Tirronen, Saska; Kadiri, Sudarsana; Alku, Paavo

Hierarchical Multi-class Classification of Voice Disorders Using Self-supervised Models and Glottal Features

Published in:
IEEE Open journal of Signal Processing

DOI:
10.1109/OJSP.2023.3242862

Published: 06/02/2023

Document Version
Publisher's PDF, also known as Version of record

IEEE Open Journal of Signal Processing

Received 27 October 2022; revised 17 January 2023; accepted 25 January 2023. Date of publication 6 February 2023; date of current version 23 February 2023. The review of this article was arranged by Associate Editor Hung-yi Lee.

Digital Object Identifier 10.1109/OJSP.2023.3242862

Hierarchical Multi-Class Classification of Voice Disorders Using Self-Supervised Models and Glottal Features

SASKA TIRRONEN, SUDARSANA REDDY KADIRI (Member, IEEE), AND PAAVO ALKU (Fellow, IEEE)

Department of Information and Communications Engineering, Aalto University, 02150 Espoo, Finland

CORRESPONDING AUTHOR: SASKA TIRRONEN. (e-mail: saska.tirronen@aalto.fi)

This work was supported in part by the Academy of Finland under Grant 330139 and in part by the Aalto University (the MEC Program for India).

ABSTRACT Previous studies on the automatic classification of voice disorders have mostly investigated the binary classification task, which aims to distinguish pathological voice from healthy voice. Using multi-class classifiers, however, more fine-grained identification of voice disorders can be achieved, which is more helpful for clinical practitioners. Unfortunately, there is little publicly available training data for many voice disorders, which lowers the classification performance on data from unseen speakers. Earlier studies have shown that the usage of glottal source features can reduce data redundancy in detection of laryngeal voice disorders. Another approach to tackle the problems caused by scarcity of training data is to utilize deep learning models, such as wav2vec 2.0 and HuBERT, that have been pre-trained on larger databases. Since the aforementioned approaches have not been thoroughly studied in the multi-class classification of voice disorders, they will be jointly studied in the present work. In addition, we study a hierarchical classifier, which enables task-wise feature optimization and more efficient utilization of data. In this work, the aforementioned three approaches are compared with traditional mel frequency cepstral coefficient (MFCC) features and one-vs-rest and one-vs-one SVM classifiers. The results in a 3-class classification problem between healthy voice and two laryngeal disorders (hyperfunctional dysphonia and vocal fold paresis) indicate that all the studied methods outperform the baselines. The best performance was achieved by using features from wav2vec 2.0 LARGE together with hierarchical classification. The balanced classification accuracy of the system was 62.77% for male speakers, and 55.36% for female speakers, which outperformed the baseline systems by an absolute improvement of 15.76% and 6.95% for male and female speakers, respectively.

INDEX TERMS Pathological voices, voice disorders, hierarchical classification, glottal source extraction, multi-class classification, Wav2vec, HuBERT.

I. INTRODUCTION

Automatic classification of voice disorders has been studied widely in the past two decades. The focus has been on detection of voice disorders (i.e., the binary classification problem) [1], [2], [3], [4], [5], [6], [7], [8], while classification of multiple voice disorders (i.e., the multi-class problem) [9], [10], [11], [12], [13] has remained less studied. In the detection problem, the automatic system distinguishes disordered voice from healthy voice. As there are many voice disorders, including both organic and functional, a multi-class classifier, which enables the classification between healthy voice and *several* different disorders, would be more useful for clinical practitioners. In the current study, a 3-class voice pathology classification problem is studied by investigating the classification between two laryngeal voice disorders (hyperfunctional dysphonia and vocal fold paresis) and healthy voice.

Traditionally, automatic detection systems have been constructed as pipelines that consist of separate feature extraction and classification steps [2], [3], [4], [5], [6], [7], [8]. In the feature extraction, the voice signal is mapped into a vector in a suitably designed feature space. The mapped vector representations are then used by a machine learning algorithm to separate healthy voices from disordered voices. In contrast, some recent studies have studied deep learning -based systems

that combine the feature extraction and classification steps into a single neural network that inputs a voice signal (or its spectrogram) and gives the classification label as output [6], [14], [15]. Such systems are often referred to as end-to-end classifiers.

In general, pipeline systems require smaller amounts of training data than end-to-end systems. This is because the classification problem of end-to-end systems is more complex, requiring the model to learn the optimal feature mapping from the data. As the amount of available training data in voice disorder databases is typically small, the focus of this paper is on pipeline systems. However, a small amount of training data is a problem for pipeline systems too, and it may cause low classification performance on unseen data. The problems caused by the scarcity of training data are particularly severe for multi-class classification tasks that call for training data representing several voice pathologies.

In the current study, we investigate three approaches to improve multi-class classification of voice disorders based on the pipeline system architecture. The *first* approach corresponds to using the glottal source signal in feature extraction. The *second* approach corresponds to using self-supervised models as pre-trained feature extractors. The *third* approach corresponds to using a hierarchical multi-class classifier architecture. The first and second approaches are related to feature extraction, and the third one is related to classification. Hypothetically, the best benefit may be gained by using the feature-based and classification-based approaches together, and combining them with data augmentation methods, as proposed by [11] and [16]. However, data augmentation is outside of the scope of the current study.

The first approach aims to take advantage of the source of voiced speech, the glottal excitation, in the feature extraction. The glottal excitation is first estimated using a glottal inverse filtering algorithm, and the estimated source signal is expressed using the mel frequency cepstral coefficient (MFCC) features. This approach helps the classifier learn more generalizable functions from small data sets, because vocal tract information, which is removed by glottal inverse filtering, may be mostly redundant for the classification of the selected disorders. Glottal source features have been studied in a few earlier studies in automatic detection of voice pathologies [7], [8], [17]. However, the glottal source has not been used previously in multi-class classification tasks, where the problem of small data is most severe.

In the second approach, we take advantage of wav2vec 2.0 [18] and HuBERT [19], which are frameworks for self-supervised learning of representations from raw speech signals. The self-supervised models were pre-trained on databases in automatic speech recognition (ASR). In the pre-training phase, the models have learned to extract features that generalize well to a variety of speech-related tasks and unseen data. Therefore, the pre-trained models are used as feature extractors by utilizing their hidden layer outputs.

Pre-trained self-supervised models have been used before to improve performance of ASR systems in recognition of disordered speech [20], [21]. The wav2vec 2.0 models have also been used, for example, for detection of aphasia [22], for detection of stuttering [23], and for speech rating of disordered children's speech [24]. Various pre-training approaches have been used to detect Alzheimer's disease [25], [26], and heart failure [27]. However, only a few studies have applied these techniques on multi-class classification of voice disorders. In [28], the pre-trained VGGish model was used for feature extraction in several multi-class problems. In [29], transfer learning methods between three disorders were studied. However, as per our knowledge, the utilization of the state-of-the-art self-supervised models as feature extractors in multi-class classification of voice disorders has not been studied before.

The third approach aims to improve the multi-class classification of voice disorders by using a hierarchical classifier architecture that combines two binary classifiers into a 3-class classifier. In the first step, a binary classification is done between healthy and *disordered* voices. In the second step, the samples that were classified as disordered are classified into the two selected laryngeal disorders. This approach is an efficient way to use training data, as each voice sample can be utilized twice in learning the two sub-problems. In this work, SVMs and fine-tuned self-supervised models are used as the two binary classifiers.

Hierarchical architectures have been used in earlier works for the classification of laryngeal voice disorders [28], [30], [31], and to classify between dysarthria, apraxia of speech, and neurotypical speech [12]. The work in [17] evaluated hierarchical sub-problems individually instead of the full multi-class problem. However, the classification of the two voice pathologies studied in the current paper (hyperfunctional dysphonia and vocal fold paresis) has not been investigated before using hierarchical classifier architectures. One important criterion for the selection of these disorders was the relatively small amount of training data of the two disorders, which makes them difficult to classify.

In summary, this work studies the effectiveness of three different approaches to alleviate the data scarcity problem in multi-class classification of voice disorders. These three approaches are highlighted with green color in Fig. 1, and they are:

1) The usage of glottal source signals in feature extraction.
2) The usage of a pre-trained self-supervised model (wav2vec 2.0 and HuBERT) as a feature extractor.
3) The usage of a hierarchical classifier.

The three approaches are compared to commonly used baselines. The glottal MFCCs and self-supervised feature extractors are compared to traditional MFCCs, which are the most popular features in voice disorder detection [2], [3], [4], [5], [7], [9], [10], [13], [17], [32]. Hierarchical classifiers are compared to SVM in the popular one-vs-one (OvO) and one-vs-rest (OvR) architectures [9].

The remaining part of the paper is structured as follows. Section II describes the proposed methods and their technical details. Section III describes the details of the experimental

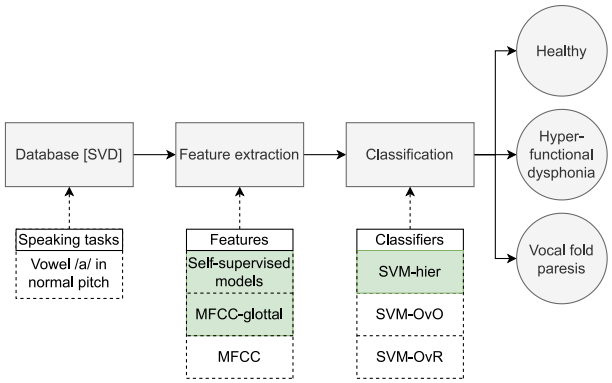

FIGURE 1. Block diagram of the pipeline system. The proposed methods for feature extraction and classification are indicated by green color.

setup, including the database, training and evaluation process, and the experiment runs. The experimental results are presented in Section IV. Finally, Section V summarizes the paper and presents final conclusions.

II. PROPOSED SYSTEM

In this work, voice disorder classification is performed by using a pipeline system that consists of separate feature extraction and classification steps. The classification is performed between three voice classes (healthy voice, hyperfunctional dysphonia, and vocal fold paresis) as illustrated in Fig. 1. The following sub-sections describe the technical details of the feature extraction and classification steps.

A. FEATURES

The voice signal is first pre-processed by re-sampling it to 16 kHz and by removing silent segments. All samples shorter than 750 ms are left out. Each utterance is normalized by dividing it with the signal's maximum absolute value. The baseline MFCC features are extracted by computing 13 coefficients with their delta and delta-delta coefficients. A frame-length of 25 ms is used with a shift of 5 ms.

As described in Section I, the first proposed approach to improve multi-class classification of voice pathologies corresponds to using MFCCs computed from glottal source waveforms (denoted as MFCC-glottal) in the feature extraction. First, the glottal source waveform is estimated using the quasi-closed phase (QCP) glottal inverse filtering method proposed in [33]. MFCCs are then extracted from the glottal waveform using the same procedure as in the baseline MFCCs.

The second approach introduced in Section I is to extract features by utilizing a self-supervised model that has been pre-trained using ASR databases. Three different self-supervised models are included in this work. Firstly, we use pre-trained wav2vec 2.0 BASE [18] that was pre-trained and fine-tuned using 960 hours of Librispeech [34]. Secondly, we use pre-trained wav2vec 2.0 LARGE [18], [35], which was pre-trained using a combination of three ASR databases (CommonVoice [36], BABEL [37], and Multilingual Librispeech [34]). Thirdly, we use HuBERT LARGE [19], which

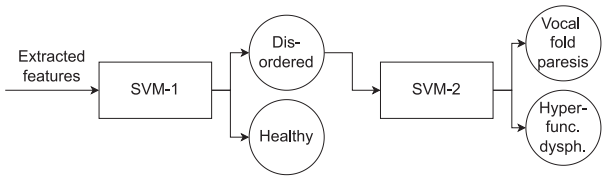

FIGURE 2. The hierarchical SVM classifier that is used in this work.

was pre-trained on the Libri-Light database [38] and further fine-tuned using 960 hours of Librispeech [34].

Both wav2vec 2.0 LARGE and HuBERT LARGE include 24 transformer blocks in their context networks and the model dimension is 1024, whereas wav2vec BASE only includes 12 transformer blocks and the model dimension is 768. For each of these models, the feature vectors are derived by computing the temporal averages of the relative positional embeddings from the output of each transformer layer of the context network. Similar computation is also done for the input of the first transformer layer. Therefore, the number of feature vectors is 25 for LARGE variations and 13 for BASE variations, and the dimension of the feature space equals the model dimension. In order to denote the feature vectors from each layer, the feature vectors are indexed in an increasing order. The input to the context network has index 0, and the output of the final embedding layer has indexes 24 and 12 for the LARGE and the BASE variations, respectively.

B. CLASSIFIERS

The third approach to improve multi-class classification of voice disorders is the use of a hierarchical classifier, which consists of two binary classifiers. Fig. 2 shows an illustration of a hierarchical classifier with two binary SVMs (SVM-hier) that is used in this work. The first classifier (SVM-1) distinguishes disordered voices from healthy voices. For the voice samples detected as disordered, the second classifier (SVM-2) classifies the pathology either as hyperfunctional dysphonia or vocal fold paresis. In addition to SVM-hier, another hierarchical system is examined that uses fine-tuned wav2vec 2.0 LARGE models as binary classifiers. This model is referred to as wav2vec-LARGE-hier. The hierarchical classifiers are compared with SVMs based on OvO (SVM-OvO) and OvR (SVM-OvR), as they have been widely used in multi-class classification [9], [10].

Hierarchical classifier, as well as the baseline OvO and OvR systems, divides the full multi-class problem into less complex sub-problems. These sub-problems are solved individually by dedicated classifiers, which effectively shares the total complexity of the task between them. This ensures that each training sample can be utilized in several parts of the architecture to learn different parts of the full problem, which can increase the utility of each training sample. This can help to train classifiers with small databases. In addition, the modularity of hierarchical classifiers can also be taken advantage of by medical practitioners. The hierarchical structure namely makes it possible to perform diagnosis as a sequence of

increasingly detailed evaluations, starting from the detection of disordered markers, and ending at a detailed diagnosis of the disorder type. Furthermore, each classifier in the hierarchy can be replaced without a need to modify or change any other classifiers. This is naturally a desirable feature of a system, as it allows for easy maintenance and continuous development.

III. EXPERIMENTAL SETUP

This section describes the experimental setup that is used in the current study. First, the voice database used in the study is described. Second, the training and testing processes of the classifiers are discussed. Finally, the details of each individual experiment are provided.

A. DATABASE

The current study uses voice data of the Saarbrücken Voice Disorders (SVD) database [39], [40]. We selected this database, because it is publicly available and covers voice samples from both genders for a variety of laryngeal voice disorders. The database contains 71 different disorders. The recordings were conducted in sessions, where each speaker conducted four speaking tasks: a pronunciation of the German sentence 'Guten Morgen, wie geht es Ihnen?' ('Good morning, how are you?'), and sustained pronunciations of three vowels (/a/, /i/, /u/). The vowels were pronounced by varying pitch in four types (low, normal, high, and low-high-low). The database contains samples from 1853 speakers in 2225 sessions.

This work includes samples of healthy voices, as well as pathological voice samples of hyperfunctional dysphonia and vocal fold paresis. These two voice disorders were selected because they are among the most prevalent voice disorders[1] [41], [42], [43]. Another reason for the selection is the small amount of data for both of the voice disorders in SVD (213 recording sessions for hyperfunctional dysphonia and 213 recording sessions for vocal fold paresis). This enables studying the 3-class classification task using pipeline classifiers in a scenario with a small amount of training data as discussed in Section I. Furthermore, in order to simulate a voice data scenario which is not only of a small size but which also could be generalized to other databases than SVD, we only selected voices that represent one popular speaking task, namely the sustained phonation of the vowel /a/ in normal pitch. We used samples from the speakers who had not had any surgeries or voice therapy prior to recordings, and who were 19-60 years old at the time of the recordings. In addition, we left out samples that were shorter than 750 ms. This resulted in data subsets that are visualized in Fig. 3.

B. TRAINING AND TESTING

All classifiers were trained by using 5-fold cross-validation (CV). All samples from each speaker were always contained within a single fold, to ensure that a model does not learn

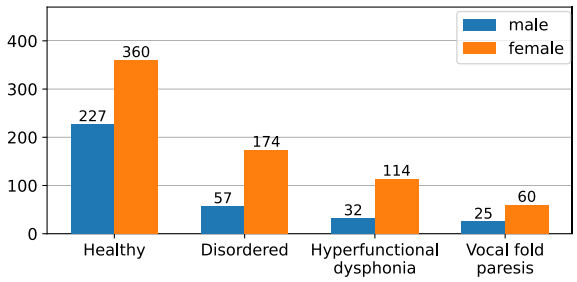

FIGURE 3. Number of recording sessions in the selected subset of the database. Included are healthy voices, and voices with hyperfunctional dysphonia and vocal fold paresis.

to classify voice samples based on speaker identity. In each iteration, one of the folds was reserved for evaluation, and the other folds were used for training. Performance metrics were computed based on the predictions that were made on the evaluation fold. The metrics include balanced classification accuracy, class-wise precision, class-wise recall, and class-wise F1 score. The 5-fold CV was performed 4 times with different random states, to get a total of 20 evaluations. For all hierarchical classifiers, the training process was performed separately for the two binary sub-problems.

As part of the cross-validation process, the number of samples in the different classes was balanced by duplicating the samples of the smallest classes. Even though this approach is most likely not optimal, it performed better in our initial tests compared to balancing the classes by leaving out samples from the majority classes. The balancing was done for each fold, which resulted in balanced data for both training and evaluation.

Some aspects of the training process were different between the experiments where the SVM-based classifiers were used and the experiments where the self-supervised models were fine-tuned and used as classifiers. When SVMs were used, the training and test features were both z-score normalized with the mean and standard deviation of the training data. Also, for each of the SVM classifiers, hyperparameters were optimized by grid-search. The searched parameter values were identical to the ones used and described in [44, p. 27–28]. For each of the fold iterations, all parameter combinations were evaluated, and the one that achieved the best mean balanced accuracy was selected.

In contrast, when the self-supervised models were fine-tuned, the input signals were pre-processed as in the pre-processing stage of feature extraction (see Section II-A). Grid-search was not applied to any hyper-parameters. Fine-tuning was conducted once for each CV iteration by minimizing cross-entropy loss by using the AdamW optimizer [45] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate was 0.0005 and it was reduced linearly. Batch size was 32 and the maximum number of epochs was 50.

C. EXPERIMENTS

The experiments consisted of two parts. The first part evaluates the two feature-based approaches glottal MFCCs and

[1]https://www.tgh.org/institutes-and-services/conditions/hyperfunctional-dysphonia

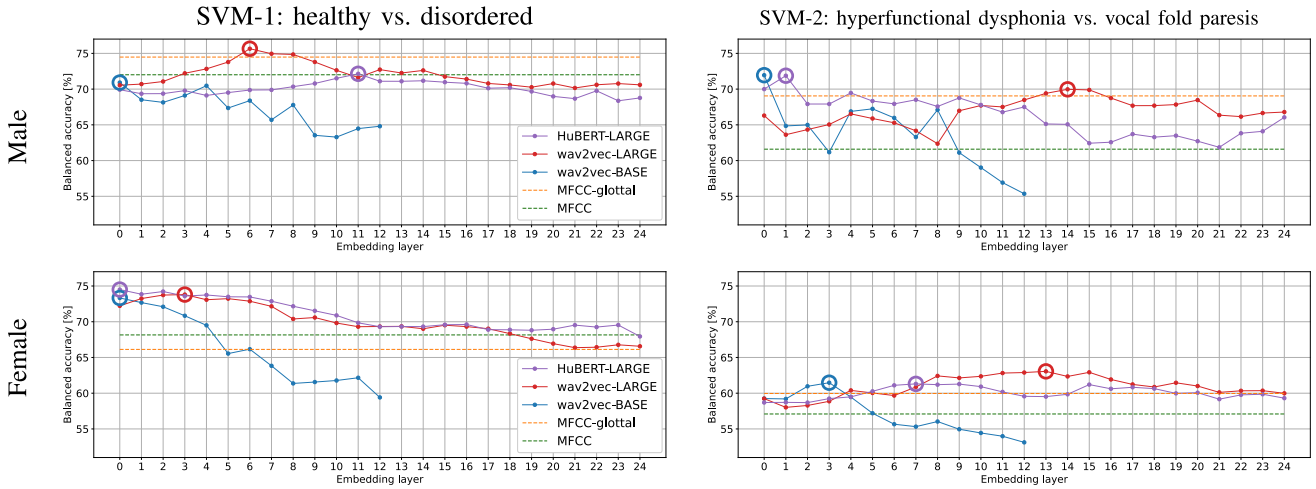

FIGURE 4. **Classification accuracies obtained in binary tasks of healthy vs. disordered (SVM-1) and hyperfunctional dysphonia vs. vocal fold paresis (SVM-2). The green dashed line represents the baseline MFCC features. The orange dashed line represents the MFCC-glottal features. The solid lines represent the self-supervised features, with the tick labels indicating the index of the corresponding layer. Index 0 refers to the input to the first embedding layer, other indexes refer to the output of the corresponding layer. For each self-supervised feature, the best values are highlighted with larger circles.**

self-supervised features, and the second part evaluates the classifier-based approach (hierarchical classification).

In the first part, the comparison of the baseline MFCC, MFCC-glottal and self-supervised features was conducted using two binary classification tasks: healthy vs. disordered, and hyperfunctional dysphonia vs. vocal fold paresis. These binary problems were selected for the feature-related experiments, because they are the two sub-problems of hierarchical classification.

In the second part, hierarchical multi-class classification was examined. The comparison was first made between SVM-hier and the baseline classifiers (SVM-OvO and SVM-OvR) by using MFCCs. In addition, we examined the effect of using the self-supervised models together with hierarchical classification. For both hierarchical steps, we selected the self-supervised model that achieved the best performances in the corresponding sub-problems (SVM-1 and SVM-2) in the first part of the experiments. These models were then used in the hierarchical framework in two alternative ways. Firstly, by extracting the self-supervised features from the models and using them together with SVM-hier. In this case, the best features were selected separately for both sub-problems, based on the results of SVM-1 and SVM-2 in the first part of the experiments. Secondly, by fine-tuning the models in the two binary sub-problems and combining them into a hierarchical multi-class classifier. In the latter case, the fine-tuning effectively replaces the manual selection of the best features, as the utility of the final embedding layer was automatically maximized.

It is worth pointing out that the number of trainable classifier parameters differs between the multi-class classification systems. This is because 3 SVMs were included in SVM-OvO and SVM-OvR, but only 2 SVMs were included in SVM-hier. Also, the self-supervised models were trained on SVD data only in the final experiment where all parameters of the self-supervised models were fine-tuned in the two binary sub-problems.

IV. RESULTS

This section describes the results of our experiments. First, the results regarding the self-supervised and MFCC-glottal features are discussed. It is followed by a discussion of the results of hierarchical classification.

A. MFCC-GLOTTAL AND SELF-SUPERVISED FEATURES

The obtained classification accuracies for all the features are shown in Fig. 4. They include evaluations in the two binary sub-problems of SVM-hier: healthy vs. disordered (SVM-1), and hyperfunctional dysphonia vs. vocal fold paresis (SVM-2). The other performance metrics than classification accuracies are shown in Table 1. As can be seen, the self-supervised features outperformed the baseline MFCCs consistently. Moreover, the MFCC-glottal features outperformed the baseline MFCCs in almost all scenarios. The best accuracies for male speakers were 75.65% for SVM-1 and 71.95% for SVM-2, and they were obtained using the wav2vec-LARGE-6 and wav2vec-BASE-0 features, respectively. The MFCC-glottal accuracies were 74.48% for SVM-1 and 69.05% for SVM-2, and the baseline MFCC accuracies were 72.01% for SVM-1 and 61.60% for SVM-2.

The best accuracies for female speakers were 74.50% for SVM-1 and 63.06% for SVM-2, and they were obtained using the HuBERT-0 and wav2vec-LARGE-13 features, respectively. The MFCC-glottal accuracies were 66.13% for SVM-1 and 59.96% for SVM-2, and the baseline accuracies were 68.15% for SVM-1 and 57.09% for SVM-2.

As the wav2vec-LARGE features were generally the best self-supervised features, they were used in the next set of experiments with the SVM-hier classifier. All performance

TABLE 1 Performance Metrics Obtained in Binary Tasks of Healthy Vs. Disordered (SVM-1) and Hyperfunctional Dysphonia Vs. Vocal Fold Paresis (SVM-2). PREC Represents Precision, REC Represents Recall, and F1 Represents F1 Score. The Numbers 0, and 1 in the Metric Names Represent the Classes, Which are Healthy (0) and Disordered (1) for SVM-1, and Hyperfunctional Dysphonia (0), and Vocal Fold Paresis (1) for SVM-2. The Mean Values Over the Folds are Reported for All Metrics. In Addition, the Standard Deviations are Reported for Accuracy, and the Best Mean Accuracy Values are Highlighted for Each Classifier and Gender. Results of All the Self-Supervised Features are Not Included, Only the Layers With the Highest Performance are Included (See Fig. 4). In the Feature Column, the Feature Names Include Their Corresponding Layer Numbers for Self-Supervised Features

Gender	Classifier	Feature	Accuracy [%]	PREC_0	REC_0	F1_0	PREC_1	REC_1	F1_1
Male	SVM-1	wav2vec-LARGE-6	**75.65 ± 5.81**	0.91	0.82	0.87	0.50	0.69	0.58
		wav2vec-BASE-0	70.92 ± 6.81	0.89	0.83	0.86	0.46	0.60	0.52
		HuBERT-LARGE-11	72.14 ± 7.93	0.89	0.85	0.87	0.50	0.59	0.54
		MFCC-glottal	74.48 ± 5.85	0.90	0.84	0.87	0.51	0.64	0.57
		MFCC	72.01 ± 7.75	0.89	0.88	0.88	0.54	0.56	0.55
	SVM-2	wav2vec-LARGE-14	69.97 ± 11.89	0.71	0.84	0.77	0.74	0.56	0.64
		wav2vec-BASE-0	**71.95 ± 12.62**	0.74	0.75	0.74	0.67	0.66	0.66
		HuBERT-LARGE-1	71.88 ± 10.56	0.73	0.80	0.76	0.70	0.62	0.66
		MFCC-glottal	69.05 ± 9.67	0.73	0.74	0.73	0.66	0.64	0.65
		MFCC	61.60 ± 8.86	0.65	0.76	0.70	0.60	0.47	0.53
Female	SVM-1	wav2vec-LARGE-3	73.80 ± 5.03	0.84	0.77	0.80	0.60	0.71	0.65
		wav2vec-BASE-0	73.33 ± 4.13	0.85	0.75	0.79	0.58	0.72	0.64
		HuBERT-LARGE-0	**74.50 ± 4.38**	0.85	0.76	0.81	0.60	0.72	0.65
		MFCC-glottal	66.13 ± 3.11	0.80	0.66	0.72	0.49	0.66	0.56
		MFCC	68.15 ± 4.59	0.81	0.68	0.74	0.51	0.68	0.58
	SVM-2	wav2vec-LARGE-13	**63.06 ± 6.77**	0.74	0.83	0.78	0.57	0.44	0.50
		wav2vec-BASE-3	61.47 ± 4.65	0.72	0.92	0.81	0.68	0.31	0.43
		HuBERT-LARGE-7	61.31 ± 5.94	0.73	0.78	0.75	0.52	0.45	0.48
		MFCC-glottal	59.96 ± 7.91	0.72	0.81	0.76	0.52	0.40	0.45
		MFCC	57.09 ± 7.48	0.71	0.74	0.72	0.45	0.42	0.43

TABLE 2 Performance Metrics for the Multi-Class Classifiers. PREC Represents Precision, REC Represents Recall, and F1 Represents F1 Score. Numbers 0, 1, and 2 in the Metric Names Represent Healthy Voice, Hyperfunctional Dysphonia, and Vocal Fold Paresis, Respectively. The Mean Values Over the Folds are Reported for All Metrics, and the Best Mean Accuracy Values are Highlighted for Each Classifier and Gender. In Addition, the Standard Deviations are Reported for Accuracy

Gender	Classifier	Feature	Accuracy [%]	PREC_0	REC_0	F1_0	PREC_1	REC_1	F1_1	PREC_2	REC_2	F1_2
Male	wav2vec-LARGE-hier	fine-tuned	**62.77 ± 10.94**	0.92	0.79	0.85	0.30	0.55	0.39	0.47	0.53	0.50
	SVM-hier	wav2vec-LARGE	61.29 ± 7.95	0.91	0.83	0.87	0.35	0.58	0.44	0.45	0.43	0.44
		MFCC-glottal	57.53 ± 6.79	0.91	0.84	0.87	0.28	0.38	0.33	0.41	0.48	0.44
		MFCC	53.76 ± 9.35	0.89	0.87	0.88	0.26	0.31	0.29	0.43	0.42	0.42
	SVM-OvR	MFCC	47.01 ± 6.13	0.87	0.84	0.86	0.23	0.30	0.26	0.33	0.28	0.30
	SVM-OvO	MFCC	46.38 ± 6.52	0.86	0.89	0.87	0.30	0.27	0.28	0.31	0.24	0.27
Female	wav2vec-LARGE-hier	fine-tuned	54.12 ± 5.23	0.83	0.70	0.76	0.37	0.44	0.4	0.32	0.49	0.38
	SVM-hier	wav2vec-LARGE	**55.36 ± 4.99**	0.84	0.78	0.81	0.39	0.50	0.43	0.43	0.39	0.41
		MFCC-glottal	49.27 ± 5.80	0.80	0.65	0.72	0.30	0.49	0.37	0.37	0.33	0.35
		MFCC	51.11 ± 7.08	0.82	0.69	0.75	0.34	0.47	0.39	0.31	0.37	0.34
	SVM-OvR	MFCC	47.70 ± 6.33	0.78	0.71	0.75	0.33	0.45	0.38	0.32	0.28	0.30
	SVM-OvO	MFCC	48.41 ± 5.67	0.79	0.72	0.75	0.34	0.47	0.39	0.34	0.27	0.30

metrics of the best layers and their respective baselines are shown in Table 1.

B. HIERARCHICAL CLASSIFICATION

The results of the experiments with hierarchical classifiers are shown in Fig. 5 and Table 2. First, the baseline classifiers, SVM-OvO and SVM-OvR, were trained and evaluated with MFCCs. For male speakers, the baseline classification accuracies were 47.01% for SVM-OvR and 46.38% for SVM-OvO. For female speakers, the baseline classification accuracies were 47.70% for SVM-OvR and 48.41% for SVM-OvO. Then, SVM-hier was trained and evaluated with MFCCs, and the results were better than those of the baselines (i.e., SVM-OvR and SVM-OvO). For male speakers, the accuracy was 53.76%, and for female speakers, the accuracy was 51.11%.

Then, the hierarchical classification was examined together with the self-supervised models. Wav2vec 2.0 LARGE was used as the self-supervised model, because it was generally

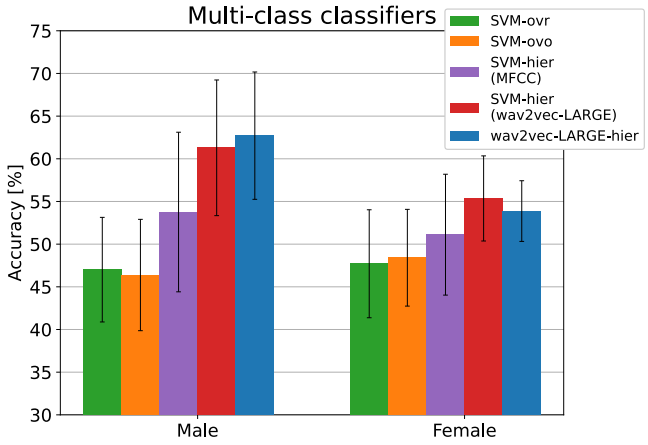

FIGURE 5. Classification accuracies obtained in multi-class classification. Heights of the bars represent the mean accuracies over the folds and the tails represent the standard deviations. Wav2vec-LARGE-hier refers to the scenario where fine-tuned wav2vec 2.0 LARGE model was used in hierarchical classification. For other models, the used features are indicated within parentheses.

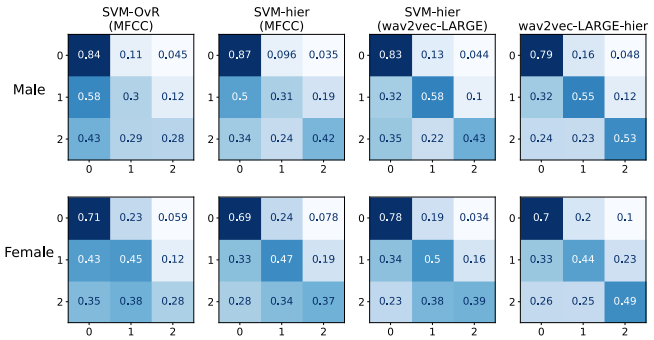

FIGURE 6. Confusion matrices of the multi-class classification systems. The horizontal axis represents the predicted classes, and the vertical axis represents the true classes. Class labels 0, 1, and 2 represent healthy voice, hyperfunctional dysphonia, and vocal fold paresis, respectively. The values are normalized over true values (rows). Wav2vec-LARGE-hier refers to the scenario where fine-tuned wav2vec 2.0 model was used in hierarchical classification. For other models, the used features are indicated within parentheses.

the best self-supervised model in Section IV-A. First, SVM-hier was used and the best wav2vec-LARGE features were selected for both hierarchical steps (SVM-1 and SVM-2) separately, based on their performance in Section IV-A. For female speakers, wav2vec-LARGE-3 was used for SVM-1 and wav2vec-LARGE-13 was used for SVM-2. For male speakers, wav2vec-LARGE-6 was used for SVM-1 and wav2vec-LARGE-14 was used for SVM-2. The resulting multi-class accuracies were 61.29% and 55.36% for male and female speakers, respectively. This was the best obtained performance for female speakers.

Finally, wav2vec 2.0 LARGE was fine-tuned for the two binary sub-problems separately, after which the fine-tuned models were combined into a hierarchical classifier, wav2vec-LARGE-hier. The obtained classification accuracies were 62.77% and 54.12% for male and female speakers, respectively. This was the best obtained performance for male speakers. Therefore, the highest absolute improvements to the baseline SVM systems were 15.76% and 6.95% for male and female speakers, respectively.

The confusion matrices for all the hierarchical systems, as well as for the SVM-OvR baseline are visualized in Fig. 6. The values in the confusion matrices are normalized over the true values (rows). It can be seen that hierarchical classification mainly increases the performance of the two smallest classes. For instance, in comparison to the baseline SVM-OvR with MFCCs, SVM-hier with the wav2vec-LARGE features increased the recall of the smallest class (vocal fold paresis) from 0.28 to 0.43 for male speakers, and from 0.28 to 0.39 for female speakers. Moreover, the recall of hyperfunctional dysphonia increased from 0.30 to 0.58 for male speakers, and from 0.45 to 0.50 for female speakers. In addition, fine-tuning further improved the performance of the smallest class without largely affecting the classification accuracy. In comparison to SVM-hier with wav2vec-LARGE features, wav2vec-LARGE-hier increased the recall of vocal fold paresis by 0.10.

V. CONCLUSION

In this paper, a 3-class voice pathology classification task was studied to automatically classify two laryngeal voice disorders (hyperfunctional dysphonia and vocal fold paresis) and healthy voice. Samples from the Saarbrücken Voice Disorders (SVD) database were used. The study examined three approaches that may alleviate the problem of data scarcity in the multi-class classification of voice disorders and, therefore, improve the classification performance of a pipeline classifier.

For the feature extraction phase, the proposed approaches corresponded to the extraction of the MFCC-glottal features, and the usage of the pre-trained self-supervised models as feature extractors. For the classification phase, a hierarchical classification approach was used. Comparisons were made to commonly used baseline approaches: MFCCs for feature extraction, and SVM-OvO and SVM-OvR for classification.

The two feature-based approaches were first evaluated in the two binary sub-problems of the hierarchical classification framework. The results indicate that both the MFCC-glottal and self-supervised features increase the classification performance in most scenarios, when comparing to the baseline MFCCs. For male speakers, there is no large difference between MFCC-glottal and the best self-supervised features (1.17% for SVM-1 and 2.90% for SVM-2), which may imply that they are equally effective methods to capture the glottal information that is discriminative between the classes. However, the glottal MFCCs performed consistently well for male speakers but not for female speakers. In fact, the glottal MFCCs were outperformed by the baseline MFCCs for SVM-1 with female speakers, with an absolute difference of 2.02%. This difference between the genders might be caused by the fact that the glottal source extraction by inverse filtering is more difficult for high-pitch speech, because less samples are

available for the estimation of vocal tract filter for each glottal cycle.

In general, the positive effect of the feature-based approaches is largest in the task of classification between the two pathologies, which is also the task which typically suffers most severely from the data scarcity problem. This finding supports our hypothesis that these approaches effectively alleviate the problems caused by scarcity of training data.

The implications are similar, when evaluating the hierarchical systems in the multi-class problem. In comparison to the baseline OvO and OvR approaches, all hierarchical systems increase the classification accuracy. The confusion matrices show that the performance improvements are almost completely caused by an improvement in the two smallest classes (hyperfunctional dysphonia and vocal fold paresis).

For both genders, the best performance was achieved by using self-supervised models together with hierarchical classification. For female speakers, the best classification accuracy (55.36%) was achieved by using the non-fine-tuned wav2vec-LARGE features, whereas for male speakers, the best accuracy (62.77%) resulted from using the fine-tuned wav2vec-LARGE models. Therefore, the total improvements to the baseline multi-class classifiers were 15.76% and 6.95% (absolute) for males and females, respectively.

Overall, the performance difference between the fine-tuned and non-fine-tuned self-supervised models was not large (1.24% for female speakers and 1.48% for male speakers). However, fine-tuning effectively balanced the performance differences between the unbalanced classes. In particular, fine-tuning resulted in an absolute improvement of 0.10 to the recall of the smallest class (vocal fold paresis), while keeping the balanced classification accuracy almost unchanged. This effect was similar for both genders, and it may indicate that fine-tuning further improves the system performance in small-data scenarios, by balancing the performance differences between the classes of different sizes.

The obtained performance metrics in the binary detection between healthy and disordered speech are generally comparable with existing studies that have used the SVD database. For example, the best classification accuracies with recordings of the vowel /a/ in normal pitch were 75.42%, 74.32 %, 67.0% in [5], [8], and [4], respectively. Some works exist that report very high accuracies. For example, the reported classification accuracies were 96.96 % in [3] and 96.5 % in [2]. However, those studies did not use class balancing or balanced classification accuracy metric, which can result overly optimistic performance values due to over-emphasizing the largest healthy class. In this study, the classes were balanced in both training and evaluation data. There is evidence showing that the performance with the SVD data can be largely dependent on the selected experimental setup [1].

The multi-class classification results of this study are not directly comparable to any existing studies, because of the differences in the used databases and included disorders. The work in [11] utilized the SVD database in multi-class classification between healthy, reflux laryngitis, hyperfunctional dysphonia and hypofunctional dysphonia, and achieved true negative and true positive rates of 92.2% and 88.9%, respectively. In [28], SVD was used in a multi-modal classification in several different multi-class classification problems, and the obtained classification accuracy in the 3-class classification was 94.3%. In [12], a 3-class classification was carried out between neurotypical speech, dysarthria and apraxia of speech, and the best balanced classification accuracy of 79.7% is reported. Similar to our work, the best performance was obtained by using a hierarchical SVM classifier.

Finally, it should be noted that this study uses the non-nested cross-validation (CV) instead of nested CV, which might result in overfitting, as discussed in [46]. The usage of nested CV has not usually been reported in studies related to detection and classification of voice disorders.

REFERENCES

[1] M. Huckvale and C. Buciuleac, "Automated detection of voice disorder in the saarbrüken voice database: Effects of pathology subset and audio materials," in *Proc. Interspeech*, 2021, pp. 1399–1403.

[2] F. Amara, M. Fezari, and H. Bourouba, "An improved GMM-SVM system based on distance metric for voice pathology detection," *Appl. Math*, vol. 10, no. 3, pp. 1061–1070, 2016.

[3] J. Y. Lee, "A two-stage approach using Gaussian mixture models and higher-order statistics for a classification of normal and pathological voices," *EURASIP J. Adv. Signal Process.*, vol. 1, pp. 1–8, 2012.

[4] D. Martínez, E. Lleida, A. Ortega, A. Miguel, and J. Villalba, "Voice pathology detection on the Saarbrüken voice database with calibration and fusion of scores using multifocal toolkit," in *Proc. Adv. Speech Lang. Technol. Iberian Lang.*, 2012, pp. 99–109.

[5] J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. Part II: Review of speaker recognition techniques and study on the effects of different variability factors," *Biomed. Signal Process. Control*, vol. 48, pp. 128–143, 2019.

[6] 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. IEEE Int. Conf. Workshop Bioinspired Intell.*, 2017, pp. 1–4.

[7] M. K. Reddy and P. Alku, "A comparison of cepstral features in the detection of pathological voices by varying the input and filterbank of the cepstrum computation," *IEEE Access*, vol. 9, pp. 135953–135963, 2021.

[8] S. R. Kadiri and P. Alku, "Analysis and detection of pathological voice using glottal source features," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 367–379, Feb. 2019.

[9] E. Vaiciukynas, A. Verikas, A. Gelzinis, M. Bacauskiene, and V. Uloza, "Exploring similarity-based classification of larynx disorders from human voice," *Speech Commun.*, vol. 54, no. 5, pp. 601–610, 2012.

[10] R. Behroozmand and F. Almasganj, "Comparison of neural networks and support vector machines applied to optimized features extracted from patients' speech signal for classification of vocal fold inflammation," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol.*, 2005, pp. 844–849.

[11] K. T. Chui, M. D. Lytras, and P. Vasant, "Combined generative adversarial network and fuzzy C-means clustering for multi-class voice disorder detection with an imbalanced dataset," *Appl. Sci.*, vol. 10, no. 13, 2020, Art. no. 4571.

[12] I. Kodrasi, M. Pernon, M. Laganaro, and H. Bourlard, "Automatic and perceptual discrimination between dysarthria, apraxia of speech, and neurotypical speech," in *Proc. IEEE Int. Conf. Acoust. Speech and Signal Process.*, 2021, pp. 7308–7312.

[13] A. A. Dibazar, T. W. Berger, and S. S. Narayanan, "Pathological voice assessment," in *Proc. IEEE Int. Conf. Eng. Med. Biol. Soc.*, 2006, pp. 1669–1673.

[14] H. Wu, J. Soraghan, A. Lowit, and G. Di-Caterina, "A deep learning method for pathological voice detection using convolutional deep belief networks," in *Proc. Interspeech*, 2018, pp. 446–450.

[15] J. C. Vásquez-Correa, J. Fritsch, J. R. Orozco-Arroyave, E. Nöth, and M. Magimai-Doss, "On modeling glottal source information for phonation assessment in Parkinson's disease," in *Proc. Interspeech*, 2021, pp. 26–30.
[16] Z. Jin et al., "Adversarial data augmentation for disordered speech recognition," in *Proc. Interspeech*, 2021, pp. 4803–4807.
[17] P. Barche, K. Gurugubelli, and A. K. Vuppala, "Towards automatic assessment of voice disorders: A clinical approach," in *Proc. Interspeech*, 2020, pp. 2537–2541.
[18] A. Baevski and H. Zhou, "Abdelrahman mohamed, and michael auli, wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 12449–12460.
[19] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.
[20] A. Hernandez, P. A. Pérez-Toro, E. Noeth, J. R. Orozco-Arroyave, A. Maier, and S. H. Yang, "Cross-lingual self-supervised speech representations for improved dysarthric speech recognition," in *Proc. Interspeech*, 2022, pp. 51–55.
[21] P. Lester, W. C. Violeta Huang, and T. Toda, "Investigating self-supervised pretraining frameworks for pathological speech recognition," in *Proc. Interspeech*, 2022, pp. 41–45.
[22] G. Chatzoudis, M. Plitsis, S. Stamouli, A.-L. Dimou, N. Katsamanis, and V. Katsouros, "Zero-shot cross-lingual aphasia detection using automatic speech recognition," in *Proc. Interspeech*, 2022, pp. 2178–2182.
[23] S. P. Bayerl, D. Wagner, E. Noeth, and K. Riedhammer, "Detecting dysfluencies in stuttering therapy using wav2vec 2.0," in *Proc. Interspeech*, 2022, pp. 2868–2872.
[24] Y. Getman et al., "Wav2vec2-based speech rating system for children with speech sound disorder," in *Proc. Interspeech*, 2022, pp. 3618–3622.
[25] Y. Zhu, X. Liang, J. A. Batsis, and R. M. Roth, "Domain-aware intermediate pretraining for dementia detection with limited data," in *Proc. Interspeech*, 2022, pp. 2183–2187.
[26] T. Wang et al., "Conformer based elderly speech recognition system for Alzheimer's disease detection," in *Proc. Interspeech*, 2022, pp. 4825–4829.
[27] D. Priyasad et al., "Detecting heart failure through voice analysis using self-supervised mode-based memory fusion," in *Proc.23rd Interspeech Conf.*, 2022, pp. 2848–2852.
[28] S. Bhattacharjee and W. Xu, "VoiceLens: A multi-view multi-class disease classification model through daily-life speech data," *Smart Health*, vol. 23, 2022, Art. no. 100233.
[29] J. Mallela et al., "Voice based classification of patients with Amyotrophic Lateral Sclerosis, Parkinson's Disease and healthy controls with CNN-LSTM using transfer learning," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 6784–6788.
[30] H. Cordeiro, J. Fonseca, I. Guimarães, and C. Meneses, "Hierarchical classification and system combination for automatically identifying physiological and neuromuscular laryngeal pathologies," *J. Voice*, vol. 31, no. 3, pp. 384.e9–384.e14, 2017.
[31] M. Nikkhah-Bahrami, H. Ahmadi-Noubari, B. S. Aghazadeh, and H. K. Heris, "Hierarchical diagnosis of vocal fold disorders," in *Advances Computer Science and Engineering*, H. Sarbazi-Azad, B. Parhami, S.-G. Miremadi, and S. Hessabi, Eds., Berlin, Germany: Springer, 2009, pp. 897–900.
[32] 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.
[33] M. Airaksinen, L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "A comparison between STRAIGHT, glottal, and sinusoidal vocoding in statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1658–1670, Sep. 2018.
[34] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," in *Proc. Interspeech*, 2020, pp. 2757–2761.
[35] Q. Xu, A. Baevski, and M. Auli, "Simple and effective zero-shot cross-lingual phoneme recognition," in *Proc. Interspeech*, 2022, pp. 2113–2117.
[36] R. Ardila et al., "Common voice: A massively-multilingual speech corpus," in *Proc. Int. Conf. Lang. Resour. Eval.*, 2019.
[37] J. F. Mark, K. M. Gales, A. K. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low-resource languages: Babel project research at cued," in *Proc. 4th Int. Workshop Spoken Lang. Technol. Under-Resourced Lang.*, 2014, pp. 16–23.
[38] J. Kahn et al., "Libri-light: A. benchmark for ASR with limited or no supervision," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 7669–7673.
[39] M. Pützer and W. J. Barry, "Saarbrüken voice database, institute of phonetics, univ. of saarland," 2007. [Online]. Available: http://www.stimmdatenbank.coli.uni-saarland.de/ (Last viewed Feb. 18, 2023).
[40] M. Pützer and W. J. Barry, "Instrumental dimensioning of normal and pathological phonation using acoustic measurements," *Clin. Linguistics Phonetics*, vol. 22, no. 6, pp. 407–420, 2008.
[41] R. E. Hillman, C. E. Stepp, J. H. V. Stan, M. Zañartu, and D. D. Mehta, "An updated theoretical framework for vocal hyperfunction," *Amer. J. Speech Lang. Pathol.*, vol. 29, no. 4, pp. 2254–2260, 2020.
[42] R. Behroozmand and F. Almasganj, "Optimal selection of wavelet-packet-based features using genetic algorithm in pathological assessment of patients' speech signal with unilateral vocal fold paralysis," *Comput. Biol. Med.*, vol. 37, no. 4, pp. 474–485, 2007.
[43] C. Walton, E. Conway, H. Blackshaw, and P. Carding, "Unilateral vocal fold paralysis: A systematic review of speech-language pathology management," *J. Voice*, vol. 31, no. 4, pp. 509–e7, 2017.
[44] S. Tirronen, "Detection and multi-class classification of voice disorders from speech recordings," Master's thesis, School of Science, Aalto University, Espoo, Finland, 2022.
[45] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2019.
[46] C. G. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, 2010.