C. Riva, and A. Boyaci, "Artificial intelligence-based COVID-19 detection using cough records," *Electrica*, vol. 21, no. 2, pp. 203–209, 2021.

[50] X.-Y. Chen, Q.-S. Zhu, J. Zhang, and L.-R. Dai, "Supervised and self-supervised pretraining based covid-19 detection using acoustic breathing/cough/speech signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2022, pp. 561–565.

[51] M. R. Kamble, J. Patino, M. A. Zuluaga, and M. Todisco, "Exploring auditory acoustic features for the diagnosis of covid-19," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2022, pp. 566–570.

[52] J. Laguarta, F. Hueto, and B. Subirana, "COVID-19 artificial intelligence diagnosis using only cough recordings," *IEEE Open J. Eng. Med. Biol.*, vol. 1, pp. 275–281, 2020.

[53] M. Pahar, M. Klopper, R. Warren, and T. Niesler, "COVID-19 detection in cough, breath and speech using deep transfer learning and bottleneck features," *Comput. Biol. Med.*, vol. 141, Feb. 2022, Art. no. 105153.

[54] M.-J. Son and S.-P. Lee, "COVID-19 diagnosis from crowdsourced cough sound data," *Appl. Sci.*, vol. 12, no. 4, p. 1795, Feb. 2022.

[55] J. P. Teixeira and P. O. Fernandes, "Acoustic analysis of vocal dysphonia," *Proc. Comput. Sci.*, vol. 64, pp. 466–473, Jan. 2015.

[56] J. H. Middendorf, "Phonotrauma in children: Management and treatment," *ASHA Leader*, vol. 12, no. 15, pp. 14–17, Nov. 2007.

[57] M. N. Syamal and M. S. Benninger, "Vocal fold paresis: A review of clinical presentation, differential diagnosis, and prognostic indicators," *Current Opinion Otolaryngology Head Neck Surgery*, vol. 24, no. 3, pp. 197–202, Jun. 2016.

[58] A. Koroulakis and M. Agarwal. (2020). *Laryngeal Cancer*. [Online]. Available: https://journals.lww.com/co-otolaryngology/toc/2014/04000

[59] M. Mauch and S. Dixon, "PYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2014, pp. 659–663.

[60] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, 2015, pp. 448–456.

[61] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[62] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 279–309.
[63] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*.
[64] J. Laroche, "Time and pitch scale modification of audio signals," in *Applications of Digital Signal Processing to Audio and Acoustics*. Cham, Switzerland: Springer, 2002, pp. 279–309.
[65] B. Ninness and S. J. Henriksen, "Time-scale modification of speech signals," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1479–1488, Apr. 2008.
[66] G. An, "The effects of adding noise during backpropagation training on a generalization performance," *Neural Comput.*, vol. 8, no. 3, pp. 643–674, Apr. 1996.
[67] L. Holmstrom and P. Koistinen, "Using additive noise in back-propagation training," *IEEE Trans. Neural Netw.*, vol. 3, no. 1, pp. 24–38, Jan. 1992.
[68] J. Timmer and M. König, "On generating power law noise," *Astron. Astrophys.*, vol. 300, pp. 707–710, Aug. 1995.
[69] F.-L. Fan, J. Xiong, M. Li, and G. Wang, "On interpretability of artificial neural networks: A survey," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 5, no. 6, pp. 741–760, Nov. 2021.
[70] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," 2015, *arXiv:1506.06579*.

IOANNA MILIARESI received the B.S. degree in physics and the M.S. degree in telecommunications engineering from the National and Kapodistrian University of Athens, Greece. She is currently pursuing the Ph.D. degree in audio processing algorithms for voice pathology classification with the University of Piraeus, Greece. From 1998 to 2003, she was a Software Engineer with the Siemens Software Center for Voice Telecommunication Networks. From 2003 to 2018, she was a Lecturer in applications with the Department of Sound and Musical Instruments Technology, Technological Institute of Ionian Islands. Since 2018, she has been a Lecturer in applications with the Department of Audio and Visual Arts, Ionian University. Her research interests include speech and audio processing.

AGGELOS PIKRAKIS (Member, IEEE) received the Diploma degree in computer engineering from the University of Patras, Greece, and the Ph.D. degree in computer science from the University of Athens, Greece. He is currently an Assistant Professor (tenured position) with the Department of Informatics, University of Piraeus, teaching courses related to machine learning and audio processing. He is the co-inventor of two AI/ML patents and he has coauthored two international textbooks in the English language and more than 70 papers in international peer-reviewed scientific journals/conferences. His research interests include solving audio analysis problems with deep neural networks, hidden Markov models, Bayesian architectures, and sequence alignment methods. His work has received award recognition and he was a recipient of the 2019 EURASIP Meritorious Service Award.