IEEE Journal of Translational Engineering in Health and Medicine

Received 24 August 2023; revised 30 November 2023; accepted 4 December 2023.
Date of publication 7 December 2023; date of current version 26 December 2023.

Digital Object Identifier 10.1109/JTEHM.2023.3340345

Multitask and Transfer Learning Approach for Joint Classification and Severity Estimation of Dysphonia

DOSTI AZIZ AND SZTAHÓ DÁVID

Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, 1117 Budapest, Hungary

CORRESPONDING AUTHOR: D. AZIZ (azizd@edu.bme.hu)

This work was supported in part by the National Research, Development and Innovation Fund of Hungary, under the K_22 Funding Scheme under Grant K143075.

ABSTRACT **Objective:** Despite speech being the primary communication medium, it carries valuable information about a speaker's health, emotions, and identity. Various conditions can affect the vocal organs, leading to speech difficulties. Extensive research has been conducted by voice clinicians and academia in speech analysis. Previous approaches primarily focused on one particular task, such as differentiating between normal and dysphonic speech, classifying different voice disorders, or estimating the severity of voice disorders. **Methods and procedures:** This study proposes an approach that combines transfer learning and multitask learning (MTL) to simultaneously perform dysphonia classification and severity estimation. Both tasks use a shared representation; network is learned from these shared features. We employed five computer vision models and changed their architecture to support multitask learning. Additionally, we conducted binary 'healthy vs. dysphonia' and multiclass 'healthy vs. organic and functional dysphonia' classification using multitask learning, with the speaker's sex as an auxiliary task. **Results:** The proposed method achieved improved performance across all classification metrics compared to single-task learning (STL), which only performs classification or severity estimation. Specifically, the model achieved F1 scores of 93% and 90% in MTL and STL, respectively. Moreover, we observed considerable improvements in both classification tasks by evaluating beta values associated with the weight assigned to the sex-predicting auxiliary task. MTL achieved an accuracy of 77% compared to the STL score of 73.2%. However, the performance of severity estimation in MTL was comparable to STL. **Conclusion:** Our goal is to improve how voice pathologists and clinicians understand patients' conditions, make it easier to track their progress, and enhance the monitoring of vocal quality and treatment procedures.

INDEX TERMS Multitask learning, dysphonia, voice pathology, deep learning, speech.

Clinical and Translational Impact Statement: By integrating both classification and severity estimation of dysphonia using multitask learning, we aim to enable clinicians to gain a better understanding of the patient's situation, effectively monitor their progress and voice quality.

I. INTRODUCTION

SPEECH has been recognized as a primary factor in human interaction. It plays a crucial role in socialization and overall well-being by enabling individuals to communicate and share ideas. Human speech can reveal necessary information about an individual's identity and health status [1]. Several psychiatric and neurogenerative conditions can impact the organs responsible for speech production, causing individuals to struggle with producing normal speech [2], [3], [4], [5]. Considering these factors, speech analysis has the potential to become a valuable clinical tool for diagnosing and monitoring a wide range of medical conditions such as Alzheimer's [4], Parkinson's [6], Depression [7] and dysphonia. By analyzing speech patterns, researchers and healthcare professionals can gain valuable insights into an individual's cognitive, motor, and emotional processes, which can provide practical diagnostic information and improve treatment procedures. Dysphonia, a condition characterized by impaired voice production, affects almost a third of the population at some

point in life. Although dysphonia is often used interchangeably with hoarseness, hoarseness is a symptom of altered voice quality reported by patients, while dysphonia is a diagnosis made by clinicians [8]. Dysphonia can affect people of all ages and genders, but it is more common in teachers [9], older adults, and individuals with significant vocal demands. It can be caused by benign or self-limited conditions but may also indicate a more severe or progressive condition that requires prompt management [10].

Dysphonia has been categorized into two groups: organic and functional dysphonia. Functional dysphonia refers to a voice condition that does not correlate with neurogenerative or anatomical factors. It originates from abnormal laryngeal function or vocal misuse or overuse, which can lead to inefficient oral communication. Any stage of voice production can be affected by this condition [11]. Organic dysphonia can be further divided into structural and neurogenic categories. Neurogenic dysphonia is caused by abnormal control, coordination, or strength of the vocal folds due to neurological diseases such as Parkinson's. Structural organic dysphonia is caused by morphological changes such as vocal cord nodules, polyps, Gastroesophageal Reflux (GERD), and vocal cord paralysis [10].

While the term functional and organic dysphonia exists in the literature, there are some contradictory arguments if they are really distinguishable. Speech-language pathology has historically used the term ''functional'' to describe voice disorders linked to psychogenic causes or personality variables. Initially, functional and psychogenic voice disorders were seen as closely related. At the same time, this has been argued by [12] as their findings indicate no significant differences in personality traits or psychological distress between individuals with organic and functional dysphonia. Suggested common causes for functional dysphonia encompass psychoneuroses, personality disorders, and faulty voice habits. These disorders were believed to occur in individuals with normal laryngeal anatomy and physiology, with stress, musculoskeletal tension, and conflicts related to sex identification as specific contributing factors [13]. Others argue that the term ''functional'' dysphonia is inadequate in terms of etiology and merely implies a ''non-organic'' disorder identified through exclusion [14]. In their scoping review, authors critically analyzed the diverse terminologies employed in classifying voice disorders, highlighting disparities among professionals and researchers. It underscores the need for a standardized classification system to facilitate effective communication among clinicians and provide appropriate treatment guidance. The study suggests categorizing hyperfunctional muscle issues as ''Muscle Tension Dysphonia,'' psychosocially-based voice disorders as ''Functional (Psychogenic) Voice Disorder,'' and disorders stemming from organic or neurogenerative causes as ''Organic Dysphonia'' [15]. The validity of the distinction between individuals with functional and organic dysphonia has yet to be thoroughly examined and tested. Therefore, it is crucial to approach this differentiation with caution, and further research is needed to establish an encompassing and standardized classification framework for voice disorders.

The conventional diagnostic approach for dysphonia involves multiple visits to healthcare professionals and voice therapists, utilizing techniques like laryngoscopy, stroboscopy, laryngeal electromyography, and auditory analysis. However, these methods often cause patient discomfort and distress [16]. Additionally, they are time-consuming, subject to variability due to their subjective nature, and expensive, imposing a significant financial burden ranging from 577 to 953 US dollars per patient per year [10].

Considering these factors and recent advancements in machine learning (ML) qualified the development of objective and reliable methods for diagnosing dysphonia. ML algorithms can detect and classify dysphonia accurately by analyzing speech signals, providing valuable diagnostic information for healthcare professionals. Speech samples were recorded from patients with voice disorders and control groups. Various acoustic features can be extracted from these speech samples, such as jitter, shimmer, harmonic-to-noise ratio (HNR), fundamental frequency, and Mel-Frequency cepstral coefficients (MFCCs) [17], [18]. Features extracted from utterances are used to train ML algorithms to classify healthy and voice-disordered persons. In [19], MFCC and Linear prediction cepstral coefficients (LPC) were extracted from sustained $/a/$ vowels and used to train Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM), achieving accuracy of 94.44%, 95.74% with HMM and GMM, respectively. A combination of VGG16 and a Support Vector Machine was proposed. Features extracted using VGG16 from sustained $/i/$ vowel used to train SVM classifier and achieved accuracy 96.7% [20]. In some other works focusing on sustained vowels, performance analysis of different ML algorithms has been conducted [21], [22]. Other paper has explored the effectiveness of algorithms such as K-nearest neighbors (KNN), SVM, and random forest [23]. Furthermore, the utilization of XGBoost, isolation forest, and DenseNet has also been investigated [24]. The [25] study predicted dysphonic speech severity from sustained vowels' time and spectral features using step-wise multiple regression, yielding mean R of 0.880 and mean R^2 of 0.775. Continuous speech samples were used for automated dysphonia severity assessment, achieving 89% accuracy and a root mean square error (RMSE) of 0.49 for binary classification and severity estimation, respectively [26]

Deep learning techniques, like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNNs), have gained traction in analyzing pathological speech due to their impressive performance in diverse domains such as image recognition [27], natural language processing [28] and speech recognition [29]. Several studies have been conducted in the research area of voice disorder. One study achieved an accuracy of 98.9% using CNN with oversampling techniques

IEEE Journal of Translational Engineering in Health and Medicine

to deal with class imbalance [30]. CNN with chromagram feature set in [31], CNN combined with RNN in [32]. In [33], a Convolutional Deep Belief Network (CDBN) was used to pre-train the weight of the CNN model, and an accuracy of 71% was achieved. In [34], a performance comparison between CNN and RNN was conducted. The study used 10-fold cross-validation and accuracies of 87.11% and 86.52% for CNN and RNN, respectively. A combination of features from EGG and speech from sustained vowel /*a*/ has been considered for distinguishing normal and pathological voices [35], [36]. Other studies considered stacked autoencoder [37] and LSTM-based autoencoder [38] using sustained vowel /*a*/ and continuous speech samples, respectively.

Limited research exists on distinguishing functional from organic dysphonia. Only [39] has performed classification between these two categories using handcrafted feature extraction and SVM. Classifying between organic and functional dysphonia is essential since the treatment procedures may differ. Benign Organic dysphonia generally may require medical interventions or surgical procedures except for vocal fold nodule, which includes voice therapy as an alternative or adjusted therapy before and after surgery. The treatment procedures for dysphonia from neurodegenerative include voice therapy, laryngeal injection with a variety of fillers, laryngeal framework surgery, and laryngeal re-innervation. Functional dysphonia may be managed through voice therapy, psychological interventions, or a collaboration [40]. Accurate classification ensures that patients receive appropriate and targeted treatment approaches.

Although previous studies have offered valuable insights into voice disorder detection and assessment, they have focused mainly on one task, such as binary, multiclass classification, or the prediction of severity scores. Additionally, most of the research has used sustained vowels, a method that might not fully capture the complexities of natural speech patterns. Furthermore, limited attention has been given to distinguishing between organic and functional dysphonia.

This paper introduces a novel approach to dysphonia evaluation by utilizing transfer and multitasking learning to distinguish between normal and dysphonia and predict severity scores simultaneously. Additionally, we performed multiclass classification between functional, organic dysphonia, and normal speech using continuous speech samples. The similarity in the effects of both conditions on speech samples adds complexity to the classification process. To address this challenge, we implemented multitask learning by incorporating the speaker's sex as an auxiliary task. This approach aims to enhance the model's classification accuracy and its overall ability to generalize. To the best of our knowledge, no previous attempts have been made to conduct both classification and severity estimation of dysphonia patients simultaneously using multitask learning. Both tasks are considered equally important as they provide valuable insights into the patient's condition and allow clinicians to gain comprehensive insights into the patient's condition and monitor their progress effectively. Our approach adopted continuous speech analysis as it faithfully captures natural speech patterns, closely corresponding to natural communication interactions in everyday scenarios. This method aligns more closely with real-life scenarios compared to analyzing isolated speech fragments, enhancing the realism and practical applicability of our findings.

II. MATERIALS AND METHODS

Five different computer vision models were used in this research. The architectures of these models were changed to support multitask learning, as discussed in section II-C. The fine-tuning was performed using 5-fold cross-validation, with a learning rate 0.0001, batch size of 4, Adam optimizer, and 50 epochs. Implementation was carried out using PyTorch [41] framework. Pre-trained versions of these networks are available in torchvision module. In each training and validation fold, two models were saved, final and best models, respectively. The final model is the fine-tuned model after 50 epochs, and the best model is the result of early stopping. We saved the best model with the lowest validation loss. The saved models(Final and Best) were independently tested on the test set.

A. DATASET

We used Hungarian speech samples from two categories of dysphonia, functional and organic, along with healthy control. All speakers were native Hungarian speakers and read a short passage titled ''The North Wind and The Sun,'' an English text translated into many languages, including Hungarian, frequently used by phoneticians. Speech samples were collected from patients who had agreed to participate in the study. The recordings were performed at the Head and Neck Surgery department of the National Institute of Oncology during the consultation. Distinguishing functional from organic dysphonia has been performed by a specialist at the same department. During the recording, various health conditions were observed, including functional dysphonia, recurrent paresis, tumors in different parts of the vocal tract, gastroesophageal reflux disease, chronic inflammation of the larynx, bulbar paresis, amyotrophic lateral sclerosis, leucoplakia, spasmodic dysphonia, and more. The speech recording was performed in a quiet (clinical office) environment with PCM audio coding, 16 kHz sampling rate, and 16-bit quantization. In total, 441 speech samples were used, including 179 healthy control samples, 179 organic dysphonia samples, and 83 functional dysphonia samples. A panel of three specialists graded the severity scores. The initial specialist directly interacted with the patients during the grading process. The other two experts evaluated the speech recordings while listening to the recording in a quiet environment. The final severity score used in this study is a rounded average score of three raters. The RBH scale, which stands for Roughness, Breathiness, and Hoarseness [42], [43], was used to determine the severity level. The scale ranges from 0 to 3; for the speech samples in this study, H was selected as the severity level, ranging from 0 to 3

TABLE 1. Sex and severity distribution of each class in the dataset.

Class	Sex	Severity Level			
		0	1	2	3
Healthy	Male	86	-	-	-
HC	Female	93	-	-	-
Oragnic	Male	-	24	24	33
OD	Female	-	34	39	25
Functional	Male	-	11	8	2
FD	Female	-	44	13	5

(0 indicates no hoarseness, ''Healthy speaker,'' and 3 means ''Severe hoarseness''). The class and severity distribution by sex is presented in Table 1.

The dataset is divided into 80% training set and a 20% test set. The training set was split into training and validation sets and used in 5-fold cross-validation using the scikit-learn library [44]. One fold was used for validation, while the remaining four were used for training the models.

Two manual seeds have been used to obtain a realistic view of our method's performance. The manual seeds were used to obtain the representative samples in the test set according to the severity level of the speech and gender of the speakers. One seed is used for joint classification and regression tasks, while the other is used for binary and multiclass classification with the speaker's sex as an auxiliary task.

B. DEEP LEARNING MODELS

Computer vision models used for the experiments are ResNet50, DenseNet, MobileNet, ConvNeXt, and EfficientNet. These models were initially used for image classification and trained on the Imagenet dataset. In this section, we will briefly explain them.

1) ResNet

Deep neural networks have benefited from weight initialization and batch normalization, which have mitigated issues like vanishing and exploding gradients. However, they are still susceptible to the degradation problem, where network accuracy plateaus or decreases as the network depth increases. To address this challenge, researchers from Microsoft proposed ResNet [45], a residual network architecture. ResNet has become a prominent solution to the degradation problem due to its unique design of skip connections, allowing for smoother gradient flow and enabling the training of much deeper networks. Figure 1 illustrates the ResNet block's architecture.

ResNet employs skip connections to enable more efficient learning of identity mapping within the neural network. As shown in Figure 1, the feature x is summed with the output of the last layer and passed through a nonlinear function, typically ReLU, establishing a direct information flow through the network. This technique has been pivotal in training large-scale neural networks and has gained widespread

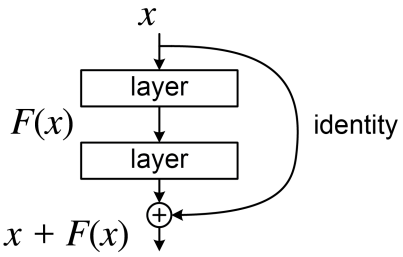

FIGURE 1. Residual block.

adoption by researchers when constructing substantial network architectures. In this experiment, ResNet50 was used, which consists of 50 layers with a skip connection between every two layers [45].

2) DenseNet

The Dense Convolutional Network (DenseNet) is built upon the residual network concept, where more connections are added between layers in the network. In DenseNet, the layers are connected using dense connectivity, which means each layer's output is concatenated with the inputs of all subsequent layers. This creates a dense connectivity pattern between the layers, allowing the network to reuse features from earlier layers and retain more information throughout the network. The DenseNet-B and DenseNet-BC architectures have demonstrated superior performance compared to other deep learning architectures because of the use of bottleneck layers between dense blocks on various image classification benchmarks. Furthermore, DenseNets use fewer parameters because of the bottleneck layers between dense blocks, making them advantageous for tasks with limited computational resources or where model size is a constraint. [46].

3) MobileNet

The aim was to design an architecture for efficient processing on mobile devices and embedded systems with limited computational resources. The authors introduce depthwise separable convolutions as a building block for the network as an alternative to the conventional convolution layer. In a depthwise convolution, a single convolutional filter is applied to each input channel to achieve a lightweight filtering process. After that, pointwise 1×1 convolution is applied to compute linear combinations of the input channels, creating new features. The resulting network has fewer parameters and requires less computation than traditional convolutional networks while still performing well on various tasks [47].

4) ConvNeXt

Vision transformers have become popular in various computer vision applications since they were first introduced in 2020. However, a group of researchers from Facebook aimed to demonstrate the continued relevance of convolutional networks by modernizing the ResNet50 architecture

toward the vision transformer's design. This involved implementing changes such as depthwise convolutions, replacing Relu with Gelu activation functions, and other changes. ConvNeXt, a family of pure ConvNeXt models constructed, have been competing with Transformers regarding accuracy and scalability [48].

5) EfficientNet

The author of this model proposed a new scaling method for designing neural network architectures. The approach involves scaling the network's depth, width, and resolution using compound scaling based on the intuition that larger input images require more layers and channels to capture detailed information. As a result, they developed a family of eight models called EfficientNetB0 to B7 that outperformed popular neural network architectures like ResNet, Inception, and MobileNet in terms of accuracy and efficiency, even with fewer parameters [49].

In fact, deep learning architectures are data-hungry and require large amounts of data and significant computation power, making it difficult for individuals or small organizations to access these resources. Transfer learning plays a crucial role in scenarios where the database size is insufficient for training deep learning (DL) from scratch. Transfer learning is a technique in which the model's parameter learned in a specific domain is transferred to another domain and fine-tuned on smaller datasets. The idea is that the pre-trained models have already learned useful features and representations that can be generalized to the new task, even if the new task's dataset is relatively small or different. Considering the limited size of our dataset, we adopted transfer learning. Different ways exist when using transfer learning, depending on datasets and performed tasks. The first scenario uses these models as a feature extractor by freezing all the model's layers. Features can be obtained before the last classifier layer. These high-dimensional features can be used to train neural networks or any other ML algorithm. Unfreezing and fine-tuning some layers of the network is another approach. Lastly is fine-tuning all neural network layers, which takes more time and computation but usually leads to better performance. Considering our domain is speech and quite different from the image in which these models have been trained, we considered the last option.

C. MULTITASKING LEARNING

Multitask Learning (MTL) was first introduced by Caruana as a learning paradigm in ML that enables learning multiple related tasks jointly to improve the generalization performance of all the tasks. The motivation for MTL was to alleviate the data sparsity problem, where each task has a limited number of labeled data. MTL aggregates the labeled data in all the tasks to obtain a more accurate learner for each task, which can help reuse existing knowledge for different tasks [50]. Since then, many researchers adopted this methodology in many areas, including computer vision [51], Natural Language Processing [52], and speech processing [53]. Researchers have found that MTL perform better than their STL counterparts. This can be attributed to MTL's ability to leverage a larger amount of data from diverse learning tasks, facilitating enhanced knowledge sharing among tasks, improved performance for each task, and mitigating risks of overfitting. The most straightforward architecture of MTL is a shared representation. In this case, all tasks are trained in parallel with the same representation feature; each task has a different output layer corresponding to the task. The loss function of the MTL network is the combination of loss functions of the tasks performed by the model. Eq. (1) shows the loss function of our multitask learning network.

$$L_{combined} = \beta_1 L_{t_1} + \beta_2 L_{t_2} \tag{1}$$

L_{t_1} and L_{t_2} are loss functions corresponding to tasks in a multitasking network. Furthermore, β_1 and β_2 are the corresponding hyperparameters that control each task's importance in the network.

We chose cross-entropy loss function for L_{t_1} for all three experiments we conducted. For the experiment of joint classification and regression of dysphonia speech, we have selected mean squared error as loss function L_{t_2}. In contrast, for binary and multiclass classification experiments using sex prediction as an auxiliary task, mean square error loss has been replaced by another cross-entropy loss function.

The choice of hyperparameters β_1 and β_2 depends on the relative importance of each task with values ranging between 0 and 1. In the joint classification and regression, we want to perform both tasks simultaneously with equal importance, so the values of β_1 and β_2 are equal to 1. Knowing the severity level of the patients along the classification task will help clinicians determine the disease's advancement and can also help monitor the progress of treatment over time. In the binary and multiclass classification cases, we investigate the impact of gradually increasing the weight β_2 assigned to the sex prediction auxiliary task in the multitasking approach. This allows us to assess its influence on the accuracy of the main classification task. The detailed explanation of the effect of β_2 values in binary and multiclass scenarios is discussed in section III.

As discussed in section II-B, five pre-trained computer vision models were used in these experiments. These models were initially developed for image classification tasks; they have single-task architecture. To use them for our work, we need to modify the architectures. The modification involves replacing the original classifier layer with a new classifier with a number of output features equal to the number of classes in our dataset. In the binary case, two output features represent healthy and dysphonia classes, while in the multiclass case, three classes represent Healthy control (HC), Functional Dysphonia (FD), and Organic Dysphonia (OD). The modification was performed so we preserved the original architecture of the models; for example, in the case of ConvNeXt, the final classifier consists of (Linear, Relu, Dropout, and Linear). So, we replaced the original sequential

model with a new sequential model with the same number of layers for the classifier, and we added a new head with the same sequential order but changed the last linear layer output to 1 for the regression task. This scenario of modification and preserving the original structure is identical to the other four models.

By specifying the value of β_1 and β_2 in Eq.(1), we allow the network to work in multitask or single task technique. For performing a single task classification $\beta_2 = 0$, in this case, the network tries to optimize parameters only related to the classification of the speech samples. Likewise, for the regression task, we update the values of β_1 and β_2 to 0 and 1, respectively. In this case, the model is only trained to predict severity estimation.

D. EXPERIMENT CASES

Overall, six cases were performed in this paper, three for STL and three for MTL. We will explain them briefly here.

- Joint classification and regression- MTL: performing classification of Healthy vs Dysphonia and predicting severity level simultaneously. For this experiment, values of β_1 and β_2 are equal to 1.
- Binary classification- STL: distinguishing between healthy and dysphonia speech sample only, $\beta_1 = 1$ and $\beta_2 = 0$.
- Regression- STL: predicting severity level of dysphonia patient, $\beta_1 = 0$ and $\beta_2 = 1$
- Binary classification- MTL with sex as an auxiliary task: trying different values of β_2
- Multiclass classification- MTL with sex as an auxiliary task: classifying Normal vs Organic vs Functional dysphonia.
- Multiclass classification- STL: same as binary classification β_2=0

E. SPEECH FEATURES: MEL SPECTROGRAM

As has been explained in section II-A, participants read a short paragraph. The length of the recording is different between files. Deep learning models used require fixed-length input features. Considering our sample's overall length, we selected 40 seconds of speech as input to the network. Speech files longer than 40 seconds were truncated, and shorter files were padded with zeros. Mel spectrogram is used as an input for the CNN model. A spectrogram is the time-frequency representation of a speech signal. It shows how the frequency of the speech changes over time. The mel spectrogram, on the other hand, is a logarithmic transformation of the frequency axis of the traditional spectrogram, which makes it more perceptually relevant to human hearing. The mel spectrogram is calculated by dividing the audio signal into short overlapping frames and then applying a Fourier transform to each frame to obtain the frequency content. The frequency axis of the spectrogram is then converted to the mel scale using a nonlinear transformation, which considers the nonlinear nature of human perception of sound. For our study, we extracted 128 mel spectrogram features using a 25ms Hamming window with 10ms overlapping. Figure 2 presents the mel spectrogram for male and female participants. From left to right, the figures show the mel spectrogram for healthy controls and dysphonia samples with severity levels ranging from 1 (SL 1) to 3 (SL 3). As observed in the figure, there is a clear distinction between the mel spectrograms of healthy individuals and those with dysphonia.

F. EVALUATION METRICS

Different evaluation metrics are needed to measure the proposed model's performance for classification and regression. Accuracy, sensitivity, specificity, and f1-score have been used to measure the performance of the classification tasks. Accuracy (Eq. 2) measures the proportion of correctly classified instances, both True Positives (TP) and True Negatives (TN), over the total number of instances.

$$Accuracy(Acc) = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

Sensitivity (Eq. 3) often refers to how well a trained model can detect the presence of a disease or condition, in our case, dysphonia. A high-sensitivity model will accurately identify most individuals with the disorder.

$$Sensitivity(Sen) = \frac{TP}{TP + FN} \tag{3}$$

Specificity (Eq. 4) is the ability of the model to correctly classify an individual who does not have the disease or condition.

$$Specificity(Spec) = \frac{TN}{TN + FP} \tag{4}$$

The F1 score (Eq. 5) is the harmonic mean of precision and recall and provides a single score that balances both metrics. Precision is the proportion of true positive predictions among all positive predictions, while recall is the proportion of true positive predictions among all actual positive cases.

$$F1score(F1) = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5}$$

where precision and recall are calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

Room Mean Square Error (RMSE) and Pearson Correlation coefficient have been used to measure how well our models perform in the case of regression. RMSE (Eq. 8) measures the average difference between the predicted and actual severity levels.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{8}$$

where y_i represents the actual value of the i-th sample's severity, and $\hat{y}_i$ represents the predicted severity of the i-th sample. The formula computes the square root of the average

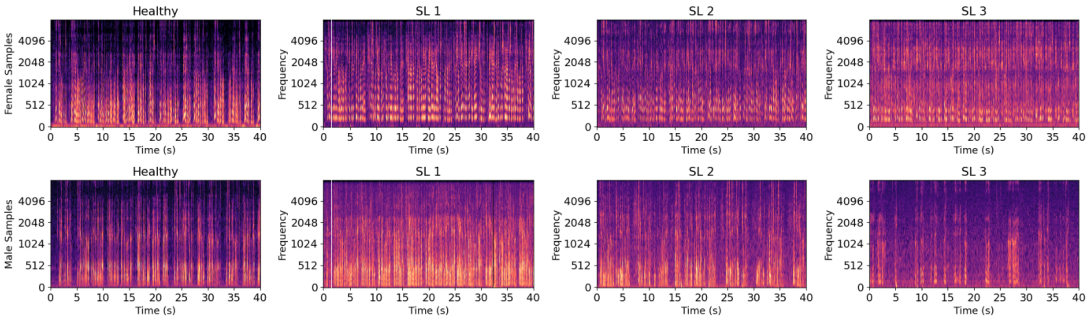

FIGURE 2. Mel spectrogram for speech samples: for female and male samples, respectively.

TABLE 2. Classification and regression using MTL and STL (MTL/STL) for 5-fold cross-validation.

Model	Type	Acc%	F1	Sen%	Spec%	RMSE	PearCorre
ResNet	Final	**90.11** / 87.86	**0.922** / 0.904	**93.21** / 90.36	**85.00** / 84.00	**0.567** / 0.567	**0.874** / 0.866
	Best	**89.44** / 87.19	**0.917** / 0.899	**92.50** / 90.71	**84.00** / 81.00	**0.547** / 0.555	0.873 / **0.873**
DenseNet	Final	**82.70** / 77.98	**0.848** / 0.810	**80.00** / 75.71	**87.00** / 82.00	0.807 / **0.653**	0.793 / **0.825**
	Best	**85.17** / 79.78	**0.873** / 0.832	**81.78** / 80.36	**91.00** / 79.00	**0.622** / 0.631	**0.841** / 0.825
MobileNet	Final	89.44 / **90.11**	0.916 / **0.919**	**92.15** / 88.93	85.00 / **92.00**	0.552 / **0.547**	**0.885**/ 0.878
	Best	**91.69** / 88.99	**0.932** / 0.915	90.71 / **93.57**	**93.00** / 81.00	0.551 / **0.546**	0.875 / **0.879**
ConvNeXt	Final	**91.23** / 86.74	**0.933** / 0.902	**96.07** / 95.71	**83.00** / 72.00	0.564 / **0.550**	0.870 / **0.879**
	Best	**89.44** / 81.58	**0.913** / 0.860	**90.36** / 90.00	**88.00** / 67.00	0.556 / **0.555**	0.870 / **0.872**
EfficientNet	Final	**76.63** / 74.83	**0.821** / 0.781	**85.00** / 73.57	62.00 / **77.00**	0.881 / **0.809**	0.685 / **0.712**
	Best	76.85 / **77.53**	0.818 / **0.820**	**83.21** / 81.43	66.00 / **71.00**	0.819 / **0.807**	0.697 / **0.708**

of the squared differences between the actual and predicted values.

The Pearson correlation coefficient(Eq. 9) measures the strength and direction of the linear relationship between the predicted and actual severity level values.

$$r(PearCorre) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (9)$$

x_i and y_i are the values of the actual and predicted severity score for the i-th sample, $\bar{x}$ and $\bar{y}$ are the sample means of the two variables, and n is the total number of samples in the test set.

III. RESULTS

The results reported in this section are the average of five results, each corresponding to models saved in each fold.

A. BINARY CLASSIFICATION AND REGRESSION (MTL VS STL)

The results presented in Table 2 demonstrate the performance of models in three different experiments, namely joint classification and severity estimation using MTL, classification, and regression using STL. The bold values represent better performance between MTL and STL, while the bold and underlined values indicate the best performance across all models (both MTL and STL). Values shown in each column correspond to MTL and STL, respectively. Overall, MTL exhibits better performance compared to STL across various metrics. MTL consistently achieves higher accuracy, F1 score, sensitivity, and specificity in all the models except in EfficientNet. Regarding sensitivity, which is the ability of the models to identify patients with dysphonia correctly, we can see that ConvNeXt in MTL archives slightly above 96%, which outperforms all other models.

A detailed analysis of the results reveals slight variations between MTL and STL approaches in terms of severity estimation performance. While there are instances where MTL achieves a slightly lower RMSE than STL, it is essential to note that these differences are not substantial. For example, in the case of ResNet, the MTL model achieves an RMSE of 0.547 in the best configuration, while the STL model also achieves an RMSE of 0.555. Similarly, the best DenseNet model yields an RMSE of 0.622 for MTL and 0.631 for STL. These differences, although present, are minimal and may not significantly impact the overall regression performance.

The MobileNet model stands out with the lowest RMSE of 0.546 when using STL, indicating its better performance in

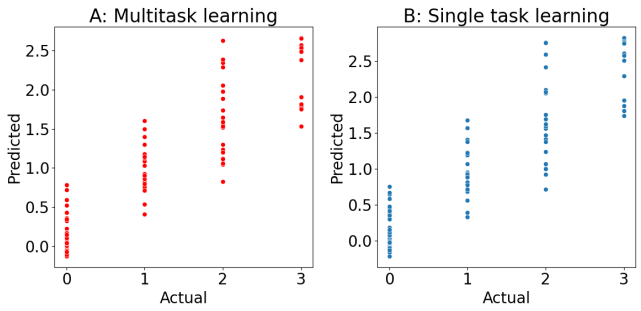

FIGURE 3. Severity estimation of MTL and STL average of 5-fold cross-validation.

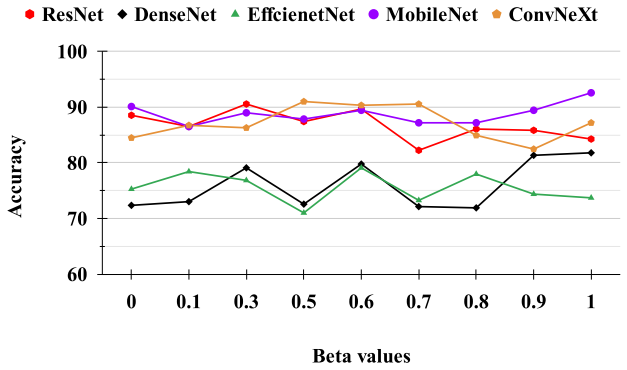

FIGURE 4. Binary classification accuracy according to different beta values.

terms of severity estimation. However, it is noteworthy that the best Pearson correlation value of 0.885 was obtained by the MobileNet model using MTL. In some scenarios, such as the MobileNet model, the STL approach demonstrates a lower RMSE of 0.547 in the final configuration compared to MTL's RMSE of 0.552. Similarly, the STL model of EfficientNet achieves an RMSE of 0.807 in its best configuration, slightly outperforming MTL's RMSE of 0.819.

The most significant difference between MTL and STL regarding RMSE is observed in the DenseNet model, where STL achieves a nearly 0.154 lower RMSE in the final model. However, in the best model configuration, MTL performs better with an almost 0.009 lower RMSE compared to STL. Figure 3 shows the scatter plot of the average severity prediction of MobilenNet by MTL and STL, respectively.

From Figure 3, x-axis represents the actual severity values in the test set, while y-axis corresponds to the predicted score by MTL and STL architecture. The difference between the MTL and STL predicted severity score could be seen from the scatter plot, especially in the case of SL2 and SL3.

These findings highlight that while there are minor variations in the RMSE values between MTL and STL for severity estimation, the overall differences are not significant. It indicates that both approaches can effectively address the regression task, with only slight performance variations observed in specific cases.

B. SEX IN BINARY CLASSIFICATION

In this experiment, the sex prediction task was incorporated as an auxiliary task in the MTL framework for classifying between dysphonia and healthy speech. The emphasis on predicting sex was controlled by varying the values of β_2 in the loss function. By starting with a value of 0, which indicates STL (only focusing on distinguishing between dysphonia and healthy speech), and incrementally increasing β_2 until reaching 1, we determined the degree of importance placed on sex prediction by the neural networks. This allowed us to analyze the impact of incorporating sex as an auxiliary task on the overall performance of the neural networks. Figure 4 presents the performance analysis of five different models using MTL and STL approaches. The lines in the graph represent individual models, while the x-axis showcases the variation in beta values, representing the weight assigned to the auxiliary task. The trends of the line are different according to the beta values. The line charts reveal that integrating sex prediction as an auxiliary task in binary classification yields improved accuracy. MobileNet with MTL achieved the highest accuracy among all the models, particularly when the beta value was set to 1. Notably, the DenseNet model exhibited a significant disparity between STL and MTL, with MTL achieving an accuracy of 81.8 compared to STL's 72.36 binary classification. Moreover, the performance of the ConvNeXt remains stable within the range of beta values between 0.5 and 0.7, which surpasses the single task model by nearly 4%.

However, it is critical to note that there are instances where STL performs better than MTL, as evidenced by the downward trend in performance for the ResNet model when the beta value exceeds 0.6. Also, in the case of EfficientNet, when a beta is equal to 0.5.

Table 3 presents the comprehensive results for the MobileNet model, considered the top-performing model among the various models evaluated in both cases. Due to the many results, we have chosen to showcase the outcomes for this specific model only. The table illustrates that the model performs better when trained using MTL compared to STL, with an improvement of just over 2%

C. SEX IN MULTICLASS CLASSIFICATION

The models were trained to classify speech samples into healthy, organic, and functional dysphonia in a multiclass classification experiment. To enhance the classification performance, we incorporated sex prediction as an auxiliary task. By considering sex as an additional aspect, the models learned to extract features related to both the speech disorder and the sex of the speaker, leading to potentially improved classification accuracy. Similar to binary classification, we explored the effects of different values of β_2. This allowed us to assess the influence of sex information on the overall performance of the models in classifying the different types of dysphonia. Figure 5 presents the trends of classification accuracy across different beta values, representing the average performance of five models. The line

TABLE 3. Binary classification for MobileNet with different beta values in 5-fold cross-validation.

Beta	F_Acc%	F_F1	F_Sen%	F_Spec%	B_Acc%	B_F1	B_Sen%	B_Spec%
0	90.11	0.921	91.43	87.88	89.89	0.9198	**92.14**	86.06
0.1	86.52	0.889	86.79	86.06	91.23	0.929	90.72	**92.12**
0.3	88.99	0.910	88.93	89.09	90.34	0.922	90.36	90.3
0.5	87.86	0.894	83.21	**95.76**	88.99	0.909	87.5	91.52
0.6	89.44	0.9152	90.71	87.27	88.99	0.912	90.72	86.06
0.7	87.19	0.894	88.57	84.85	88.99	0.910	88.22	90.3
0.8	87.19	0.896	87.86	86.06	89.89	0.918	89.64	90.3
0.9	89.44	0.916	91.43	86.06	90.56	0.924	91.07	89.7
1	**92.58**	**94.12**	**95**	88.49	**91.24**	**0.929**	91.43	90.91

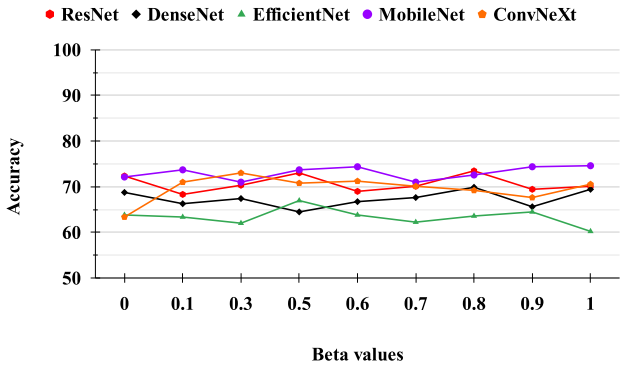

FIGURE 5. Multiclass classification accuracy according to the different beta values.

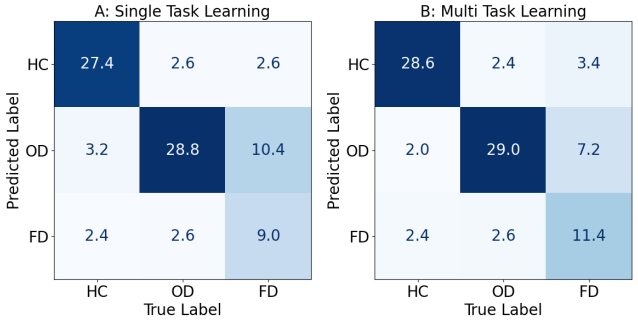

FIGURE 6. Confusion matrices of MobileNet in MTL and STL.

chart reveals that the MobileNet model achieved the highest accuracy of 77.53% when beta was set to 0.8. Comparatively, the best accuracy attained using STL was 73.26%, indicating an improvement of more than 4% when employing the MTL. In case of ConvNeXt and ResNet, we can see that MTL outperforms STL in most beta values.

Nevertheless, it is interesting in the case of EfficientNet; regardless of the beta value used, the accuracy achieved through MTL was consistently lower than the STL. This finding suggests that for the EfficientNet model in particular, incorporating the sex prediction task as an auxiliary task did not contribute to improved performance in the multiclass case. By analyzing the results across various beta values, we gained insights into the significance of sex in the context of multiclass dysphonia classification. Table 4 shows the detailed metrics of the best-performing model in both STL and MTL.

As shown in Table 4, MTL approach with a beta value of 0.8 consistently outperforms the counterpart approach in all metrics for both the final and early stopping models. The MTL achieved significant improvements of more than 5% in both sensitivity and F1 scores compared to the STL model.

Confusion Matrices from Figure 6A and B represent the performance of a classification model in classifying instances into three classes: HC (Healthy Control), OD (Organic Dysphonia), and FD (Functional Dysphonia) in STL and MTL, respectively. Each column represents the actual class, and each row represents the predicted class. The values in the matrix indicate the average number of instances in five folds classified into that class. The diagonal values correspond to the number of correct predictions. Summing the values of each column will determine the number of samples in that specific class in the test set.

From the analysis of the confusion matrices, it is evident that MTL approach outperforms STL in the classification of dysphonia categories. It shows higher accuracy in correctly predicting instances across all three classes, with a notable improvement in distinguishing functional dysphonia. Specifically, MTL accurately classifies 11.4 instances of functional dysphonia, compared to only the STL's nine correct predictions out of 22 total samples. This highlights the effectiveness of incorporating sex prediction as an auxiliary task in the MTL approach for better classification accuracy. Furthermore, the MTL demonstrates a lower number of incorrect classifications between organic and functional dysphonia, indicating its superior ability to differentiate between these two categories.

TABLE 4. Accuracy of MobileNet 5-fold multiclass classification.

Beta	F_Acc%	F_Sen%	F_Spec%	F_F1	B_Acc%	B_Sen%	B_Spec%	B_F1
0	73.26	69.50	86.20	69.10	72.140	68.10	85.60	67.60
0.1	75.05	71.00	87.00	69.90	73.710	70.70	86.40	71.10
0.3	71.69	67.40	85.30	66.60	71.010	66.50	84.70	65.40
0.5	71.91	68.30	85.50	67.40	73.710	69.60	86.30	69.10
0.6	74.16	70.60	86.60	70.60	74.380	70.80	86.70	70.80
0.7	72.58	69.20	85.90	68.90	71.010	66.80	85.00	66.50
0.8	**77.53**	**74.60**	**88.50**	**74.60**	72.580	70.00	86.20	69.70
0.9	70.79	66.90	84.90	66.20	74.380	**71.60**	**86.80**	**71.90**
1	72.81	70.40	86.20	69.90	**74.610**	70.40	86.70	69.90

IV. DISCUSSION

Although the research area of distinguishing normal from dysphonic voice and estimating the severity of dysphonia has long been studied in the literature, these methods only focused on one task at a time. The objective of this study is to investigate the performance of MTL compared to STL in dysphonia classification and severity estimation tasks. Our findings shed light on the benefits and limitations of incorporating MTL approaches in these tasks.

In the first experiments, we aimed to achieve two objectives: classifying normal from dysphonic speech and estimating the severity score together. We consider both tasks to be equally important. Predicting the severity scores and differentiating between healthy speech and dysphonia provide practical insights into the patient's condition. This comprehensive approach enables clinicians to understand the patient's situation better and effectively monitor their progress in both the treatment stages and the progressiveness of the condition. Results from joint learning indicate that MTL models demonstrated better performance across various evaluation metrics, including accuracy of 91.69%, F1 score of 93.20%, and specificity of 93.57% in case of MobileNet and F1 score of 93.3% with ConvNeXt. The significant improvement in these metrics suggests that jointly considering classification and regression tasks using MTL facilitates more effective dysphonia classification. The ability of MTL models to leverage shared representations and learn task-specific features simultaneously leads to their enhanced performance.

Regarding severity estimation, both MTL and STL approach effectively tackled the regression task. The minor variations observed concerning RMSE indicate that both approaches demonstrate similar accuracy in estimating severity of the condition. The occurrence of negative transfer, where the performance of one task negatively impacts another in multitask learning, was not significant. While the difference between STL and MTL is insignificant, the best-performing model, MobileNet, showed a slight advantage of 0.005 RMSE for STL. However, MTL outperformed STL with a difference of 0.006 concerning the Pearson correlation metric. This finding suggests that the choice of MTL may not impact the regression performance compared to the STL.

When using sex prediction as an auxiliary task, our results indicated that integrating this additional task improves the accuracy of dysphonia classification both in binary and multiclass scenarios. The weight assigned to the auxiliary task played an essential role in determining the extent of this improvement. The line charts demonstrated clear trends, with certain models achieving their highest accuracy at specific beta values. In the classification of healthy vs. dysphonia speech, the best performance among all computer vision models was achieved by MobileNet with an accuracy of 92.58% and F1 score of 94.12%. Moreover, in multiclass classification experiments, MTL outperforms STL and achieves 77.53% accuracy compared to 73.26% in STL. MTL demonstrates improved abilities in differentiating between organic and functional dysphonia, potentially aiding clinicians in more accurately classifying and managing specific types of voice disorders using non-invasive and cost-effective methods. However, compared to binary classification, both STL and MTL approaches face challenges in distinguishing between organic and functional dysphonia in general, which suggests these disorders affect speech similarly. By examining the confusion matrix of the MTL approach, as shown in Figure 6-B, it becomes evident that MTL has better capabilities for differentiating between these two classes. More research needs to be performed to understand the reason behind this confusion between the two types of dysphonia. By learning from multiple tasks, models can effectively learn the shared information and extract robust features related to dysphonia evaluation. This shared representation learning improves the model's ability to generalize to unseen data and enhances its performance in real-world scenarios. MTL also offers advantages regarding computational costs and training time. Rather than training two separate models, one for each specific task, MTL allows both tasks to be performed within

the same framework. These findings suggest that including sex prediction as an auxiliary task provides valuable contextual information that complements the dysphonia classification task.

In clinical practice, classification and severity estimation of dysphonia specialists rely on standards like GRBAS, CAPE-V, RBH, and other diagnostic frameworks. These existing standards serve as a benchmark for clinicians in assessing voice disorders. The advantage of this method is that it provides clinicians with novel computational and cost-effective methodologies along these standards. MTL offers valuable insights to clinicians in tandem with their existing diagnostic approaches and experiences. Furthermore, the methods described here can serve as pre-screening for pre-diagnosis stages, such as general practitioner examination, or can even be used at home on mobile devices. This could shed light on a possible pathological affection, thus bringing the meeting with an expert to facilitate correct treatment. Ultimately, the successful integration of computational methodologies into dysphonia assessment should be used as a complement, rather than replace, enabling a collaborative approach that leverages both technological advancements and clinical expertise.

V. CONCLUSION

In this paper, we presented an approach that combines transfer learning and multitask learning for the binary classification and severity estimation of dysphonia, as well as multiclass classification. By leveraging five computer vision deep learning architectures, we fine-tuned them on our dataset after modifying their architecture to adapt multitask learning. A significant advantage of these deep learning models is their ability to learn feature representations without manual feature engineering, which is often required in traditional ML methods. Our experimental results revealed the benefits of multitask learning in the joint classification and severity estimation of dysphonia. Compared to single-task learning counterparts, the multitask learning models demonstrated more promising performance in distinguishing between healthy and dysphonic speech while maintaining a comparable level of accuracy in severity estimation. This demonstrates the effectiveness of leveraging shared knowledge and interdependencies between tasks to enhance overall performance. Moreover, we found that multitask learning facilitated better feature representation learning, enabling the models to discriminate between organic and functional dysphonia effectively. This improved capability to distinguish between the two types of dysphonia highlights the advantage of using sex of the speakers as an auxiliary task in MTL.

In conclusion, our proposed approach demonstrates promising results in dysphonia classification and severity estimation. By leveraging deep learning architectures and exploiting the interdependencies between tasks, we achieve enhanced performance and contribute to a better understanding of dysphonia-related factors. Future research is to expand the range of additional tasks to improve the performance of multitask learning further. Additionally, evaluating the generalizing ability of our approach on larger, more diverse, and multilingual datasets would provide valuable insights. These efforts will contribute to advancing the field of dysphonia assessment and its clinical applications.

REFERENCES

[1] P. Belin, S. Fecteau, and C. Bédard, "Thinking the voice: Neural correlates of voice perception," *Trends Cognit. Sci.*, vol. 8, no. 3, pp. 129–135, Mar. 2004.

[2] G. Szatloczki, I. Hoffmann, V. Vincze, J. Kalman, and M. Pakaski, "Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease," *Frontiers Aging Neurosci.*, vol. 7, p. 195, Oct. 2015.

[3] A. S. Cohen, J. E. McGovern, T. J. Dinzeo, and M. A. Covington, "Speech deficits in serious mental illness: A cognitive resource issue?" *Schizophrenia Res.*, vol. 160, nos. 1–3, pp. 173–179, Dec. 2014.

[4] A. S. Cohen and B. Elvevåg, "Automated computerized analysis of speech in psychiatric disorders," *Current Opinion Psychiatry*, vol. 27, no. 3, pp. 203–209, 2014.

[5] M. L. Poole, A. Brodtmann, D. Darby, and A. P. Vogel, "Motor speech phenotypes of frontotemporal dementia, primary progressive aphasia, and progressive apraxia of speech," *J. Speech, Lang., Hearing Res.*, vol. 60, no. 4, pp. 897–911, Apr. 2017.

[6] A. Suppa et al., "Voice in Parkinson's disease: A machine learning study," *Frontiers Neurol.*, vol. 13, Feb. 2022, Art. no. 831428.

[7] B. Hajduska-Dér, G. Kiss, D. Sztahó, K. Vicsi, and L. Simon, "The applicability of the beck depression inventory and Hamilton depression scale in the automatic recognition of depression based on speech signal processing," *Frontiers Psychiatry*, vol. 13, p. 1767, Aug. 2022.

[8] M. M. Johns, R. T. Sataloff, A. L. Merati, and C. A. Rosen, "Article commentary: Shortfalls of the American academy of otolaryngology—Head and neck surgery's *clinical practice guideline: Hoarseness (dysphonia)*," *Otolaryngol.-Head Neck Surg.*, vol. 143, no. 2, pp. 175–177, Aug. 2010.

[9] E. Nerrière, M.-N. Vercambre, F. Gilbert, and V. Kovess-Masféty, "Voice disorders and mental health in teachers: A cross-sectional nationwide study," *BMC Public Health*, vol. 9, no. 1, pp. 1–8, Dec. 2009.

[10] R. J. Stachler et al., "Clinical practice guideline: Hoarseness (dysphonia) (update)," *Otolaryngol.-Head Neck Surg.*, vol. 158, no. S1, pp. S1–S42, Mar. 2018.

[11] L. Crevier-Buchman, T. Ch, A. Sauvignet, S. Brihaye-Arpin, and M.-C. Monfrais-Pfauwadel, "Diagnosis of non-organic dysphonia in adult," *Revue Laryngologie-Otologie-Rhinologie*, vol. 126, no. 5, pp. 353–360, 2005.

[12] A. Millar, I. J. Deary, J. A. Wilson, and K. MacKenzie, "Is an organic/functional distinction psychologically meaningful in patients with dysphonia?" *J. Psychosomatic Res.*, vol. 46, no. 6, pp. 497–505, Jun. 1999.

[13] A. E. Aronson, *Clinical Voice Disorders, An Interdisciplinary Approach / Arnold E. Aronson ; [Medical Ill., Floyd E. Hosmer]*. New York, NY, USA: B. C. Decker, 1990.

[14] N. P. Connor and D. M. Bless, *Functional and Organic Voice Disorders* (Cambridge Handbooks in Language and Linguistics). Cambridge, U.K.: Cambridge Univ. Press, 2013.

[15] C. L. Payten, G. Chiapello, K. A. Weir, and C. J. Madill, "Frameworks, terminology and definitions used for the classification of voice disorders: A scoping review," *J. Voice*, vol. 2022, p. 89, Mar. 2022.

[16] C. Sapienza and B. Hoffman, *Voice Disorders*. San Diego, CA, USA: Plural Publishing, Inc., 2020.

[17] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis—Jitter, shimmer and HNR parameters," *Proc. Technol.*, vol. 9, pp. 1112–1122, Jan. 2013.

[18] Z. Kh. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, vol. 10, pp. 122136–122158, 2022.

[19] S. Jothilakshmi, "Automatic system to detect the type of voice pathology," *Appl. Soft Comput.*, vol. 21, pp. 244–249, Aug. 2014.

[20] J. Reid, P. Parmar, T. Lund, D. K. Aalto, and C. C. Jeffery, "Development of a machine-learning based voice disorder screening tool," *Amer. J. Otolaryngol.*, vol. 43, no. 2, Mar. 2022, Art. no. 103327.

[21] D. R. A. Leite, R. M. de Moraes, and L. W. Lopes, "Different performances of machine learning models to classify dysphonic and non-dysphonic voices," *J. Voice*, vol. 2022, pp. 1–10, Dec. 2022.

[22] L. Verde, G. De Pietro, and G. Sannino, "Voice disorder identification by using machine learning techniques," *IEEE Access*, vol. 6, pp. 16246–16255, 2018.

[23] Z. Dankovičová, D. Sovák, P. Drotár, and L. Vokorokos, "Machine learning approach to dysphonia detection," *Appl. Sci.*, vol. 8, no. 10, p. 1927, Oct. 2018.

[24] P. Harar, Z. Galaz, J. B. Alonso-Hernandez, J. Mekyska, R. Burget, and Z. Smekal, "Towards robust voice pathology detection: Investigation of supervised deep learning, gradient boosting, and anomaly detection approaches across four databases," *Neural Comput. Appl.*, vol. 32, no. 20, pp. 15747–15757, Oct. 2020.

[25] S. N. Awan and N. Roy, "Toward the development of an objective index of dysphonia severity: A four-factor acoustic model," *Clin. Linguistics Phonetics*, vol. 20, no. 1, pp. 35–49, Jan. 2006.

[26] M. G. Tulics and K. Vicsi, "The automatic assessment of the severity of dysphonia," *Int. J. Speech Technol.*, vol. 22, no. 2, pp. 341–350, Jun. 2019.

[27] M. Wu and L. Chen, "Image recognition based on deep learning," in *Proc. Chin. Autom. Congr. (CAC)*, 2015, pp. 542–546.

[28] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, "Natural language processing advancements by deep learning: A survey," 2020, *arXiv:2003.01200*.

[29] U. Kamath, J. Liu, and J. Whitaker, *Deep Learning for NLP and Speech Recognition*, vol. 84. Cham, Switzerland: Springer, 2019.

[30] J.-N. Lee and J.-Y. Lee, "An efficient SMOTE-based deep learning model for voice pathology detection," *Appl. Sci.*, vol. 13, no. 6, p. 3571, Mar. 2023.

[31] R. Islam and M. Tarique, "A novel convolutional neural network based dysphonic voice detection algorithm using chromagram," *Int. J. Electr. Comput. Eng.*, vol. 12, no. 5, p. 5511, Oct. 2022.

[32] 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.

[33] 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*, Sep. 2018.

[34] 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.

[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] A. N. Omeroglu, H. M. A. Mohammed, and E. A. Oral, "Multi-modal voice pathology detection architecture based on deep and handcrafted feature fusion," *Eng. Sci. Technol., Int. J.*, vol. 36, Dec. 2022, Art. no. 101148.

[37] L. Chen and J. Chen, "Deep neural network for automatic classification of pathological voice signals," *J. Voice*, vol. 36, no. 2, pp. 288.e15–288.e24, Mar. 2022.

[38] D. Sztahó, K. Gábor, and T. Gábriel, "Deep learning solution for pathological voice detection using LSTM-based autoencoder hybrid with multi-task learning," in *Proc. 14th Int. Joint Conf. Biomed. Eng. Syst. Technol.*, 2021, pp. 135–141.

[39] M. G. Tulics, L. J. Lavati, K. Mészáros, and K. Vicsi, "Possibilities for the automatic classification of functional and organic dysphonia," in *Proc. Int. Conf. Speech Technol. Hum.-Comput. Dialogue (SpeD)*, Oct. 2019, pp. 1–6.

[40] C. Robotti et al., "Treatment of relapsing functional and organic dysphonia: A narrative literature review," *Acta Otorhinolaryngol. Ital.*, vol. 43, p. S84, Apr. 2023.

[41] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[42] R. Schönweiler, M. Hess, P. Wübbelt, and M. Ptok, "Novel approach to acoustical voice analysis using artificial neural networks," *J. Assoc. Res. Otolaryngol.*, vol. 1, pp. 270–282, Jan. 2000.

[43] M. Ptok, C. Schwemmle, C. Iven, M. Jessen, and T. Nawka, "[On the auditory evaluation of voice quality]," *HNO*, vol. 54, pp. 793–802, Oct. 2006.

[44] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12 no. 10, pp. 2825–2830, 2012.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[46] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[47] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[48] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11966–11976.

[49] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[50] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[51] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee, "12-in-1: Multi-task vision and language representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10434–10443.

[52] J. Worsham and J. Kalita, "Multi-task learning for natural language processing in the 2020s: Where are we going?" *Pattern Recognit. Lett.*, vol. 136, pp. 120–126, Aug. 2020.

[53] B. T. Atmaja, A. Sasou, and M. Akagi, "Speech emotion and naturalness recognitions with multitask and single-task learnings," *IEEE Access*, vol. 10, pp. 72381–72387, 2022.