cancers

MDPI

Article

Lightweight Deep Learning Model for Assessment of Substitution Voicing and Speech after Laryngeal Carcinoma Surgery

Rytis Maskeliūnas [1], Audrius Kulikajevas [1], Robertas Damaševičius [1,*], Kipras Pribuišis [2], Nora Ulozaitė-Stanienė [2] and Virgilijus Uloza [2]

[1] Faculty of Informatics, Kaunas University of Technology, 51368 Kaunas, Lithuania; rytis.maskeliunas@ktu.lt (R.M.); audrius.kulikajevas@ktu.edu (A.K.)
[2] Department of Otorhinolaryngology, Lithuanian University of Health Sciences, 50061 Kaunas, Lithuania; Kipras.pribuisis@lsmuni.lt (K.P.); Nora.ulozaite@lsmuni.lt (N.U.-S.); Virgilijus.ulozas@lsmuni.lt (V.U.)
* Correspondence: robertas.damasevicius@ktu.lt

Simple Summary: A total laryngectomy involves the full and permanent separation of the upper and lower airways, resulting in the loss of voice and inability to interact vocally. To identify, extract, and evaluate replacement voicing following laryngeal oncosurgery, we propose employing convolutional neural networks for categorization of speech representations (spectrograms). With an overall accuracy of 89.47 percent, our technique has the greatest true-positive rate of any of the tested state-of-the-art methodologies.

Abstract: Laryngeal carcinoma is the most common malignant tumor of the upper respiratory tract. Total laryngectomy provides complete and permanent detachment of the upper and lower airways that causes the loss of voice, leading to a patient's inability to verbally communicate in the postoperative period. This paper aims to exploit modern areas of deep learning research to objectively classify, extract and measure the substitution voicing after laryngeal oncosurgery from the audio signal. We propose using well-known convolutional neural networks (CNNs) applied for image classification for the analysis of voice audio signal. Our approach takes an input of Mel-frequency spectrogram (MFCC) as an input of deep neural network architecture. A database of digital speech recordings of 367 male subjects (279 normal speech samples and 88 pathological speech samples) was used. Our approach has shown the best true-positive rate of any of the compared state-of-the-art approaches, achieving an overall accuracy of 89.47%.

Keywords: laryngeal carcinoma; substitution voicing; voice analysis; convolutional neural networks; deep learning

Citation: Maskeliūnas, R.; Kulikajevas, A.; Damaševičius, R.; Pribuišis, K.; Ulozaitė-Stanienė, N.; Uloza, V. Lightweight Deep Learning Model for Assessment of Substitution Voicing and Speech after Laryngeal Carcinoma Surgery. *Cancers* **2022**, *14*, 2366. https://doi.org/10.3390/cancers14102366

Academic Editors: Muhammad Fazal Ijaz and Marcin Woźniak

Received: 11 April 2022
Accepted: 4 May 2022
Published: 11 May 2022

1. Introduction

Laryngeal carcinoma remains the most common malignant tumor of the upper respiratory tract worldwide as reported by Steuer et al. [1]. Literature reports an incidence of around 5 cases per 100,000 inhabitants but National Cancer Institute's Cancer registry reported 18.3 cases per 100,000 Lithuanian citizens [2]. The most current American Cancer Society estimates for laryngeal cancer in the United States for 2022 are: estimated 12,470 new cases of laryngeal cancer, and predicted 3820 deaths from laryngeal cancer [3]. Although the overall incidence is declining, laryngeal cancer is one of the few oncological diseases in which the 5-year survival rate has decreased over the past 40 years, from 66% to 63%. This may be attributed to more conservative treatment protocols, as well as factors that might delay the patient's follow-up, mainly—the lack of medical care availability near the patient's place of residence as described by the report in Journal of Clinical Oncology [4]. Programs that require less specialized medical care and provide patients with reliable

follow-up means might help to improve the 5-year survival rate, as well as, increase patient safety during the pandemics [5]. Software that reduces the need for specialized medical care might free up medical facilities for COVID-19 patients. Additionally, this software might reduce the workload of specialized medical personnel and make them available for COVID-19 related tasks. Fewer nonessential trips to outpatient facilities lead to a lower risk of infection during pandemics [6]. This can potentially be achieved without incurring additional costs to the healthcare system.

Chemoradiotherapy and surgery are usually feasible treatment choices for patients with early (stage I-II) laryngeal cancer. The extent of surgery is primarily determined by the tumor's spread. Depending on the tumor stage, surgical treatment results in locoregional cancer control comparable to that provided by laryngeal radiation or chemoradiation therapy or even higher survival rates, cancer can be achieved for patients who undergo surgical treatment for advanced-stage laryngeal [1].

After laryngeal oncosurgery that may include extended cordectomy (removal of the vocal fold), partial or total laryngectomy patients lose one or even both vocal folds. As a consequence, the voice is generated by a single vocal fold oscillating with the remaining laryngeal and pharyngeal structures or alaryngeal (oesophageal or tracheoesophageal) speech is used. These conditions can be considered as substitution voicing (SV), which is defined as the voicing without two true vocal folds [7]. In SV, involuntary aphonic (unvoiced) segments of speech coexist with rough-voiced ones. Various degrees of speech impairment or even a complete inability to speak after laryngeal oncosurgery are the most important complaints expressed by patients and may lead to their social isolation [8].

During the current pandemic, a lot of specialized medical care facilities and personnel have been dedicated to fighting COVID-19 [9]. This in turn led to delayed diagnostics for primary laryngeal cancer patients and follow-up for patients after treatment [10]. This resulted in the need of more radical cancer treatments and increased patient mortality which otherwise could have been avoided. More than half of laryngeal cancer patients present with stage III or higher at the first appointment. For patients with those stages, total laryngectomy is usually advised for favorable locoregional cancer control and an optimal 5-year survival rate [11]. Total laryngectomy is also performed when the patient is not eligible for conservative techniques like chemotherapy and radiotherapy or in case of their failure. Total laryngectomy provides complete and permanent detachment of the upper and lower airways. This separation causes the loss of voice, smell, xerostomia, and altered taste. Total laryngectomy leads to a patient's inability to verbally communicate in the postoperative period. Patients after laryngectomy often have to rely on pen and paper or other forms of written text to communicate anywhere from 2 weeks to 6 months after the initial surgery. This is especially troubling during the COVID -19 pandemic when patients have to rely on text messaging to contact their families and have trouble receiving basic social or telemedicine care simply because they can not use the phone by themselves [12].

According to Pereira da Silva et al., loss of voice has a significant influence on the quality of life of laryngeal cancer patients [13]. It has an impact on their communication, social life, and even their ability to keep a job. Furthermore, failure to communicate effectively generates worry, and 40–57% of these people develop a serious depressive condition [14]. As a result, it is critical to give trustworthy voice and speech rehabilitation choices to laryngectomized patients. Because of its ease of use, high success rate in generating speech, and quick training period, vocal prosthesis has become a popular way of rehabilitation [15]. Although effective, all established speech restoration techniques provide patients with distinctly distorted speech patterns, which are perceived as unhealthy by both the patient and society. This is due to the fact that substitution voicing generated speech features high irregularity, frequency shifts, and aperiodicity, together with frequent speech phonatory breaks [16]. This problem often becomes more apparent when the patient has to speak in a loud environment or over the phone [17]. Practitioners often rely on expert opinion on the perceived voice quality measurements, classification, and diagnosis of voice pathology. The problem is that often the procedure is time consuming and can be subject to parameter

sensitivity [18]. Latest digitization trends have pushed towards a major improvement in computer-assisted medical techniques. Thus, following established practice, the acoustic prosodic properties of the speech signal have to be modulated by a variety of health-related effects [19], leading to changes in a human voice and the automated detection of pathologies using machine learning has attracted significant medical attention [20].

Many approaches for detecting voice pathology have been proposed in recent research in the above-mentioned literature [21]. However, these systems only attempted to distinguish normal voices from diseased sounds, indicating that there is a research gap in terms of voice illness detection in relation to laryngeal cancer. There are circumstances in machine learning algorithms when speech signals cannot ensure high accuracy and cause time consumption in pathology monitoring systems. As a result, there is an urgent need for a research that highlights the most essential concerns and challenges confronting vocal pathology systems, as well as the importance of illness identification in voice pathology. To our knowledge, not much data on the application of artificial intelligence (AI) technologies for SV assessment exists in the literature (see Section 2). As a result, implementing AI-based models for objective assessment and classification of SV could potentially open up new avenues in research and clinical practice, paving the way for the development of a useful and reliable tool for evaluating SV following laryngeal oncosurgery. Existing deep learning voice analysis approaches generally tend to apply some form of recurrent gates for temporal voice signal analysis, these methods tend to suffer from poor performance and are notoriously difficult to train. It is noticeable, that there is no working AI prototype for SV assessment. As a result, using an AI-based models to objectively assess and classify SV could possibly open up new avenues for study and clinical use. To begin with, a well-designed algorithm might standardize SV evaluation across numerous oncology canters, allowing data sets in different patient groups to be simply compared. The same data sets could be used to improve the algorithm in the future. Instead of the existing methods, but not very efficient already applied methods, requiring prior medical knowledge for signal analysis, we aim to exploit modern areas of machine learning (deep learning) research to extract, measure and objectively classify substitution voicing and speech after laryngeal oncosurgery from the audio signal. The objective estimates obtained can be simplified and used by general practitioners and patients. This would be especially valuable when movement is limited or specialized medical centers are difficult to find, as it was during the peak of the COVID-19 pandemic. Last but not least, AI saves time and does not retire—the knowledge gained via its use is always available and does not expire.

In this paper, we propose using convolutional neural networks (CNNs), generally applied for image classification for the analysis of audio signals by transforming the audio signals waveform into Mels spectrogram and using it as an input in a re-purposed lightweight image classification network. This approach allowed us to achieve the overall accuracy of 89.47% with a simpler network architecture, allowing the approach to be used on computing devices having only Central Processing Unit (CPU) but without a dedicated Graphical Processing Unit (GPU) for the classification of subjects voice pathology.

The paper is structured as follows: Section 2 discusses the state-of-the-art works. The dataset used in this study and the deep neural architecture are described in Section 3. The experimental results are presented and analyzed in Section 4. Finally, the results of this study are discussed in Section 5. The paper concludes with Section 6.

2. State of the Art Analysis

A chaotic nature of the substitution voicing signal makes evaluation of substitution voicing improper or even impossible with standard methods of acoustic voice analysis used in clinical settings. Multiparametric models for evaluating voice quality and dysphonia severity are sufficiently reliable and valid because of their correlations to auditory-perceptual evaluation and high reliability and validity in voice pathology detection [22]. Currently, two multiparametric acoustic indices based on sustained vowels and on continuous speech analysis have gained popularity in research and clinical settings to objectively

estimate dysphonia: i.e., the Cepstral Spectral Index of Dysphonia (CSID) and the Acoustic Voice Quality Index (AVQI) [23,24]. Both indices may provide reasonable estimates of dysphonia severity and represent valid acoustic metrics for objectifying abnormal overall voice quality [25,26]. However, the use of these indices for assessing SV could be unreliable or technically impossible due to irregular and rather chaotic origin of SV signal. There is no data in the literature about the use of CSID for SV assessment. Only the recent study by van Sluis et al. [27] employed the AVQI to evaluate acoustic voice quality in patients who had undergone total laryngectomy. However, the authors noted that a specific AVQI cut-off value and the discriminative power of this index for SV (tracheoesophageal speech) after laryngeal oncosurgery have to be determined in future research studies. The AMPEX algorithm developed by Van Immerseel and Martens allows automatic reliable analysis of running speech, recognizing regularity patterns for pitch values <100 Hz and differentiating between noise and voicing at low frequencies [28]. Despite the feasibility of AMPEX as a tool for evaluating highly irregular speech has been supported by several studies, this algorithm has not yet gained wider clinical recognition [7,29].

Consequently, to perform automatic voice pathology classification and diagnosis, it is important to obtain reliable signal properties, which is essential for the reliability of the result. The clinical interpretation of vocal features is often conducted before the process of pathology detection [30]. Judging from the analysis of other studies, it is clear that from a technological point of view, many researchers distinguish signal processing functions such as Mel Frequency Coefficients, waveform packet transformations, others use multiple voice analysis tools for a variety of physiological and etiological reasons [31–33]. Multiple parameters are used to determine speech roughness, including height, vibration, and flicker, and other methods are often used, such as Harmonic to Noise Ratio, Normalized Noise Energy, and Smooth-to-Noise Ratio [34].

There are two types of possible features to analyze disease impact on voice/speech signal: temporal and spectral [35]. The temporal features (time-domain features) are used to extract and have an easy physical interpretation of a signal (energy, zero-crossing rate, maximum amplitude, minimum energy, time of the ending transient or Log-Attack-Time Descriptor) and are sensitive to articulation. The spectral features (frequency-based features) are obtained by converting the time-based signal into the frequency domain using the Fourier Transform. They might be more efficient for automatic classification because they are not dependent on articulation [36]. The most popular frequency descriptors are fundamental frequency, frequency components, spectral centroid, spectral flux, spectral density, irregularity of spectrum, brightness, etc. [37]. These features can be used to identify changing features in human speech, where the Mel Frequency Cepstral Coefficients are often used in human voice analysis [38]. Methodology from standard speech analysis could be adapted, i.e., using OpenSMILE features [39,40], Essential descriptors, MPEG7 descriptors, jAudio, YAAFE, Tsanas features [41], Wavelet Time Scattering features [42] and Random Forest supervised learning algorithms to detect the symptoms [43] and also to fuse information in the form of soft decisions, obtained using various audio feature sets from separate modalities [44]. In addition, Cepstral Separation Difference could be applied for quantification of speech impairment [45]. Feature extraction using signal-to-noise ratio, harmonic-to-noise ratio, glottal to noise excitation, vocal fold excitation ratio, and empirical mode decomposition excitation ratio methods with Random Forests and support vector machines for classification algorithms can also be used [46].

Alternative approaches could be adopted through Syllable-level Features, Low-Level Descriptor Features, Formant Features, Phonotactic Features with SVM classifier, features extracted using Principal Component Analysis and Linear Discriminant Analysis), SVM, Adaptive Boosting (AdaBoost), K-Nearest Neighbor (KNN) and Adaptive Resonance Theory-Kohonen Neural Network classifiers and the likes. In addition, dimensional reduction techniques such as linear discriminant analysis, principal component analysis, kernel PCA, feeder discriminant ratio, singular value decomposition, and so on are used to find suitable latent variables for classification [47]. Other researchers have taken into account the

characteristics of human voice and hearing systems. Aicha et al. [48] used glottal waveform with feature selection using PCA and classification using SVM. Fontes et al. proposed a low-complexity approach using correntropy spectral density [49]. MPEG-7 features are most commonly used for indexing video and audio media and were investigated for this purpose [50]. Hossain et al. have demonstrated that the low-level functions of MPEG-7 sound are effective in diagnosing pathological voice using support vector machines [51]. Vaziri et al. distinguished between a healthy voice and a pathological voice using nonlinear dynamics performance and voice acoustic disturbances [52].

A wide variety of statistical, machine learning based, and other types of algorithms are now widely used for the detection of pathological voice based on the computed acoustic features of the input signal [53]. Pathology classification methods can be sorted into two categories [54]. The first category is "classical" methods, often based on k-nearest neighbor methods and Hilbert-Huang Transforms [55], random forests [56], support vector machines [57], Gaussian mixture models [58], latent Markov models [59], Dynamic time warping [60], discriminative paraconsistent machines [61] and so on. Often these methods are used in combination with traditional features, as illustrated by Ghulam et al., who singled out MFCC from long-term voice voice samples as characteristics and found a significant increase in accuracy in diagnosing pathological voices using the Gaussian mixture model [62]. Other researchers treated voice signals as normal vibration signals when classifying, e.g., Cordeiro et al. calculated the spectral envelope peaks of the voice signal as a function of the classification of pathological voices [63]. Alternatively, Saeedi et al. proposed a pathological voice recognition method based on wave transformation, which calculated the parameters of a wave filter bank using a genetic algorithm [64].

"Modern" side of pathology detection is often related to traditional dense neural networks [65], the more advanced CNNs [66] and very popular recurrent neural networks [67]. Deep learning, which transforms intelligent signal analysis so that algorithms can under certain conditions, theoretically might reach near-medical (expert) capabilities in a variety of voice pathology classification tasks. Chen et al. used 12 Mel frequency cepstral coefficients of each voice sample as row features for their deep learning implementation [68]. Miliaresi et al. suggest to analyze various properties of the voice signal window as low-level descriptors (LLDs) by extracting and analyzing variable-length fragments from the speech signal using the prisms of the main tone, energy, and spectrum [69] and using this data to train the deep learning models. Furthermore, a number of functional elements, such as moments, extremes, percentiles, and regression parameters, will then be applied to each LLD [70], to form a set of aggregate features for a healthy and unhealthy human voice. These statistical summaries can also be combined to form tensors for the training of AI (deep learning) algorithms, where multipath learning and learning transfer could be applied according to the multifunctional LSTM-RNN paradigm [71]. Kim et al. [72] collected features from voice samples of a vowel sound of /a:/ and computed the Mel-frequency cepstral coefficients (MFCCs) using the software package for speech analysis in phonetics (PRAAT), which were used identify between patients with laryngeal cancer and healthy controls. Depending on the features extracted, some authors suggest to an investigation of [53]. Alternatively, it is possible to try to introduce kernel-based extreme learning machines [73] and data preprocessing [74]. Or involves a combination of the k-means clustering-based feature weighting method and a complex-valued artificial neural network [75].

3. Materials and Methods

3.1. Clinical Evaluation and Equipment

All participants of the study were evaluated by clinical voice specialists performing video laryngostroboscopy (VLS) at the Department of Otorhinolaryngology of the Lithuanian University of Health Sciences (LUHS), Kaunas, Lithuania. VLS was performed using the XION EndoSTROB DX device (XION GmbH, Berlin, Germany) with a 70° rigid endoscope. VLS is routine in clinical practice and did not cause any additional discomfort or delays for the participants.

Speech recordings of the phonetically balanced Lithuanian sentence 'Turėjo senelė žilą oželį' ('The grandmother had a little grey goat') were obtained using a T-series silent room for hearing testing (T-room, CA Tegner AB, Bromma, Sweden) via a D60S Dynamic Vocal (AKG Acoustics, Vienna, Austria) microphone placed 10.0 cm from the mouth with an about 90° microphone-to-mouth angle. Speech recordings were made at a rate of 44,100 samples per second and exported as uncompressed 16-bit deep WAV audio files.

3.2. Dataset

A database of digital speech recordings of 367 male subjects (279 normal speech samples and 88 pathological speech samples) was used. Subjects' age ranged from 18 to 80 years. The control group comprised 279 healthy male volunteers (mean age 38.1 ± 12.7 years) with the voices evaluated as healthy by the clinical voice specialists. The control group (class 0) subjects had no present or preexisting speech, neurological, hearing, or laryngeal disorders and were free of common cold or upper respiratory infection at the time of speech recording. Furthermore, no pathological alterations in the larynx of the subjects of the normal voice subgroup group were found during VLS. The pathological speech subgroup consisted of 88 (64.1 ± 6.9 years) male patients who used substitution voicing (SV) after oncosurgery. This subgroup included 43 patients after extended cordectomy (class 1), 17 patients after partial vertical laryngectomy (class 2), and 28 patients after total laryngectomy who used tracheoesophageal prosthesis (TEP) (class 3). The pathological speech subgroup patients were recruited from consecutive patients who were diagnosed with the before-mentioned conditions. Speech recordings were obtained at least 6 months after the surgery to ensure a reasonable amount of time for the laryngeal tissue to heal and speech rehabilitation programs to end. A comparison cochleagrams of each class are illustrated in Figure 1. We use the cochleagrams of sound signals for time-frequency analysis and feature extraction instead of the more traditional spectrograms. The signal is initially passed via a gammatone filter, which is designed to mimic the auditory filters found in the human cochlea. The filtered signal is then divided into small windows, with the energy in each window summed and normalized to produce the cochleagram image's intensity values.

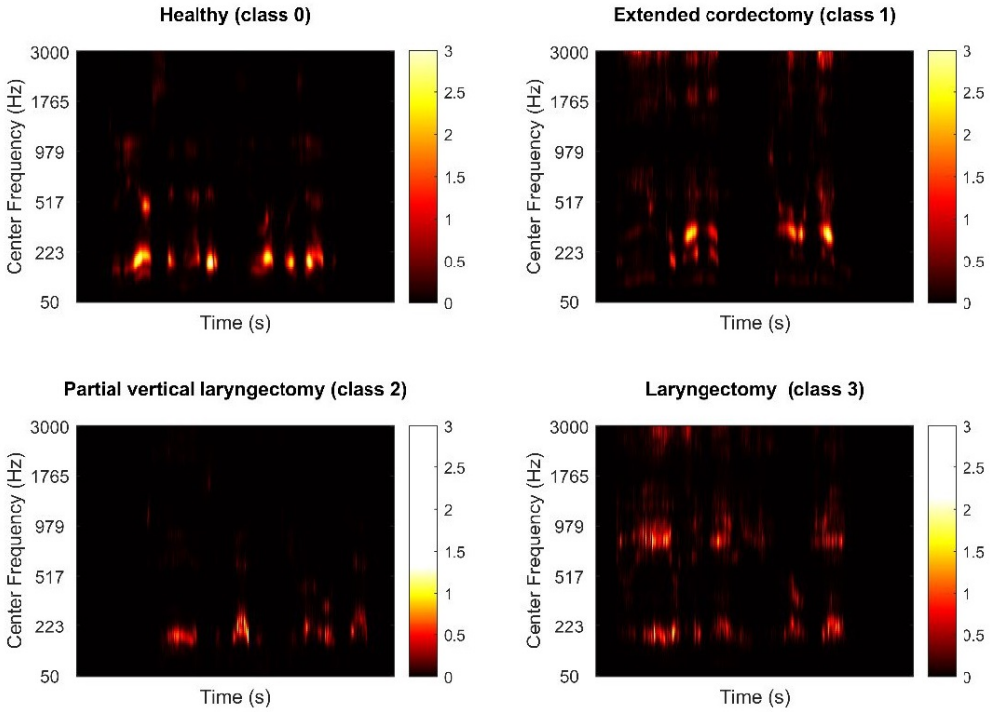

Figure 1. Cochleagrams of each class.

3.3. Data Analysis

Table 1 summarizes the voice features captured in the dataset.

Table 1. Summary of voice features.

Feature	Description
PVF	Percentage of voiced frames
PVS	Percentage of voiced speech frames
AVE	Mean voicing evidence of voiced frames
PVFU	Percentage of unreliable voiced frames
MD	Average F0 modulation
MDc	MD only in frames with a "reliable" F0 estimate. Vocal frequency estimate F0 is considered reliable if it deviates less than 25% from the average over all voiced frames.
Jitter	F0-jitter in all voiced frame pairs (=2 consecutive frames)

Figure 2 shows the histograms of database feature value distributions among classes. The analysis was supported by one-way ANOVA statistical test, which revealed statistically significant differences between classes in PVF ($p < 0.001$), PVS ($p < 0.001$), AVE ($p < 0.001$), PVFU ($p < 0.001$), MD ($p < 0.001$), MDc ($p < 0.01$), and Jitter ($p < 0.001$) values. There was no statistically significant difference in Tmax values.

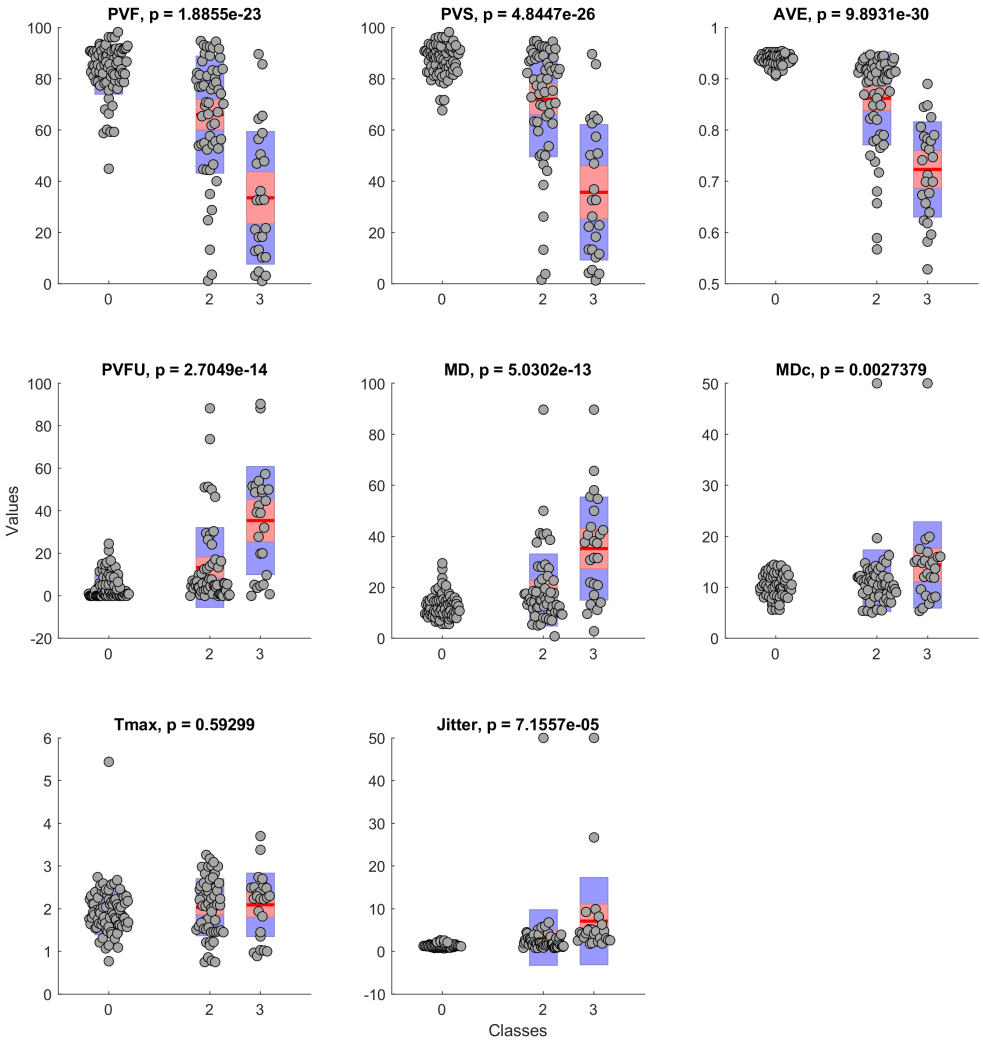

Figure 2. Histogram of feature value distribution among classes with *p*-value from ANOVA test.

Figure 3 shows the correlation between feature values among classes in database. The strong correlation was found between PVS and PVF ($R = 0.963$, $p < 0.001$), PVS and AVE ($R = 0.942$, $p < 0.001$), and MD and PVFU ($R = 0.898$, $p < 0.001$). This shows a strong co-linearity property in the database, which makes it difficult to use for training classical machine learning models [76].

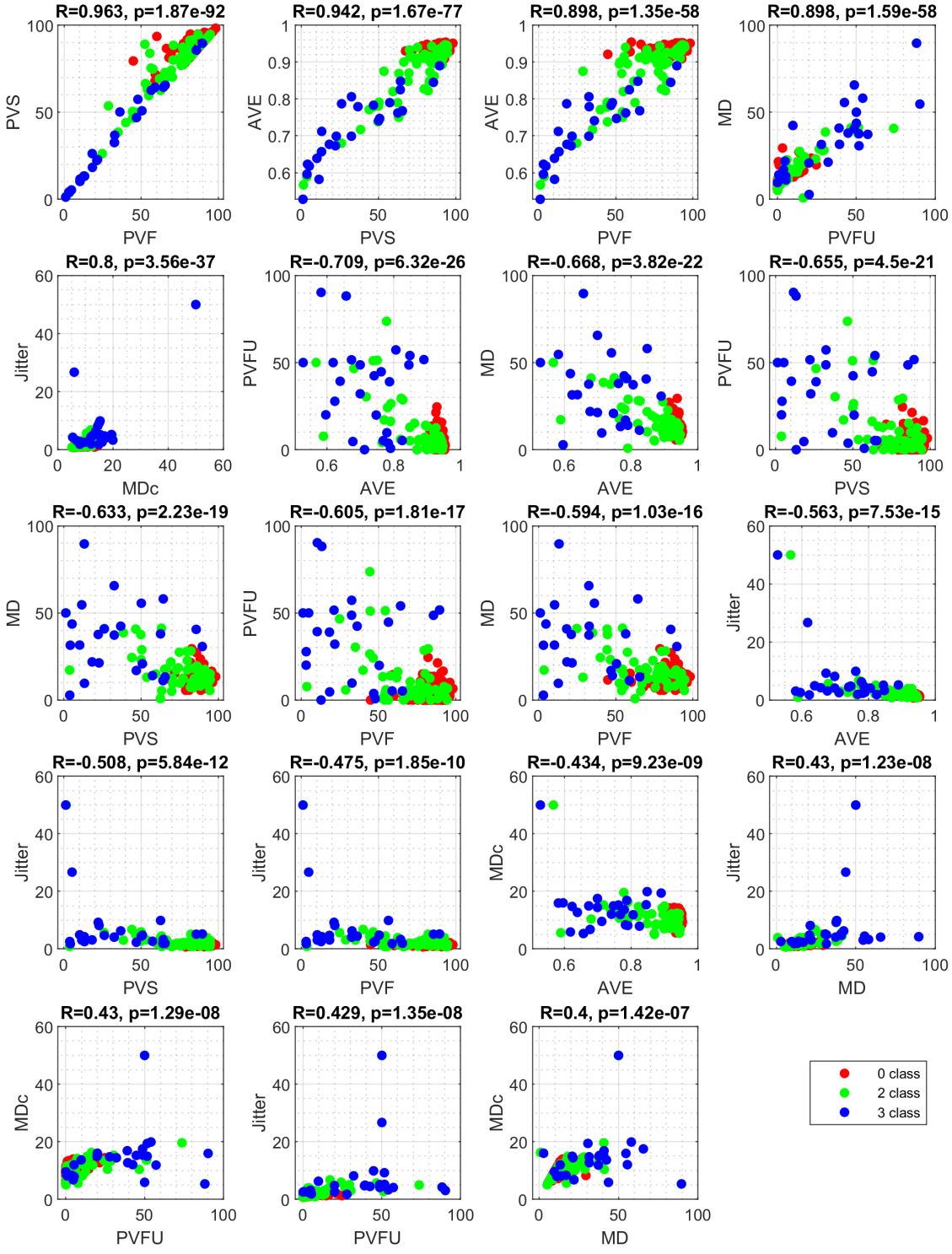

Figure 3. Correlation between feature values among classes. Correlation value (R) and its significance (p) are given. The plots are arranged by decreasing statistical significance of the determination coefficient (R^2). Only plots with significant correlations are shown.

3.4. Architecture

Figure 4 shows our approach deep neural network architecture. Our approach takes an input of Mel-frequency spectrogram (MFCC) as an input with a total of 80 coefficients. Therefore, given a waveform, the converted MFCC spectroctrogram gives an input of $N \times 80 \times 1$ where N is the sequence length. Each of the layer blocks starts with a convolutional network with stride 2, this reduces the input dimensionality by half. Layers 2, 3 and 4 internally contain skip connections (dashed lines), these allow for a better gradient flow. The fourth and final layer is then connected to fully-connected that has 4 neuron output, each of the neurons is belongs to one of four voice classes. The network is trained using initial learning rate of $lr = 10^{-4}$ with the batch size of $n = 16$, to reduce memory requirements training was performed on half-precision floating points. Because the sequence length between the audio files was not equal the each of the batch audio files have been padded with zeroes to equalize the sequence length. The network was trained for 3000 epochs using Adam optimizer [77] and cosine annealing with warm restarts every 500 epochs, which would adjust the learning rate in the range of $lr = [10^{-7}; 10^{-4}]$, cosine annealing was chosen for it has demonstrated the ability to achieve better recall rates due to potentially jumping out of local minimums [78]. The hyper-parameter values were chosen during empirical experiments. Over-fitting was avoided by employing an early stopping process and batch normalization.

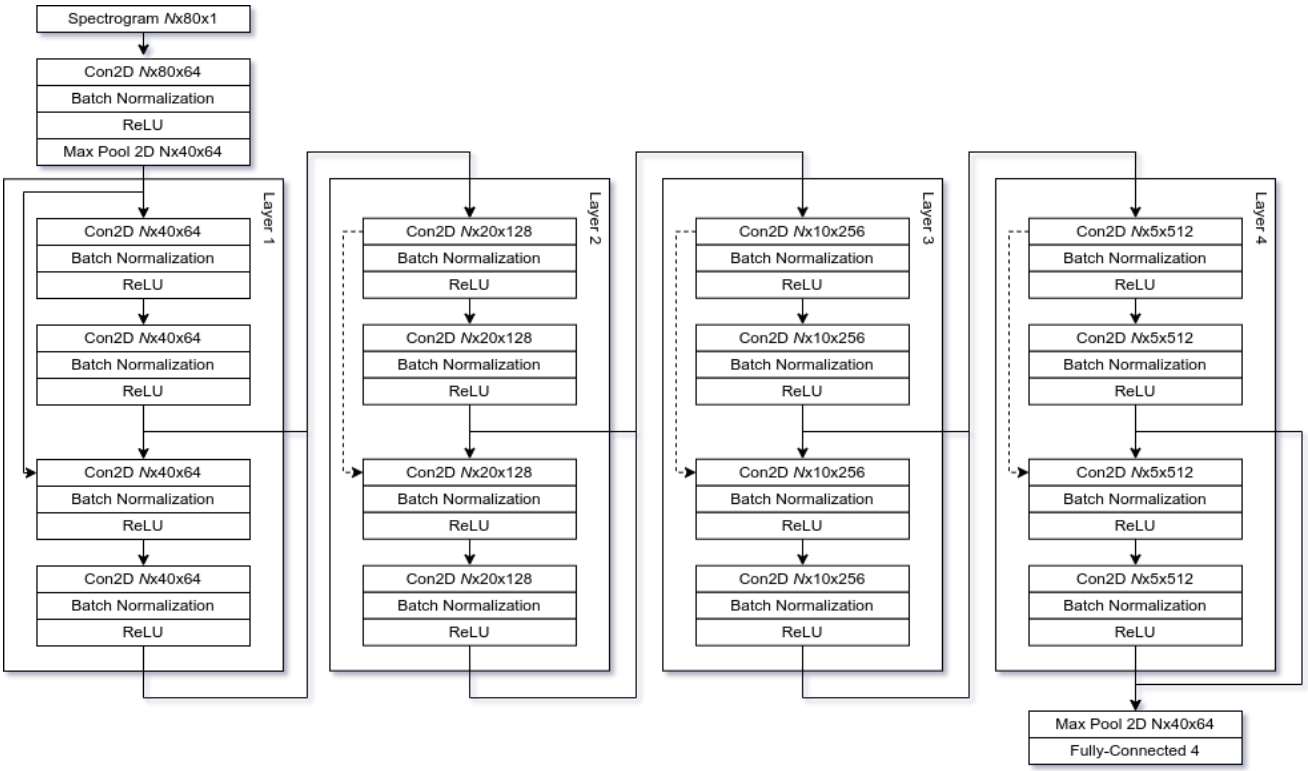

Figure 4. Our approach, here N is the sequence length, dashed lines are skip connections.

3.5. Implementation

In Figures 5–7 we can see how our approach works for evaluating subject's voice class. In order for the subject to evaluate their voice, firstly they need to make a voice recording using their microphone, the audio waveform is sampled using mono-channel 8000 Hz sampling rate (as 8 kHz still retains voice information (as stipulated by most standards, including telephony), a down-sampling (from 44 kHz to 8 kHz) was performed to optimize the required quantity of data and reduce network overhead while taking VRAM limits into account.). After the voice waveform is recorded, it is then converted into Mels-frequency diagram using 80 coefficients. Normally, this would be around 40 MFCC samples, however

the system kept too little information in our situation (as substitution voicing loses a lot of information in relation to "healthy" speech), therefore 80 MFCC samples was the best determined option. The MFCC spectrogram is then used as an input in our neural network, where one of four classes are predicted: healthy, one-voice fold pathology, two-voice fold pathology, and finally nonspecific voice pathology.

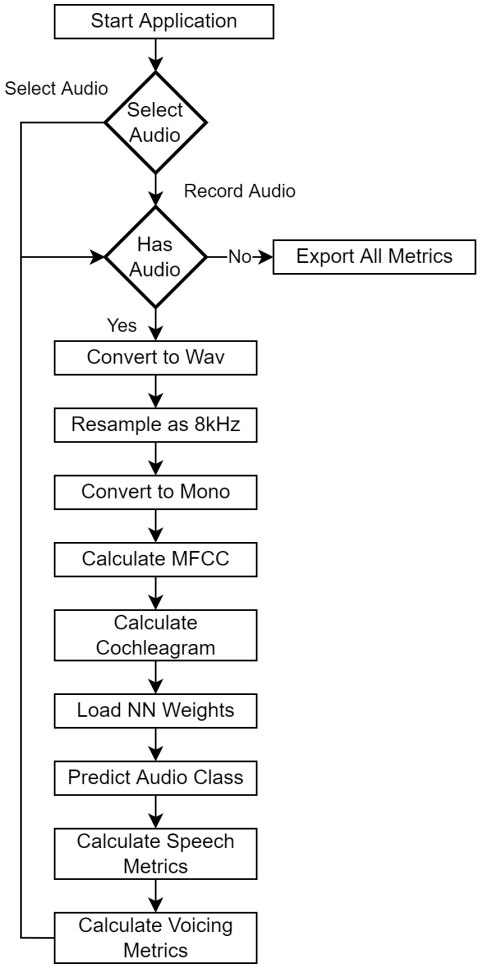

Figure 5. Architecture of the system.

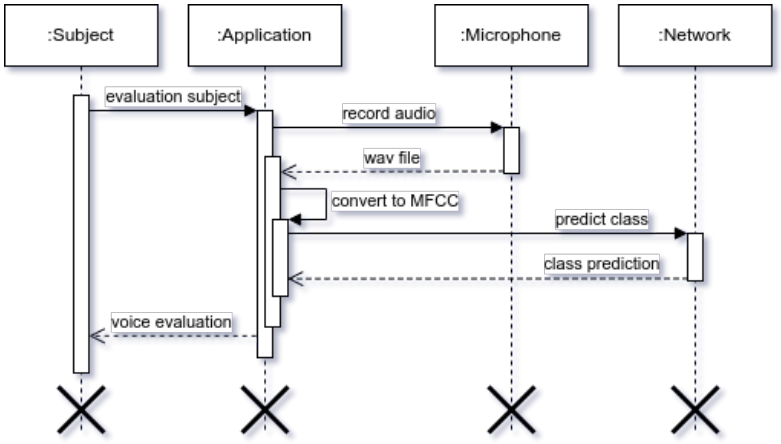

Figure 6. Voice evaluation sequence diagram.

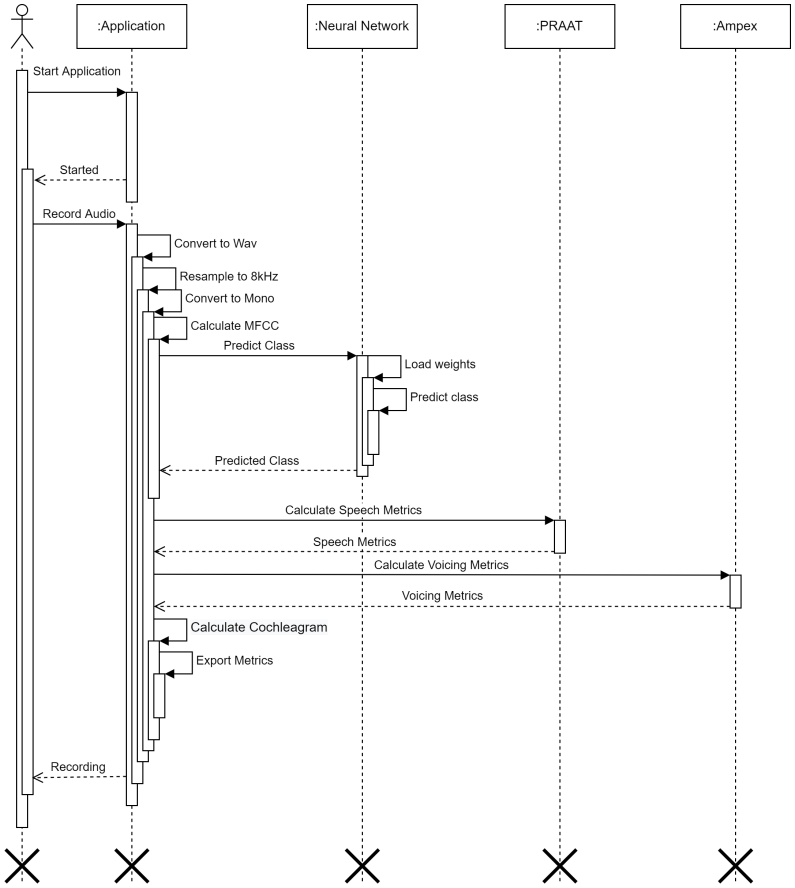

Figure 7. Composition of the voice evaluation sequence processes.

4. Experimental Evaluation and Results

4.1. Setup

To test our minimalistic CPU optimized approach, we have used augmented the dataset and used 147 recordings containing no voice pathology (normal voice), 111 voice recordings of mass lesions of one single vocal fold, 57 recordings of mass lesions in both vocal folds, and finally 67 recordings containing nonspecific voice pathology from the dataset collected in Lithuanian University of Health Sciences (see Section 3.2). The training set is divided using 80:20 rule, where 80% of the recordings of each class separately are used for training, and the remaining are used for validation. Additionally, because the dataset is highly unbalanced, we have dropped the data points in classes that have an excess of data, this allows all classes to have an identical amount of data samples, reducing the probability that the network will overfit using any of the underlying classes. To evaluate and compare our approach versus state of the art, we have used confusion matrices as they best reflect the results in multiclass problems by allowing us to evaluate true-positive versus false-positive rates.

4.2. Metrics

We used accuracy, precision, recall and F1-score as fitness measures. These are defined as follows:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \times 100\% \quad (1)$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (2)$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (3)$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}. \quad (4)$$

where TP (true positives) is the number of voice pathology samples that were labeled correctly, TN (true negatives) is the number of non-pathotology voice samples that were labeled correctly. FP (false positives) is the number of voice pathology samples that were labeled incorrectly as being not voice pathology samples, and FN (false negatives) is the number of not-pathology samples that were miss classified as pathology samples.

4.3. Results

In addition to our approach, we have tested three additional approaches, ResNet-101 [79], a state-of-the-art image classification network, Wav2Letter [80] and M5 [81] as state-of-the-art audio classification networks using the identical training procedure and datasets. The confusion matrices for our approach can be seen in Figure 8, for ResNet-101 can be seen in Figure 9, Wav2Letter in Figure 10, and finally M5 confusion matrix can be seen in Figure 11. Here Class 0 represents normal voice; Class 1 represents SV after cordectomy; Class 2 represents SV after partial laryngectomy; Class 3 represents SV using TEP. As we can see, our approach has shown the best true positive rate of any of the compared state-of-the-art approaches. Giving an overall accuracy of 89.47%.

	Predicted 0	Predicted 1	Predicted 2	Predicted 3
True 0	93.10%	3.45%	0.00%	3.45%
True 1	8.70%	82.61%	8.70%	0.00%
True 2	0.00%	9.09%	90.91%	0.00%
True 3	7.69%	0.00%	0.00%	92.31%

Figure 8. Confusion Matrix for our approach.

	Predicted 0	Predicted 1	Predicted 2	Predicted 3
True 0	86.21%	10.34%	0.00%	3.45%
True 1	21.74%	78.26%	0.00%	0.00%
True 2	0.00%	45.45%	54.55%	0.00%
True 3	15.38%	0.00%	0.00%	84.62%

Figure 9. Confusion Matrix for ResNet-101 model.

	Predicted 0	Predicted 1	Predicted 2	Predicted 3
True 0	34.48%	62.07%	3.45%	0.00%
True 1	8.70%	91.30%	0.00%	0.00%
True 2	9.09%	36.36%	54.55%	0.00%
True 3	0.00%	100.00%	0.00%	0.00%

Figure 10. Confusion Matrix for Wav2Letter model.

	Predicted 0	Predicted 1	Predicted 2	Predicted 3
True 0	55.17%	20.69%	0.00%	24.14%
True 1	8.70%	56.52%	8.70%	26.09%
True 2	0.00%	54.55%	45.45%	0.00%
True 3	23.08%	7.69%	0.00%	69.23%

Figure 11. Confusion Matrix for M5 model.

In Figure 12 we can see the model accuracy comparison side-by-side for each of the approaches broken down by class, additionally we can see our approach result breakdown in Table 2, as we can see, the accuracy for all of each of the individual classes is above 90%.

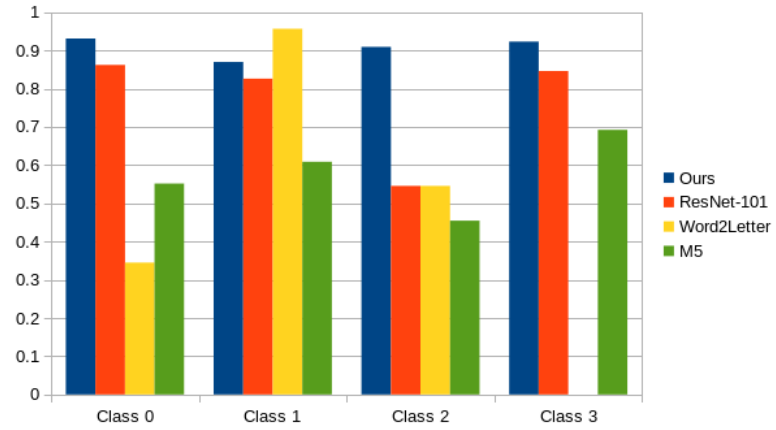

Figure 12. Comparison of performance between different models: ResNet-101, Word2Letter, M5 and our model.

Table 2. Our result approach breakdown by class.

Class	n (Truth)	n (Classified)	Accuracy	Precision	Recall	F1 Score
0—normal voice	30	29	93.42%	0.93	0.9	0.92
1—SV after cordectom	21	23	92.11%	0.83	0.9	0.86
2—SV after partial laryngectomy	12	11	96.05%	0.91	0.83	0.87
3—SV using TEP	13	13	97.37%	0.92	0.92	0.92

To analyze the predictions of models more precisely, we used t-distributed stochastic neighbor embedding (t-SNE), a statistical method for visualizing high-dimensional data by mapping it to a two-dimensional embedding. The results are presented in Figure 13. It shows that the classes are well separated while the miss-classifications using the best model (resnet18) are few.

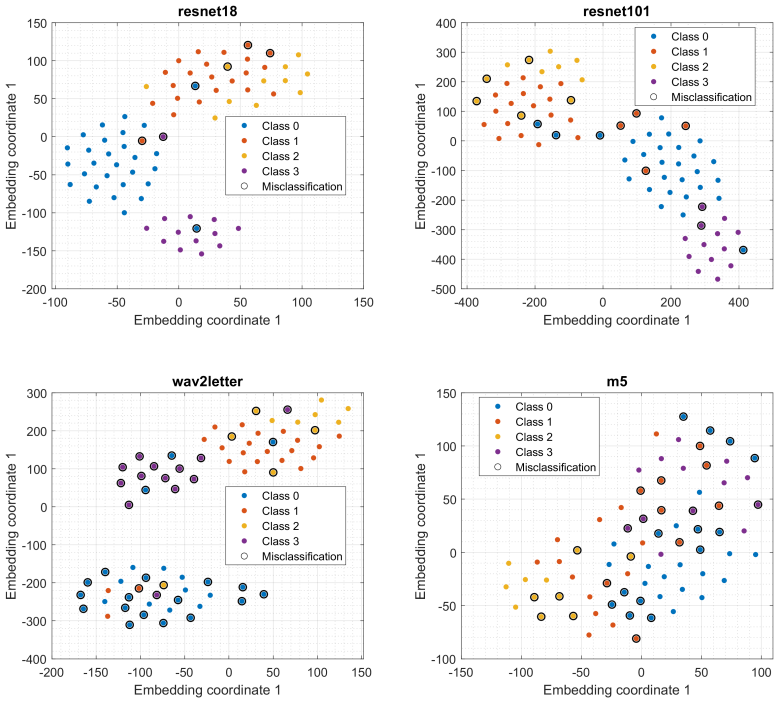

Figure 13. Comparison between t-SNE embeddings of different model predictions.

5. Discussion

This work provides a technique for automatically assessing if a voice is healthy or whether its quality has changed owing to a pathological condition. Because these spread swiftly, automatic detection is necessary, yet it is frequently underestimated. Machine learning is making a significant contribution to illness diagnosis and early detection in cardiology, pulmonology, liver tumor segmentation, and other fields of healthcare. As a consequence, machine learning might be employed effectively in a computer or mobile healthcare system to automatically identify and detect irregularities in a person's speech for early diagnosis.

For the study of speech audio signals, we propose employing well-known CNN models that have been used for image classification. Our method uses a Mel-frequency spectrogram (MFCC) as an input to a deep neural network architecture while achieving very good classification results. Our outcomes demonstrate that a deep learning model after training using a pathological speech database, voice alone might be utilized for common vocal fold illness identification using a deep learning technique. This AI-based technique might be therapeutically effective for screening general vocal fold illness using the voice. A brief assessment and a general health examination are part of the strategy. It can be used during telemedicine in places where primary care facilities do not have laryngoscopic capabilities. It might aid physicians in pre-screening patients by allowing invasive exams to be done only in situations involving issues with automatic recognition or listening, as well as expert evaluations of other clinical examination findings that raise concerns about the existence of diseases.

The biggest issue that each patient suffers, especially those who live in distant areas, is the lack of physicians and care in emergency circumstances. As a result, there is a need to provide a new framework in such remote locations by utilizing telecommunication means and artificial intelligence methods for automated voice analysis in the context of remotely-provided telehealth services [82]. Telehealth is a successful paradigm for diagnosing and treating voice issues in remote locations, as an alternative to face-to-face consultations. Telehealth consultations have been found to contribute to medical diagnosis for a variety of vocal problems, with diagnostic decision outcomes comparable to in-person consultations [83]. There are several instances in which patients require long-term monitoring. In this sense, the provision of continuous monitoring is critical. Because laryngeal cancer is a potentially fatal disease, new and effective methods for laryngeal cancer early detection are desperately needed. The method provided in this study enables an effective and noninvasive way for diagnosing laryngeal carcinoma.

6. Conclusions

In this paper we used cutting-edge deep learning research to objectively categorize, extract, and assess substitution voicing after laryngeal oncosurgery from audio signals. For the study of speech audio signals, we propose employing well-known CNNs that have been used for image classification. Our method uses a Mel-frequency spectrogram as an input to a deep neural network architecture. A database of 367 male participants' digital voice recordings (279 normal speech samples and 88 abnormal speech samples) was employed. Our method has the highest true-positive rate of any of the assessed state-of-the-art methods, with an overall accuracy of 89.47%.

Author Contributions: All authors have contributed equally to this manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research has received funding from European Regional Development Fund (project No. 13.1.1-LMT-K-718-05-0027) under grant agreement with the Research Council of Lithuania (LMTLT). Funded as European Union's measure in response to COVID-19 pandemic.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The dataset used in this study is available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Steuer, C.E.; El-Deiry, M.; Parks, J.R.; Higgins, K.A.; Saba, N.F. An update on larynx cancer. *CA A Cancer J. Clin.* **2017**, *67*, 31–50. [CrossRef] [PubMed]
2. Sass, V.; Gadeyne, S. Social Disparities in Survival from Head and Neck Cancers in Europe. In *Social Environment and Cancer in Europe*; Launoy, G., Zadnik, V., Coleman, M.P., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 141–158. [CrossRef]
3. American Cancer Society. Key Statistics for Laryngeal and Hypopharyngeal Cancers. *Cancer.org* **2022**. Available online: https://www.cancer.org/cancer/laryngeal-and-hypopharyngeal-cancer/about/key-statistics.html (accessed on 20 January 2022).
4. Groome, P.A.; O'Sullivan, B.; Irish, J.C.; Rothwell, D.M.; Schulze, K.; Warde, P.R.; Schneider, K.M.; Mackenzie, R.G.; Hodson, D.I.; Hammond, J.A.; et al. Management and Outcome Differences in Supraglottic Cancer Between Ontario, Canada, and the Surveillance, Epidemiology, and End Results Areas of the United States. *J. Clin. Oncol.* **2003**, *21*, 496–505. [CrossRef] [PubMed]
5. Anthony Jnr, B. Implications of telehealth and digital care solutions during COVID-19 pandemic: A qualitative literature review. *Inf. Health Soc. Care* **2021**, *46*, 68–83. [CrossRef] [PubMed]
6. Sharifi, M.; Asadi-Pooya, A.A.; Mousavi-Roknabadi, R.S. Burnout among Healthcare Providers of COVID-19; a Systematic Review of Epidemiology and Recommendations. *Arch. Acad. Emerg. Med.* **2021**, *9*, e7. [CrossRef]
7. Moerman, M.; Martens, J.P.; Dejonckere, P. Multidimensional assessment of strongly irregular voices such as in substitution voicing and spasmodic dysphonia: A compilation of own research. *Logop. Phoniatr. Vocol.* **2015**, *40*, 24–29. [CrossRef]
8. Semple, C.; Parahoo, K.; Norman, A.; McCaughan, E.; Humphris, G.; Mills, M. Psychosocial interventions for patients with head and neck cancer. *Cochrane Database Syst. Rev.* **2013**. [CrossRef]
9. Kumar, V.; Singh, D.; Kaur, M.; Damaševičius, R. Overview of Current State of Research on the Application of Artificial Intelligence Techniques for COVID-19. *PeerJ Comput. Sci.* **2021**, *7*, 1–34. [CrossRef]
10. Thomas, A.; Manchella, S.; Koo, K.; Tiong, A.; Nastri, A.; Wiesenfeld, D. The impact of delayed diagnosis on the outcomes of oral cancer patients: A retrospective cohort study. *Int. J. Oral Maxillofac. Surg.* **2021**, *50*, 585–590. [CrossRef]
11. Noel, C.W.; Li, Q.; Sutradhar, R.; Eskander, A. Total Laryngectomy Volume During the COVID-19 Pandemic: Looking for Evidence of Stage Migration. *JAMA Otolaryngol. Neck Surg.* **2021**, *147*, 909. [CrossRef]
12. Singh, A.; Bhardwaj, A.; Ravichandran, N.; Malhotra, M. Surviving COVID-19 and multiple complications post total laryngectomy. *BMJ Case Rep. CP* **2021**, *14*, e244277. doi: 10.1136/bcr-2021-244277. [CrossRef]
13. Pereira da Silva, A.; Feliciano, T.; Vaz Freitas, S.; Esteves, S.; Almeida e Sousa, C. Quality of Life in Patients Submitted to Total Laryngectomy. *J. Voice* **2015**, *29*, 382–388. [CrossRef] [PubMed]
14. Zilcha-Mano, S.; Goldstein, P.; Dolev-Amit, T.; Ben David-Sela, T.; Barber, J.P. A randomized controlled trial for identifying the most suitable treatment for depression based on patients' attachment orientation. *J. Consult. Clin. Psychol.* **2021**, *89*, 985–994. [CrossRef] [PubMed]
15. Brook, I.; Goodman, J.F. Tracheoesophageal Voice Prosthesis Use and Maintenance in Laryngectomees. *Int. Arch. Otorhinolaryngol.* **2020**, *24*, e535–e538. [CrossRef] [PubMed]
16. Mattys, S.L.; Davis, M.H.; Bradlow, A.R.; Scott, S.K. Speech recognition in adverse conditions: A review. *Lang. Cogn. Process.* **2012**, *27*, 953–978. [CrossRef]
17. Uscher-Pines, L.; Sousa, J.; Raja, P.; Mehrotra, A.; Barnett, M.L.; Huskamp, H.A. Suddenly Becoming a "Virtual Doctor": Experiences of Psychiatrists Transitioning to Telemedicine During the COVID-19 Pandemic. *Psychiatr. Serv.* **2020**, *71*, 1143–1150. [CrossRef] [PubMed]
18. Hossain, M.S.; Muhammad, G.; Alamri, A. Smart healthcare monitoring: A voice pathology detection paradigm for smart cities. *Multimed. Syst.* **2019**, *25*, 565–575. [CrossRef]
19. Cummins, N.; Baird, A.; Schuller, B.W. Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods* **2018**, *151*, 41–54. [CrossRef]
20. Lee, J.Y. Experimental Evaluation of Deep Learning Methods for an Intelligent Pathological Voice Detection System Using the Saarbruecken Voice Database. *Appl. Sci.* **2021**, *11*, 7149. [CrossRef]
21. Al-Dhief, F.T.; Latiff, N.M.A.; Malik, N.N.N.A.; Salim, N.S.; Baki, M.M.; Albadr, M.A.A.; Mohammed, M.A. A Survey of Voice Pathology Surveillance Systems Based on Internet of Things and Machine Learning Algorithms. *IEEE Access* **2020**, *8*, 64514–64533. [CrossRef]
22. Barsties, B.; De Bodt, M. Assessment of voice quality: Current state-of-the-art. *Auris Nasus Larynx* **2015**, *42*, 183–188. [CrossRef]
23. Awan, S.N.; Roy, N.; Dromey, C. Estimating dysphonia severity in continuous speech: Application of a multi-parameter spectral/cepstral model. *Clin. Linguist. Phon.* **2009**, *23*, 825–841. [CrossRef] [PubMed]
24. Maryn, Y.; De Bodt, M.; Roy, N. The Acoustic Voice Quality Index: Toward improved treatment outcomes assessment in voice disorders. *J. Commun. Disord.* **2010**, *43*, 161–174. [CrossRef] [PubMed]

25. Barsties v. Latoszek, B.; Mathmann, P.; Neumann, K. The cepstral spectral index of dysphonia, the acoustic voice quality index and the acoustic breathiness index as novel multiparametric indices for acoustic assessment of voice quality. *Curr. Opin. Otolaryngol. Head Neck Surg.* **2021**, *29*, 451–457. [CrossRef]
26. Lee, J.M.; Roy, N.; Peterson, E.; Merrill, R.M. Comparison of Two Multiparameter Acoustic Indices of Dysphonia Severity: The Acoustic Voice Quality Index and Cepstral Spectral Index of Dysphonia. *J. Voice* **2018**, *32*, 515-e1. [CrossRef] [PubMed]
27. van Sluis, K.E.; van Son, R.J.J.H.; van der Molen, L.; MCGuinness, A.J.; Palme, C.E.; Novakovic, D.; Stone, D.; Natsis, L.; Charters, E.; Jones, K.; et al. Multidimensional evaluation of voice outcomes following total laryngectomy: A prospective multicenter cohort study. *Eur. Arch.-Oto-Rhino-Laryngol.* **2020**, *278*, 1209–1222. [CrossRef]
28. Manfredi, C.; Giordano, A.; Schoentgen, J.; Fraj, S.; Bocchi, L.; Dejonckere, P. Validity of jitter measures in non-quasi-periodic voices. Part II: The effect of noise. *Logop. Phoniatr. Vocol.* **2011**, *36*, 78–89. [CrossRef]
29. Dejonckere, P.H.; Moerman, M.B.J.; Martens, J.P.; Schoentgen, J.; Manfredi, C. Voicing quantification is more relevant than period perturbation in substitution voices: An advanced acoustical study. *Eur. Arch.-Oto-Rhino-Laryngol.* **2012**, *269*, 1205–1212. [CrossRef]
30. Muhammad, G.; Alhamid, M.; Hossain, M.; Almogren, A.; Vasilakos, A. Enhanced Living by Assessing Voice Pathology Using a Co-Occurrence Matrix. *Sensors* **2017**, *17*, 267. [CrossRef]
31. Jiang, J.; Li, Y. Review of active noise control techniques with emphasis on sound quality enhancement. *Appl. Acoust.* **2018**, *136*, 139–148. [CrossRef]
32. Avila, A.R.; Gamper, H.; Reddy, C.; Cutler, R.; Tashev, I.; Gehrke, J. Non-intrusive Speech Quality Assessment Using Neural Networks. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 631–635. [CrossRef]
33. Gamper, H.; Reddy, C.K.A.; Cutler, R.; Tashev, I.J.; Gehrke, J. Intrusive and Non-Intrusive Perceptual Speech Quality Assessment Using a Convolutional Neural Network. In Proceedings of the 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 20–23 October 2019; pp. 85–89. [CrossRef]
34. v. Latoszek, B.B.; Maryn, Y.; Gerrits, E.; Bodt, M.D. A Meta-Analysis: Acoustic Measurement of Roughness and Breathiness. *J. Speech Lang. Hear. Res.* **2018**, *61*, 298–323. [CrossRef]
35. Muhammad, G.; Melhem, M. Pathological voice detection and binary classification using MPEG-7 audio features. *Biomed. Signal Process. Control* **2014**, *11*, 1–9. [CrossRef]
36. Yin, D.; Luo, C.; Xiong, Z.; Zeng, W. PHASEN: A Phase-and-Harmonics-Aware Speech Enhancement Network. *AAAI Conf. Artif. Intell.* **2020**, *34*, 9458–9465. [CrossRef]
37. Yuanbo, W.; Changwei, Z.; Ziqi, F.; Yihua, Z.; Xiaojun, Z.; Zhi, T. Voice Pathology Detection and Multi-classification Using Machine Learning Classifiers. In Proceedings of the 2020 International Conference on Sensing, Measurement Data Analytics in the Era of Artificial Intelligence (ICSMD), Xi'an, China, 15–17 Ocober 2020; pp. 319–324. [CrossRef]
38. Fang, S.H.; Tsao, Y.; Hsiao, M.J.; Chen, J.Y.; Lai, Y.H.; Lin, F.C.; Wang, C.T. Detection of Pathological Voice Using Cepstrum Vectors: A Deep Learning Approach. *J. Voice* **2019**, *33*, 634–641. [CrossRef] [PubMed]
39. Guimaraes, M.T.; Medeiros, A.G.; Almeida, J.S.; Falcao Y Martin, M.; Damasevicius, R.; Maskeliunas, R.; Cavalcante Mattos, C.L.; Reboucas Filho, P.P. An Optimized Approach to Huntington's Disease Detecting via Audio Signals Processing with Dimensionality Reduction. In Proceedings of the International Joint Conference on Neural Networks, Glasgow, UK, 3 Ocober 2020.
40. Narendra, N.; Alku, P. Automatic assessment of intelligibility in speakers with dysarthria from coded telephone speech using glottal features. *Comput. Speech Lang.* **2021**, *65*, 101117. [CrossRef]
41. Arora, S.; Tsanas, A. Assessing Parkinson's Disease at Scale Using Telephone-Recorded Speech: Insights from the Parkinson's Voice Initiative. *Diagnostics* **2021**, *11*, 1892. [CrossRef]
42. Lauraitis, A.; Maskeliunas, R.; Damaševičius, R.; Krilavičius, T. Detection of Speech Impairments Using Cepstrum, Auditory Spectrogram and Wavelet Time Scattering Domain Features. *IEEE Access* **2020**, *8*, 96162–96172. [CrossRef]
43. Braga, D.; Madureira, A.M.; Coelho, L.; Ajith, R. Automatic detection of Parkinson's disease based on acoustic analysis of speech. *Eng. Appl. Artif. Intell.* **2019**, *77*, 148–158. [CrossRef]
44. Qian, Y.; Chen, Z.; Wang, S. Audio-Visual Deep Neural Network for Robust Person Verification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1079–1092. [CrossRef]
45. Patil, A.T.; Patil, H.A.; Khoria, K. Effectiveness of energy separation-based instantaneous frequency estimation for cochlear cepstral features for synthetic and voice-converted spoofed speech detection. *Comput. Speech Lang.* **2022**, *72*, 101301. [CrossRef]
46. Jalali-najafabadi, F.; Gadepalli, C.; Jarchi, D.; Cheetham, B.M. Acoustic analysis and digital signal processing for the assessment of voice quality. *Biomed. Signal Process. Control* **2021**, *70*, 103018. [CrossRef]
47. Jothi, K.R.; Sivaraju, S.S.; Yawalkar, P.J. AI based Speech Language Therapy using Speech Quality Parameters for Aphasia Person: A Comprehensive Review. In Proceedings of the 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 5–7 November 2020; pp. 1263–1271. [CrossRef]
48. Aicha, A.B. Noninvasive Detection of Potentially Precancerous Lesions of Vocal Fold Based on Glottal Wave signal and sVM Approaches. *Procedia Comput. Sci.* **2018**, *126*, 586–595. [CrossRef]
49. Fontes, A.I.R.; Souza, P.T.V.; Neto, A.D.D.; Martins, A.d.M.; Silveira, L.F.Q. Classification System of Pathological Voices Using Correntropy. *Math. Probl. Eng.* **2014**, *2014*, 1–7. [CrossRef]
50. Alías, F.; Socoro, J.; Sevillano, X. A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds. *Appl. Sci.* **2016**, *6*, 143. [CrossRef]

51. Hossain, M.S.; Muhammad, G. Healthcare Big Data Voice Pathology Assessment Framework. *IEEE Access* **2016**, *4*, 7806–7815. [CrossRef]
52. Vaziri, G.; Giguère, C.; Dajani, H.R. Evaluating noise suppression methods for recovering the Lombard speech from vocal output in an external noise field. *Int. J. Speech Technol.* **2019**, *22*, 31–46. [CrossRef]
53. Hegde, S.; Shetty, S.; Rai, S.; Dodderi, T. A Survey on Machine Learning Approaches for Automatic Detection of Voice Disorders. *J. Voice* **2019**, *33*, 947.e11–947.e33. [CrossRef]
54. Zhang, D.; Wu, K. *Pathological Voice Analysis*; Springer: Singapore, 2020. [CrossRef]
55. Chen, L.; Wang, C.; Chen, J.; Xiang, Z.; Hu, X. Voice Disorder Identification by using Hilbert-Huang Transform (HHT) and K Nearest Neighbor (KNN). *J. Voice* **2021**, *35*, 932.e1–932.e11. [CrossRef]
56. Uloza, V.; Padervinskis, E.; Vegiene, A.; Pribuisiene, R.; Saferis, V.; Vaiciukynas, E.; Gelzinis, A.; Verikas, A. Exploring the feasibility of smart phone microphone for measurement of acoustic voice parameters and voice pathology screening. *Eur. Arch. Oto-Rhino* **2015**, *272*, 3391–3399. [CrossRef]
57. Amami, R.; Smiti, A. An incremental method combining density clustering and support vector machines for voice pathology detection. *Comput. Electr. Eng.* **2017**, *57*, 257–265. [CrossRef]
58. Lee, J.Y. 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.* **2012**, *2012*, 252. [CrossRef]
59. Pham, M.; Lin, J.; Zhang, Y. Diagnosing Voice Disorder with Machine Learning. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 5263–5266. [CrossRef]
60. Hammami, I.; Salhi, L.; Labidi, S. Voice Pathologies Classification and Detection Using EMD-DWT Analysis Based on Higher Order Statistic Features. *IRBM* **2020**, *41*, 161–171. [CrossRef]
61. Fonseca, E.S.; Guido, R.C.; Junior, S.B.; Dezani, H.; Gati, R.R.; Mosconi Pereira, D.C. Acoustic investigation of speech pathologies based on the discriminative paraconsistent machine (DPM). *Biomed. Signal Process. Control* **2020**, *55*, 101615. [CrossRef]
62. Muhammad, G.; Alhussein, M. Convergence of Artificial Intelligence and Internet of Things in Smart Healthcare: A Case Study of Voice Pathology Detection. *IEEE Access* **2021**, *9*, 89198–89209. [CrossRef]
63. Cordeiro, H.T.; Ribeiro, C.M. Spectral envelope first peak and periodic component in pathological voices: A spectral analysis. *Procedia Comput. Sci.* **2018**, *138*, 64–71. [CrossRef]
64. Erfanian Saeedi, N.; Almasganj, F.; Torabinejad, F. Support vector wavelet adaptation for pathological voice assessment. *Comput. Biol. Med.* **2011**, *41*, 822–828. [CrossRef]
65. Vásquez-Correa, J.; Klumpp, P.; Orozco-Arroyave, J.R.; Nöth, E. Phonet: A Tool Based on Gated Recurrent Neural Networks to Extract Phonological Posteriors from Speech. In Proceedings of the Interspeech 2019, ISCA, Graz, Austria, 15–19 September 2019; pp. 549–553. [CrossRef]
66. Wu, H.; Soraghan, J.; Lowit, A.; Di Caterina, G. Convolutional Neural Networks for Pathological Voice Detection. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; pp. 1–4. [CrossRef]
67. Areiza-Laverde, H.J.; Castro-Ospina, A.E.; Peluffo-Ordóñez, D.H. Voice Pathology Detection Using Artificial Neural Networks and Support Vector Machines Powered by a Multicriteria Optimization Algorithm. In *Applied Computer Sciences in Engineering*; Figueroa-García, J.C., López-Santana, E.R., Rodriguez-Molano, J.I., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; Volume 915, pp. 148–159. [CrossRef]
68. Chen, L.; Chen, J. Deep Neural Network for Automatic Classification of Pathological Voice Signals. *J. Voice* **2020**, *36*, 288.E15–288.E24. [CrossRef]
69. Miliaresi, I.; Poutos, K.; Pikrakis, A. Combining acoustic features and medical data in deep learning networks for voice pathology classification. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 18–21 January 2021; pp. 1190–1194. [CrossRef]
70. Gómez García, J.A. Contributions to the Design of Automatic Voice Quality Analysis Systems Using Speech Technologies. Ph.D. Thesis, Universidad Politécnica de Madrid, Madrid, Spain, 2018. [CrossRef]
71. Syed, S.A.; Rashid, M.; Hussain, S.; Zahid, H. Comparative Analysis of CNN and RNN for Voice Pathology Detection. *BioMed Res. Int.* **2021**, *2021*, 1–8. [CrossRef]
72. Kim, H.; Jeon, J.; Han, Y.J.; Joo, Y.; Lee, J.; Lee, S.; Im, S. Convolutional Neural Network Classifies Pathological Voice Change in Laryngeal Cancer with High Accuracy. *J. Clin. Med.* **2020**, *9*, 3415. [CrossRef]
73. Wahengbam, K.; Singh, M.P.; Nongmeikapam, K.; Singh, A.D. A Group Decision Optimization Analogy-Based Deep Learning Architecture for Multiclass Pathology Classification in a Voice Signal. *IEEE Sens. J.* **2021**, *21*, 8100–8116. [CrossRef]
74. Raj, J.R.; Jabez, J.; Srinivasulu, S.S.; Gowri, S.; Vimali, J.S. Voice Pathology Detection Based on Deep Neural Network Approach. *IOP Conf. Ser. Mater. Sci. Eng.* **2021**, *1020*, 012001. [CrossRef]
75. Fan, Z.; Wu, Y.; Zhou, C.; Zhang, X.; Tao, Z. Class-Imbalanced Voice Pathology Detection and Classification Using Fuzzy Cluster Oversampling Method. *Appl. Sci.* **2021**, *11*, 3450. [CrossRef]
76. Tolo̧şi, L.; Lengauer, T. Classification with correlated features: Unreliability of feature ranking and solutions. *Bioinformatics* **2011**, *27*, 1986–1994. [CrossRef] [PubMed]
77. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. 2017. Available online: http://xxx.lanl.gov/abs/1412.6980 (accessed on 20 January 2022).

78. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. 2017. Available online: http://xxx.lanl.gov/abs/1608.03983 (accessed on 20 January 2022).
79. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. 2015. Available online: http://xxx.lanl.gov/abs/1512.03385 (accessed on 20 January 2022).
80. Collobert, R.; Puhrsch, C.; Synnaeve, G. Wav2Letter: An End-to-End ConvNet-Based Speech Recognition System. 2016. Available online: http://xxx.lanl.gov/abs/1609.03193 (accessed on 20 January 2022).
81. Dai, W.; Dai, C.; Qu, S.; Li, J.; Das, S. Very Deep Convolutional Neural Networks for Raw Waveforms. 2016. Available online: http://xxx.lanl.gov/abs/1610.00087 (accessed on 20 January 2022).
82. Vanagas, G.; Engelbrecht, R.; Damaševičius, R.; Suomi, R.; Solanas, A. EHealth Solutions for the Integrated Healthcare. *J. Healthc. Eng.* **2018**, *2018*, 3846892. [CrossRef] [PubMed]
83. Payten, C.L.; Nguyen, D.D.; Novakovic, D.; O'Neill, J.; Chacon, A.M.; Weir, K.A.; Madill, C.J. Telehealth voice assessment by speech language pathologists during a global pandemic using principles of a primary contact model: An observational cohort study protocol. *BMJ Open* **2022**, *12*, e052518. [CrossRef]