Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Quantitative analysis of automatic voice disorder detection studies for hybrid feature and classifier selection[☆]

Jong Bub Lee[a], Hyun Gyu Lee[a,b,*]

[a] Department of Electrical and Computer Engineering, Inha university, Incheon, 22212, Republic of Korea
[b] College of Medicine, Inha university, Incheon, 22212, Republic of Korea

ARTICLE INFO

Dataset link: https://stimmdb.coli.uni-saarland.de/index.php4#target

Keywords:
Voice disorder detection
Machine learning
Speech analysis
Healthcare

ABSTRACT

Owing to the development of machine learning, particularly deep learning, researchers have focused on automatic voice-disorder detection. However, voice-disorder datasets vary significantly in terms of the number of patients per disorder, and different conditions are targeted in different studies. Therefore, conducting direct comparisons of performances across related studies is complicated. Hence, we compare conventional machine learning, deep learning, and multimodal methods by establishing a fixed dataset and an evaluation pipeline using the Saarbrücken voice database, which is the most commonly used database for automatic voice-disorder detection. In addition, we propose an automatic voice-disorder detection method that combines features and classifiers. Experimental results show mean unweighted average recall differences of 8% and 15% on the abovementioned two datasets, respectively, and that the proposed combination improves them by 1.5% and 0.5%, respectively.

1. Introduction

Voice is the primary method of human communication and provides important information regarding gender, age, emotions, physical conditions, and cultural environments [1,2]. A voice disorder occurs when the voice quality, pitch, or loudness is different or inappropriate with respect to an individual's age, sex, cultural background, or geographic environment. Additionally, voice disorders occur when an individual is uncomfortable with an abnormal voice that does not satisfy their daily needs, even if others do not consider the voice to be different or deviant [3]. Typical voice disorders result from vocal cord cysts, vocal polyps, and vocal paralysis, which cause incomplete contact between the vocal cords and result in hoarseness [1,4]. Voice disorders are typically encountered in professions that require excessive use of the voice, such as lecturers, lawyers, and singers; specifically, 25% of people in these professions worldwide are affected by voice disorders [5,6]. People with voice disorders may experience anxiety, depression, and difficulties communicating with others, which can result in various social and personal complications [7]. Therefore, accurate early detection and treatment of voice disorders are important.

Laryngoscopy, an invasive method used to differentiate between normal and disordered voices, is a preferred method as it allows direct visualization of the vocal cords. However, it presents limitations in providing feedback on the results of each voice treatment [1,8,9]. Invasive testing methods require well-trained professionals and expensive equipment, which can delay accurate diagnosis and subsequent treatment in areas without adequate healthcare facilities [7,9]. In addition, invasive testing methods are time-consuming and can be painful or traumatic for some patients. Consequently, patients may not seek early treatment if their voice is not significantly affected, which will worsen the situation [7]. By contrast, non-invasive methods have received significant attention and have been investigated comprehensively in speech-language pathology because of their use of acoustic information [1,10]. Non-invasive methods for diagnosing voice disorders rely on either perceptual or objective assessments. Perceptual assessments are reliable as they require the speech-language pathologist to hear and evaluate a patient's voice directly; however, they are challenging as they involve subjective elements in decision-making [11,12]. Meanwhile, objective assessment methods or automated voice-disorder diagnostic methods use signal processing techniques to extract temporal

[☆] This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [No. 2022-0-00641, XVoice: Multi-Modal Voice Meta Learning], [No. RS-2022-00155915, Artificial Intelligence Convergence Innovation Human Resources Development (Inha University)], the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2022R1F1A1071574), and INHA UNIVERSITY Research Grant.

[*] Corresponding author at: Department of Electrical and Computer Engineering, Inha university, Incheon, 22212, Republic of Korea.
E-mail addresses: bub3690@inha.edu (J.B. Lee), hglee@inha.ac.kr (H.G. Lee).

https://doi.org/10.1016/j.bspc.2024.106014
Received 17 August 2023; Received in revised form 21 December 2023; Accepted 29 January 2024
Available online 6 February 2024

Table 1
Overview of voice-disorder detection and classification studies performed on SVD. $\mathbf{x}_{CD}$: Conventional dysphonic feature; $\mathbf{x}_{SB}$: Spectrogram-based feature; $\mathbf{x}_{OS}$: OpenSMILE feature; $\mathbf{x}_{GS}$: Glottal source feature.

Literature	Model	$\mathbf{x}_{CD}$	$\mathbf{x}_{SB}$	$\mathbf{x}_{OS}$	$\mathbf{x}_{GS}$	Subset of disorder	Record	Published score
ML								
2012 [14]	GMM	✓				All 71 pathologies excluding damaged files	All vowels on all pitch levels	67.0%(ACC)
2014 [15]	GMM-SVM	✓				1 pathology	/a,u/ produced at normal	81.1%(ACC)
2015 [16]	GMM	✓				11 functional and organic dysarthric pathologies	/a/ produced at normal	99.8%(ACC)
2015 [17]	SVM(PCA)	✓				Selected 4 voice pathologies	/a/ produced at normal	86.0%(ACC)
2016 [18]	ELM,GMM,SVM	✓				Vocal fold pathology	–	95.0%(ACC)
2017 [19]	GMM	✓				Selected 5 voice pathologies	/a/ produced at normal	80.2%(ACC)
2016 [20]	RF with PCA	✓				Selected 4 voice pathologies and others	each /a, i, u/ produced at high, normal, low	99.0%(ACC)
2017 [21]	Naive Bayes	✓				Selected 2 voice pathologies	/a/ produced at normal	90.0%(ACC)
2018 [22]	Isolation forest	✓				Patients over 19 years old and under 60 years old.	/a/ produced at normal	61.7%(ACC)
2018 [22]	XGBOOST	✓			✓	Patients over 19 years old and under 60 years old.	/a/ produced at normal	73.3%(ACC)
2018 [23]	DT, BC, LMT (PCA)	✓				Selected samples from all 71 pathologies	/a/ produced at normal	83.5%(ACC)
2018 [23]	SVM	✓				Selected samples from all 71 pathologies	/a/ produced at normal	85.7%(ACC)
2017 [24]	SVM				✓	Selected 3 voice pathologies	/a/ produced at normal	93.2%(ACC)
2019 [25]	SVM				✓	Selected samples from all 71 pathologies	Phrase	76.1%(ACC)
2019 [26]	SVM(JOLL4R)	✓				Selected 2 voice pathologies	/a/ produced at any two of the four different pitch types	87.8%(ACC)
2020 [11]	SVM			✓	✓	Selected 5 from organic and non-organic pathologies	All vowels on all pitch levels	85.2%(ACC)
2021 [27]	SVM			✓		Organic pathologies, All 71 pathologies	Phrase	86.7%(UAR)
DL								
2018 [28]	CDBN		✓			Selected 6 voice pathologies	/a/ produced at normal	71.0%(ACC)
2019 [29]	Alexnet		✓			Selected 4 voice pathologies	/a/ produced at normal	94.2%(ACC)
2020 [30]	Resnet34		✓			Selected samples from all 71 pathologies	/a/ produced at normal	94.5%(ACC)
2021 [31]	CNN, LSTM	✓				12 Common voice pathologies	/a/ produced at normal	87.1%(ACC)
2021 [32]	1D CNN, LSTM		✓			Selected 4 voice pathologies	/a/ produced at normal	40.0%(F1)
2022 [7]	VGG 19 with SVM	✓	✓			All 71 pathologies	/a/ produced at normal	87.6%(ACC)
Multimodal								
2019 [33]	Bimodal GMM		✓			Structural pathologies	/a/ produced at normal	94.2%(ACC)
2021 [34]	Bimodal Xception with LSTM		✓			Structural pathologies	/a/ produced at normal	95.6%(ACC)
2022 [7]	Bimodal AlexNet, with feature selection, SVM	✓	✓			All 71 pathologies	/a/ produced at normal	90.1%(ACC)
2022 [35]	Bimodal Resnet18 with multimodal transfer module		✓			9 pathologies	/a/ produced at normal	100%(ACC)

and spectral acoustic features from a patient's voice for evaluation [13]. Objective methods are the most widely investigated in voice-disorder diagnosis because they are faster and more effective, and the acoustic features used are highly correlated with perceptual evaluation [11].

Three main objective assessment methods exist. The first method uses a machine learning (ML) classifier. After extracting acoustically handcrafted features from speech signals, an ML classifier is used to perform the diagnosis [11]. The second method, based on deep learning (DL), primarily uses a pretrained convolutional neural network (CNN) [36] to extract acoustic features and perform a diagnosis. The CNN, which is a type of DL algorithm designed for processing and analyzing structured image data, uses an audio spectrogram as an image for voice disorder assessment. The third is a multimodal method that uses both electroglottography (EGG) signals and acoustic features, and it has been shown to be useful in diagnosing voice disorders. The EGG signal is generated based on the relative change in the contact surface during glottic movement. This process holds significant importance in capturing information related to vocal cord vibrations [35].

Several problems have been indicated regarding the voice-disorder data used in objective assessment methods. Voice-disorder datasets vary widely in terms of the number of patients per disorder, and different studies have targeted different conditions, which makes it difficult to directly compare the performances obtained in those studies. In addition, the existence of multiple voice samples for each patient can cause the classifier to be biased toward the patient during cross-validation [27]. These problems make it challenging for researchers to objectively compare the performances acquired from related studies and determine the best method [22,37]. Therefore, this study aims to objectively evaluate voice-disorder detection methods and select the best method to be used as a basis for further investigations.

The contributions of this study are as follows:

1. Studies pertaining to voice-disorder detection from the perspectives of models, classifiers, and features are introduced.

2. An objective comparison of related studies based on ML, DL, and multimodal methods on a fixed dataset from the Saarbrücken voice database (SVD) is performed [38].
3. An accuracy level higher than that achieved previously is demonstrated by combining features and classifiers for automatic voice disorder detection.

The remainder of this paper is organized as follows: Section 2 introduces previous studies pertaining to conventional ML, DL, and multimodal methods. Section 3 describes the detailed composition of the SVD dataset, a subset of diseases based on the clinical classification scheme used in the experiment, and an evaluation pipeline. In addition, the feature extraction methods, classifiers, and DL feature extractors used in the experiment are described. In Section 4, the best studies for each method introduced in Section 2 are selected and evaluated, and an automatic voice-disorder detection method is proposed by combining the methods. Finally, Section 5 presents a summary and discussion of the results.

2. Related work

Automatic voice-disorder detection has been extensively investigated using ML, with an emphasis on identifying and comparing suitable features and models. Table 1 lists the used features, ML models, subsets of diseases, record materials, and classification performance for each study conducted on the SVD. The main types of features used for voice-disorder detection were conventional dysphonic, spectrogram-based, Open Source Speech and Music Interpretation by Large-space Extraction (OpenSMILE) [39,40], and glottal source features. The detailed construction of each feature used in this study is described in Section 3.

Studies utilizing ML have primarily used handcrafted conventional dysphonic features as classifier inputs. Melfrequency cepstral coefficients (MFCCs) are typically used as input features. After extracting the features, several classical classifiers are used to detect voice disorders. The classical supervised ML classifiers used are Gaussian mixture models, support vector machines (SVMs), random forest, and naive Bayes [14–21,26]. The SVM performed the best among classical ML classifiers that do not require dimensionality reduction methods such as principal component analysis (PCA) [23]. Methods other than the classical ML classifiers, such as XGBoost [41], have been investigated, and an isolation forest was used to indicate that normal label data cannot be 100% normal. Even though speech databases other than the SVD were used for training, the performance yielded was insufficient, with an accuracy of 73%, thus indicating that more data are required for speech disorder detection [22]. Meanwhile, glottal source features and the SVM were used in some studies to extract features by estimating the glottal signal using speech, which yielded performances equal to or better than those achieved by MFCCs [24,25]. Additionally, the OpenSMILE feature extraction toolkit, which integrates feature extraction algorithms from both the speech processing and music information retrieval communities, has been shown to perform better than MFCCs or spectrogram-based features, and that using glottal source features can further improve the performance [11,27].

Studies that utilize DL have mainly considered CNN models to automatically extract acoustic features from spectrogram-based features [28–32]. However, overfitting occurs when an insufficient amount of data is used, whereas using pretrained CNNs results in relatively good performance [29,30]. To compensate for the above-mentioned shortcoming, some studies have combined conventional dysphonic features with a CNN for training and classification using an SVM [7].

Multimodal methods have been investigated to train two parallel CNNs using both voice and EGG as input. Late fusion methods concatenate the outputs of two pretrained networks and classify them using long short-term memory (LSTM) and SVMs [7,34]. Additionally, researchers have investigated a multimodal transfer module (MMTM) that performs squeeze excitation operations between feature maps in the middle layer via an intermediate fusion method [35,42].

Table 2
Configuration of SVD database used in the experiment; #1 and #2 indicate that two subsets of disorders used in the experiment.

Subset		Types	# of patients	# of samples
		Control	630	638
All #2	Organic #1	Structural	269	325
		Neurogenical	164	271
	Non organic	Functional	234	291
		Psychogenic	79	54
		Others	341	409

As can be observed in Table 1, although several relevant studies have been conducted, each study involved a different subset of diseases and different types of pronounced speech. Therefore, in this study, we selected ML, DL, and multimodal methods to perform a comparison on a fixed dataset. The selection criterion was the most recent study with non-overlapping features and models. Specifically, we selected the ML, DL, and multimodal methods reported in [7,11,23,27,29–31], and [7,34,35], respectively.

3. Experimental setup

3.1. Dataset

The voice-disorder database used in our experiments was the SVD [38], which is a German database available for free. It comprises voice and EGG recordings from 1356 patients with 71 distinct disorders. It is important to note that a single patient may exhibit more than one disorder simultaneously. The recordings include sustained vowel pronunciations (/a/, /i/, and /u/) recorded at various pitches (high, low, normal, and rising–falling pitches) and pronunciations of the phrase "Guten Morgen, wie geht es Ihnen?" ("Good morning, how are you?"). The voice and EGG recordings were sampled at 16-bit precision at 50,000 samples per second [43]. For our experiments, both the voice and EGG recordings were downsampled to 16,000 samples per second to facilitate computational efficiency.

The dataset used in our experiments consists of 2225 samples of phrase pronunciations. It is essential to acknowledge that not all patients had their vowel pronunciations recorded. We opted to focus on phrase data for our experiments owing to their higher discriminative power, as supported by previous research [27]. We established a dataset using samples from patients who provided both phrase and vowel data recorded in voice and EGG so that we could combine different pronunciations in future studies.

According to the American Speech Language Hearing guidelines, voice disorders are categorized into organic, functional, and psychogenic voice disorders [3]. Organic voice disorders are further categorized into structural and neurogenic voice disorders. Non-organic voice disorders are classified into functional and psychogenic voice disorders. If none of these apply, then the category is "other". Table 2 lists the subsets of disorders and the number of samples used in the experiments. To ensure the robustness and relevance of our experimental results, we deliberately focused on two fixed disease categories: organic voice disorders and the comprehensive set of all voice disorders. This strategic decision was motivated by the potential for misleading outcomes when incorporating non-organic voice disorders into the analysis. Non-organic disorders, as supported by previous studies [4,11,27], primarily manifest as voice fatigue without significantly affecting the overall voice quality; furthermore, organic and non-organic voice disorders require different treatment methods. The classification of the disorders in the SVD database sample is organized similarly as in a previous study [27], and the names of the diseases in each category are included in Supplementary Table A.1.

3.2. Evaluation metrics

Voice-disorder detection is a binary classification problem where a speech sample is categorized as either disordered or non-disordered. Voice-disorder detection performance is measured based on accuracy, sensitivity, specificity, and unweighted average recall (UAR) [44]. The dataset of the SVD was unbalanced, with a 2.1:1 ratio of disordered to non-disordered data. Therefore, even after performing a stratified K-fold cross-validation, the validation and test data were unbalanced. Performance evaluation based on accuracy (ACC) has been performed in previous studies; however, ACC is affected significantly by the majority class in class-imbalanced sets. Therefore, UAR was used to evaluate the disordered and non-disordered data by weighting them equally. The evaluation metrics were calculated as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

$$SN = \frac{TP}{TP + FN}, \tag{2}$$

$$SP = \frac{TN}{TN + FP}, \tag{3}$$

$$UAR = \frac{SP + SN}{2}, \tag{4}$$

where true positives (TPs) and false negatives (FNs) occur when the spoken sample is an actual disordered voice, whereas true negatives (TNs) and false positives (FPs) occur when the voice sample is a true non-disordered voice.

3.3. Experimental pipeline

The data were split into training and testing data at a ratio of 8:2 based on the number of patients to ensure that the same patients were not included in the testing data. The training data were split into stratified five-fold validation data at a ratio of 8:2 ratio to ensure that the same patients were not included in the validation data. The training data were randomly oversampled to balance the proportions of normal and voice-disordered data. As the number of voice samples per patient varied, the performance of the testing set was affected significantly by random sampling. Therefore, the experiments were conducted using a dataset comprising five random samples; subsequently, the performance was measured based on the average values of the evaluation metrics.

3.4. Feature extraction

3.4.1. Conventional dysphonic features

Two types of conventional dysphonic features, $\mathbf{x}_{\mathrm{CD}}$, have been used to diagnose voice disorders. First, speech language pathologists use acoustic indicators, which provide an objective measure of the irregularity, wobbliness, or noisiness of a sound's waveform and can include the F0, jitter, shimmer, and harmonic-to-noise ratio [23,45]. Second, speaker characteristics – features based on speech and speaker recognition – are determined. When combined, these features can be applied not only to speaker characteristics, but also to voice-disorder diagnosis. The typical features are pitch, spectral slope, MFCC, and LPC [7]. To represent the global characteristics of the speaker, each feature was compressed into 11 statistics (mean, standard deviation, maximum, minimum, kurtosis, skewness, q1, q2, q3, iqr, and minimum) along the time axis. A total of 14 features were used in the experiment, including six other features.

3.4.2. Spectrogram-based feature

The features extracted via DL using spectrograms as input are known as spectrogram-based features, $\mathbf{x}_{\mathrm{SB}}$. The spectrograms used as input to the DL feature extractor are the short-time Fourier transform (STFT) and Mel-spectrograms, which extract both speech and EGG signals. In this study, the STFT spectrogram was computed by applying an STFT to the input signal using Hanning windows with a length of 25 ms and 25% overlap. To reduce the effect of high-frequency noise, a Mel filter was applied to the spectrogram to compute the Mel-spectrogram. In this case, 128 Mel filters were used. To prevent low-amplitude sounds from being undetected, we converted the amplitude units to logscale decibel units [35]. Furthermore, to ensure the stable convergence of the DL network, we performed the min–max normalization on each batch and then used the result as input to the network. Fig. 1 shows the EGG Mel-spectrograms of a normal patient and a patient with a voice disorder.

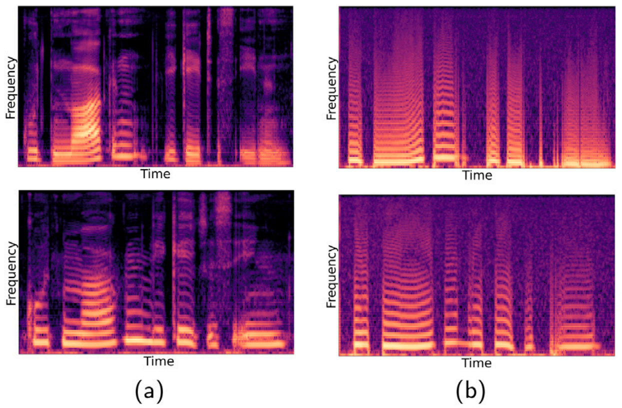

Fig. 1. Samples of Mel-spectrogram: (a) Mel-spectrogram of sound signal; (b) Mel-spectrogram of EGG signal, where control and disordered samples are shown in the first and second rows, respectively.

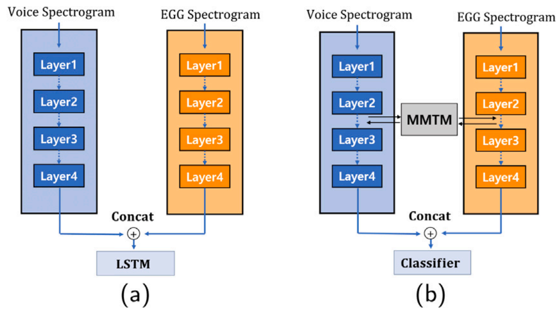

Fig. 2. Data fusion methods used by the Multimodal feature extractor: (a) architecture of long short-term memory (LSTM) late fusion; CNN processes spectrograms from both voice and EGG, and the resulting features are concatenated for multimodal feature extraction using LSTM, (b) Architecture of multimodal transfer module (MMTM) [42]; MMTM extracts multimodal features from voice and EGG spectrograms using squeeze and excitation within the CNN process.

3.4.3. OpenSMILE feature

OpenSMILE is an acoustic feature-extraction toolkit that is available for free. The OpenSMILE features, $\mathbf{x}_{\mathrm{OS}}$, used in our experiments constitute the baseline feature extractor used in the Interspeech 2016 Computational Paralinguistics Challenge and are also known as ComParE features [39]. ComParE comprises 6373 features calculated via various statistical operations on 64 low-level descriptors and their respective delta values [46]. The ComParE features are uniformly distributed for each fold using QuantileTransformer in the Python Scikit-learn package [47].

3.4.4. Glottal source feature

The glottal source feature, $\mathbf{x}_{\mathrm{GS}}$, is one of the features used in speech signal processing. It refers to the feature extracted from the glottal flow waveform, which is used to describe and measure the operation of the phonation mechanism [11]. To estimate the glottal

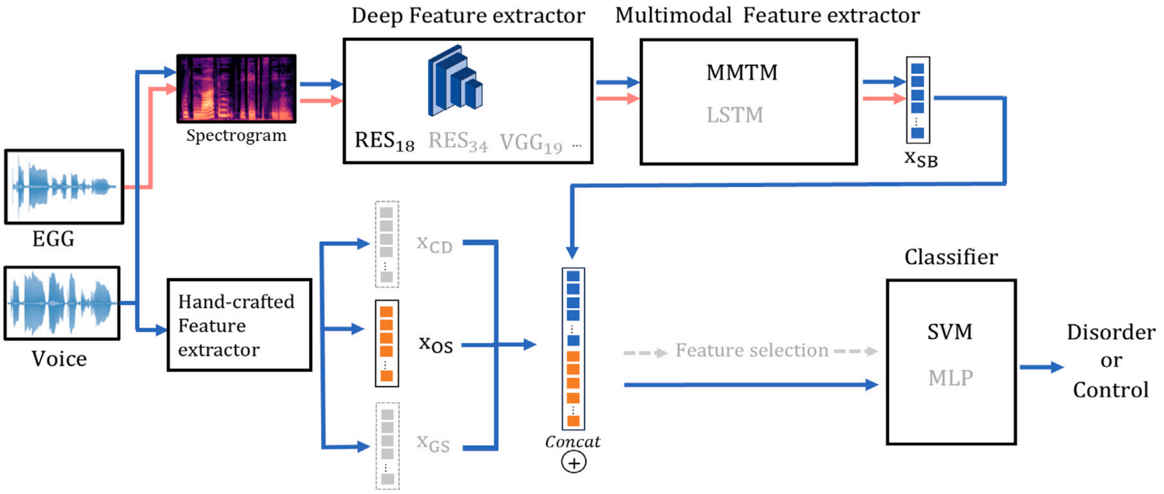

Fig. 3. Proposed combination method. Bold writing in black represents result of best combination method; gray writing and dotted arrows indicate features that are not combined. Red arrows indicate process of EGG feature extraction and blue arrows indicate process of voice feature extraction. Spectrograms from both EGG and voice undergo feature extraction using deep learning feature extractors and a multimodal feature extractor. At the same time, the raw voice signal is feature-extracted using a hand-crafted feature extractor. And then both hand-crafted and deep features concatenated before the classifier to detect voice disorder.

flow waveform in this experiment, we used the quasi-closed phase as a glottal inverse filtering algorithm [11,48]. Based on the estimated glottal flow waveform, we extracted two sets of features, i.e., time- and frequency-domain features. The time-domain feature set comprises opening quotients (OQ1 and OQ2), which represent the ratio of the vocal tract opening time; closing quotients (CQ and CQa), which represent the ratio of the vocal tract closing time; speech quotients (SQ1 and SQ2), which represent the ratio of the difference between the opening and closing speeds; amplitude and normalized amplitude quotients (AQ and NAQ), which represent the ratio of the maximum-to-minimum amplitude; and the quasi-open quotient (QOQ), which represents the difference between the open and closed time of the vocal folds. The frequency-domain features include the difference between the lowest glottal harmonics (H1-H2), which represents the difference in the harmonic amplitudes; the parabolic spectral parameter, which captures the spectral decay characteristics; and the harmonic richness factor, which indicates the degree of harmonic richness [1,11]. The glottal source features were extracted using the Matlab code in [11].

In conclusion, the experiment employed the following number of features for each audio processing method: $\mathbf{x}_{CD}$, 20 features; $\mathbf{x}_{SB}$, 128 times the frame size; $\mathbf{x}_{OS}$, 6373 features; and $\mathbf{x}_{GS}$, 192 features.

3.5. Machine learning model

3.5.1. Machine learning classifier

As can be observed in Table 1, the most commonly used ML classifier for voice-disorder detection in recent years is the SVM. In fact, the SVM has been shown to perform better than other ML classifiers without feature compression or selection [23]. SVMs are trained to determine the optimal hyperplane that maximizes the separation between two classes of support vectors. In addition, when the data are not linearly separable, a kernel can transform the samples into a higher-dimensional space [7]. The kernels used in this study were radial basis functional and polynomial kernels. A hyperparameter search was performed on the five-fold validation set, and box constraint C was used for the hyperparameter search.

3.5.2. Deep learning feature extractor

The DL feature extractors used for voice-disorder detection include CNNs and LSTM modules. Compared with one-dimensional CNNs that use voice waveforms as input, two-dimensional (2D) CNNs that use temporal and spectral 2D images perform better in terms of voice disorder detection [49]. CNNs perform better than LSTM modules that model time-series data [31]. However, owing to the small amount of data in the SVD, the 2D CNN was overfitted; therefore, we used a pretrained network. To use the STFT and Mel-spectrograms as inputs to the pretrained network, a one-channel 2D time-spectral image was duplicated into three channels, or the first and second derivatives were used as additional channels. The extracted features were classified into two classes using a multilayer perceptron (MLP).

3.5.3. Multimodal feature extractor

Three methods can be used to fuse voice and EGG data in DL networks. Early fusion combines these data before being input into the network; intermediate-level fusion combines intermediate results from two DL networks; and late fusion combines decisions before classification [50]. In this study, the MMTM method was used for intermediate-level fusion [42], whereas the LSTM and SVM classifiers were used for late fusion [34]. Fig. 2 illustrates the data fusion methods used by the multimodal feature extractors. Fig. 2.a shows the structure of the late-fusion network using the LSTM. The features from the speech and EGG samples passing through the CNN are concatenated to extract continuous multimodal features using LSTM. Fig. 2.b shows the MMTM structure applied between the feature maps at the intermediate stage of the bimodal CNN. The MMTM utilizes squeeze and excitation to recalibrate multimodal features. In addition, multimodal features are concatenated for classification using an SVM [7].

3.6. Combination method

Fig. 3 shows the proposed combination method. The features and methods presented in this figure are those proposed in previous studies. The experiments are ordered from left to right. The combination method with the best results was investigated. We investigated the feature extractors presented in previous studies and then compared those used for image classification. For the multimodal feature extractor, we compared the MMTM with intermediate-level fusion and the LSTM with late fusion. After feature extraction, we investigated four sets of features: $\mathbf{x}_{SB}$ using the DL feature extractor and handcrafted features $\mathbf{x}_{CD}$, $\mathbf{x}_{OS}$, and $\mathbf{x}_{GS}$. Subsequently, the extracted features were concatenated for comparison. Feature selection compares the input features with partial selection or no selection with the classifier. In this study, we used the SelectKBest feature selection method from the Python Scikit-learn package [47]. This method utilizes mutual information to measure the dependency between independent and dependent variables [51]. In the Classifier section, a comparison of the performances of SVM and MLP classifiers is presented.

Table 3
Reimplementation of selected studies and evaluation based on data from Subset1 and Subset2. Data of Subset 1: Organic voice disorder vs. control; Data of Subset 2: All voice disorder vs. control.

Literature	Method	Feature	Published score	Subset 1 score	Subset 2 score
ML					
[23]	Polynomial SVM	F0, Jitter, Shimmer, HNR, 13 MFCC and first, second derivatives	85.77% ACC	67.9% UAR	65.5% UAR
[27]	RBF SVM, Gaussianisation	OpenSMILE ComParE	86.73% UAR	**86.2% UAR**	**80.3% UAR**
[11]	RBF SVM	OpenSMILE ComParE, Glottal features	85.2% ACC	85.7% UAR	**80.5% UAR**
DL					
[30]	RESNET 34	Mel-Spectrogram and first, second derivatives	96.1% ACC	84.6% UAR	76.2% UAR
[31]	RESNET 34	13 MFCC, pitch, rolloff, ZCR, energy entropy, spectral flux, spectral centroid, and energy.	87.1% ACC	82.3% UAR	76.7% UAR
[29]	3 parallel alexnet with 3-layer mlp fusion	Mel-Spectrogram and first, second derivatives	94.2% ACC	82.6% UAR	72.1% UAR
[7]	VGG19 with SVM	STFT, 30 MFCC, LPC, F0, Spectral slope	87.6% ACC	79.0% UAR	75.7% UAR
Multimodal					
[34]	Bimodal Xception with LSTM	Mel spectrogram	95.6% ACC	84.7% UAR	75.5% UAR
[7]	Bimodal AlexNet, SVM, Handcrafted Feature fusion, Feature selection	STFT, 30 MFCC, LPC, F0, Spectral slope	90.1% ACC	84.0% UAR	77.2% UAR
[35]	Bimodal Resnet18, Multimodal transfer module	Mel spectrogram	100% ACC	85.9% UAR	76.2% UAR

4. Result

4.1. Quantitative analysis of representative related studies

Table 3 presents the results of reimplementing the methods selected from the ML, DL, and multimodal studies and the experimental results on the two data subsets based on the experimental pipeline shown in Section 3.3. The data from Subset 1 refers to the organic voice disorder and non-disordered classification problems, and the data from Subset 2 refers to all the voice disordered and non-disordered classification problems. The average published score was 90.84%; however, the reimplementation results for Subset 1 and Subset 2 were 82.29% and 75.59% in terms of the average UAR. This shows that the performance of the same method differs significantly on different datasets, thus suggesting that the researchers did not utilize all the data but instead performed an experiment using a subset. This also implies that the same speaker may have been present in the testing and training data. Compared with Subset 1, Subset 2 showed a UAR reduction of 6.7%. Although Subset 2 contained more types of disorders, which resulted in a wider feature distribution, the issue of sparsity occurred owing to the small number of samples per disease. Therefore, Subset 2 performed worse than Subset 1 owing to its generalization difficulty, even though the number of samples increased.

For all the algorithms in Table 3, we did not know exactly how many samples and what type of labels the original authors used, so we could not directly compare them to the published scores. What we attempt to present in Table 3 is that we fixed the datasets to objectively compare the performance of existing studies that experimented with different subsets of the SVD dataset, and we could directly see which ones performed better on the same dataset. In addition, we performed hyperparameter tuning for each algorithm on our proposed subsets because we were concerned that the original authors' hyperparameters might not work well on new subsets.

The results of the ML, DL, and multimodal reimplementation are interpreted next.

4.1.1. Machine learning classifier

In ML, classification is performed using an SVM on conventional dysphonic, OpenSMILE, and glottal source features [11,23,27]. In Subset 1, the OpenSMILE feature was the best, with a UAR of 86.2%; and in Subset 2, the fusion of OpenSMILE and glottal source features yielded the best result with a UAR of 80.5%, which was also the best result achieved using the ML, DL, and multimodal methods. This shows that OpenSMILE, which is a manually extracted feature, is more discriminative than DL-based feature extractors, which automatically extract features when data are scarce. In particular, because the SVM uses only support vectors to define decision boundaries, it exhibits high generalization performance even under unbalanced class distributions and thus performs well in Subset 2, which contains a small amount of data per disease. However, when conventional dysphonic features were used, the performance was the worst, with UAR values of 67.9% and 65.5% for Subset 1 and Subset 2, respectively.

4.1.2. Deep learning feature extractor

In DL, spectrogram-based and conventional dysphonic features are extracted and classified using a pretrained CNN [7,29–31]. Conventional dysphonic features are combined with the output of the CNN and passed to the classifier. When training ResNet34 on the Mel-spectrogram and its delta values, the best performance was achieved in Subset 1, with a UAR of 84.6%. The Mel-spectrogram features demonstrated better discriminative ability than the spectrogram and MFCC, indicating that using the entire speech as input is superior to separating it into three parallel AlexNet inputs. Compared to handcrafted features, the algorithms based on DL in Subset 1 do not show a significant drop in performance. However, when we move to Subset 2, where there is an increase in classes with a small amount of data, methods using DL suffer from insufficient learning compared to approaches using OpenSMILE, resulting in a significant performance degradation. In a method that combines DL-based algorithms with handcrafted features [31], a performance improvement of 0.5% was observed in the UAR compared with that using the Mel-spectrogram only. Nonetheless, because many DL feature extractors beyond ResNet34, AlexNet, and VGG19 have not yet been investigated, further research is necessary to identify superior feature extractors.

4.1.3. Multimodal feature extractor

In multimodal methods, STFT, or Mel-spectrograms, are used to extract features from a bimodal CNN to perform classification. The two images in Fig. 2 show the structure of a bimodal CNN, where the two CNNs do not share weights. In Subset 1, the intermediate-level fusion level using the MMTM was 85.9%, which was superior to that achieved by the late fusion method using the LSTM and SVM. Fig. 2.a shows the structure of the bimodal CNN using the MMTM. The best performance was achieved when the MMTM in the output of Layer 2 in ResNet18 was used. Subset 2 performed the best, with a UAR of 77.2%, when fusing the spectrogram and conventional dysphonic features. This indicates that, similar to the results of the voice-only DL feature extractor, overfitting occurred in the sparse samples per disease. Additionally, this implies that SVMs other than MLPs can be used as classifiers for feature fusion in DL feature extractors. However, feature extractors other than Xception, AlexNet, and ResNet18 should be investigated as multimodal feature extractors.

Table 4
Result of feature and method combination. Data of Subset 1: Organic voice disorder vs. control; Data of Subset 2: All voice disorder vs. control. $\mathbf{x}_{SB}$: Spectrogram-based feature; $\mathbf{x}_{OS}$: OpenSMILE feature; The spectrogram-based feature used in this table is the Mel spectrogram.

Method	Feature	Subset 1 UAR	Subset 2 UAR	Average
Sound only				
SVM (RBF) [27]	$\mathbf{x}_{OS}$	86.2%	80.3%	83.3%
Resnet 18	$\mathbf{x}_{SB}$(sound)	84.4%	76.8%	80.6%
Resnet18 + MLP	$\mathbf{x}_{SB}$(sound) + $\mathbf{x}_{OS}$	86.5%	78.4%	82.5%
Resnet18 + SVM(RBF)	$\mathbf{x}_{SB}$(sound) + $\mathbf{x}_{OS}$	86.7%	**80.9%**	83.8%
Resnet18 + SVM(RBF) + Feature selection	$\mathbf{x}_{SB}$(sound) + $\mathbf{x}_{OS}$	87.2%	80.7%	84.0%
Multimodal				
Resnet18 + MMTM	$\mathbf{x}_{SB}$(sound,EGG)	85.9%	76.2%	81.1%
Resnet18 + MMTM + MLP	$\mathbf{x}_{SB}$(sound,EGG) + $\mathbf{x}_{OS}$	**88.0%**	79.3%	83.7%
Resnet18 + MMTM + SVM	$\mathbf{x}_{SB}$(sound,EGG) + $\mathbf{x}_{OS}$	87.3%	**81.0%**	84.2%
Resnet18 + MMTM + SVM + Feature selection	$\mathbf{x}_{SB}$(sound,EGG) + $\mathbf{x}_{OS}$	87.5%	**81.0%**	**84.3%**

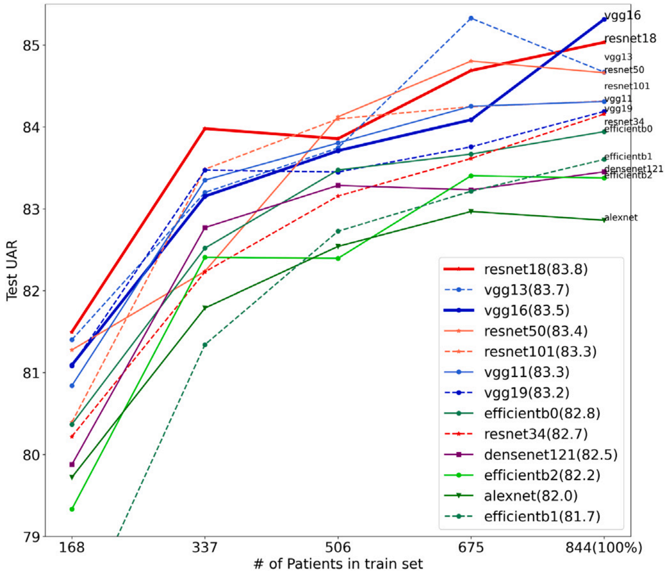

Fig. 4. Comparison of deep learning feature extractor experiments; x-axis represents number of patients in training data, and y-axis represents test UAR (%) performance. Values in parentheses are the average values of test UAR with respect to the x-axis.

4.2. Comparison of deep learning feature extractors

Previous studies pertaining to DL feature extractors are not extensive, which makes it difficult to select the best extractor. Therefore, we performed experiments on various CNNs using a three-channel Mel-spectrogram as the input to select the best extractor. The CNNs used were ResNet, EfficientNet, DenseNet, VGG, and AlexNet [36,52–55]. Fig. 4 shows a comparison of the performances of the networks as the number of patients increases, with the x-axis representing the number of patients in the training data and the y-axis representing the test UAR (%). The numbers in parentheses to the right of the network names are the average values of the test UAR as the x-axis varies. The result shows that ResNet18, which contains fewer parameters as compared with the other CNNs, outperformed the deeper-layered ResNet and DenseNet. This is because the greater the number of network parameters, the worse the performance owing to the insufficient amount of data. When all the training data were used, VGG16 demonstrated the best performance (85.3%). However, as shown in Fig. 4, ResNet18 exhibited the highest average sum of variation along the x-axis, indicating that it is a more robust feature extractor for sparse data than VGG16. Therefore, ResNet18 was utilized as the feature extractor to compare the combination methods.

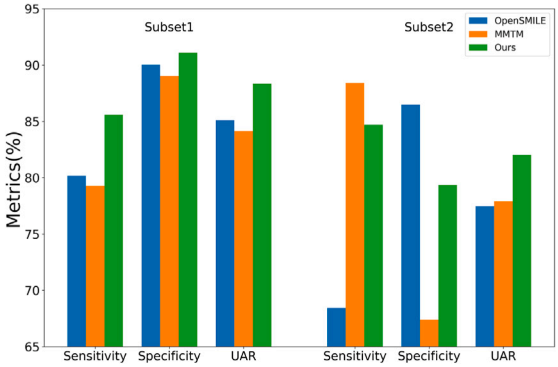

Fig. 5. Detailed comparison of performance metrics for different methods; x-axis represents metrics, and y-axis represents metrics values (%). Left bars show the comparison results on Subset1 and the right bars on Subset2.

4.3. Best component analysis for voice disorder detection

In Sections 4.1 and 4.2, we presented the selection of the best methods and a comparison of the performances of DL feature extractors on the two data subsets. We excluded conventional dysphonic and glottal source features, which exhibited unsatisfactory performances, and used both spectrogram-based and OpenSMILE features combined. Table 4 presents a comparison of the results yielded using the methods selected based on Fig. 3. Based on the data of Subset 1, after adding the EGG feature extractor, the UAR increased by 1.5% from 84.4% when only the DL feature extractor was used, and the best performance was achieved (88.0% UAR) when the MLP was used as a classifier by adding OpenSMILE features. However, based on the data of Subset 2, the performance deteriorated or remained the same despite the addition of EGG features. This finding implies that EGGs are underfeatured in disorders other than organic voice disorders, as can similarly be observed in Fig. 1.b, which shows no significant difference between the normal and disordered EGG Mel-spectrogram samples. The feature selection method improved performance slightly, but not sufficiently to attain a significant difference. Based on the data in Subset 2, the best performance was achieved when the SVM was fused with OpenSMILE features, where UAR values of 81.0% and 80.9% were observed. Therefore, using the spectrogram and OpenSMILE as features, ResNet18 and the MMTM as feature extractors, and the SVM as the classifier yielded the best performance, as shown in bold in Fig. 3.

4.4. Detailed result analysis

For a detailed analysis of the results, we conducted a comparison using a single seed on the test set. Fig. 5 compares our method to the two best models in ML and DL for specificity, sensitivity, and

UAR. The left bars show the comparison results on Subset 1 and the right bars on Subset 2. Our hybrid method demonstrated a notable improvement, particularly achieving a 6% higher sensitivity in Subset 1 compared to that of MMTM. This enhancement was most prominent in cases of structural disorders, while other subsets exhibited nearly identical results. Moreover, compared with the SVM with OpenSMILE features, our method demonstrated comparable improvements, notably in the prediction of structural disorders. This suggests that the hybrid method enhances the sensitivity to predict voice disorders, especially capturing subtle details in the presence of noise. In Subset 2, our method outperformed the MMTM by 12% in specificity, albeit at the expense of a 4% reduction in sensitivity. When compared with the SVM with OpenSMILE features, our method demonstrated an 11% improvement in sensitivity. This indicates that the deep feature extractor faces challenges in handling ambiguity and weak symptoms associated with non-organic voice disorders. Nevertheless, learning with hybrid features proved advantageous in overcoming these challenges.

5. Conclusion

This paper identifies the problems of existing voice disorder detection research and comprehensively summarizes models, classifiers, and features. Using the SVD, we built a fixed dataset and evaluation pipeline to objectively compare traditional ML, DL, and multimodal methods, and we proposed an automatic voice disorder detection method with improved performance by combining features and classifiers. We evaluated previous studies on datasets with two different subsets of disorders and found that there is a difference in average performance between the published results and ours. This indicates that previous studies focused on specific disorders and failed to perform accurate evaluations owing to the presence of duplicate speakers in the training and evaluation data. A combination of DL feature extractors and OpenSMILE features improved performance on an organic voice disorder dataset. A combination of SVM classifiers improved performance on all voice disorder datasets.

Although many DL methods have been improved, handcrafted features yield better results than DL feature extractors owing to the scarcity of voice disorder samples. To improve the performance of the DL feature extractors, a large amount of data is needed. In the future, we plan to collect more voice disorder databases other than the SVD. Then, we will use self-supervised learning to improve the feature extractors and increase the robustness for out-of-domain databases.

CRediT authorship contribution statement

Jong Bub Lee: Conceptualization, Data curation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Hyun Gyu Lee:** Funding acquisition, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jong Bub Lee reports financial support was provided by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT). Jong Bub Lee reports financial support was provided by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT).

Data availability

public dataset: https://stimmdb.coli.uni-saarland.de/index.php4#target.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.bspc.2024.106014.

References

[1] H.J. Park, B.J. Shin, Usefulness of glottal inverse filtering analysis in pathological voice, J. Speech-Lang. Hear. Disord. 30 (1) (2021) 41–48.

[2] J. Gómez-García, L. Moro-Velázquez, J. Godino-Llorente, On the design of automatic voice condition analysis systems. part I: Review of concepts and an insight to the state of the art, Biomed. Signal Process. Control 51 (2019) 181–199.

[3] B.-K. Julie M, J.N. Craig, Voice disorders, 2023, URL: https://www.asha.org/practice-portal/clinical-topics/voice-disorders/. Publisher: American Speech-Language-Hearing Association.

[4] J.C. Stemple, E.R. Hapner (Eds.), Voice Therapy: Clinical Case Studies, fifth ed., Plural Publishing, Inc, San Diego, CA, 2019.

[5] M. Alhussein, G. Muhammad, Voice pathology detection using deep learning on mobile healthcare framework, IEEE Access 6 (2018) 41034–41041.

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

[7] A.N. Omeroglu, H.M. Mohammed, E.A. Oral, Multi-modal voice pathology detection architecture based on deep and handcrafted feature fusion, Eng. Sci. Technol. Int. J. 36 (2022) 101148.

[8] D.D. Mehta, R.E. Hillman, Current role of stroboscopy in laryngeal imaging, Curr. Opin. Otolaryngol. Head Neck Surg. 20 (6) (2012) 429–436.

[9] R. Islam, M. Tarique, E. Abdel-Raheem, A survey on signal processing based pathological voice detection techniques, IEEE Access 8 (2020) 66749–66776.

[10] E.B. Holmberg, R.E. Hillman, J.S. Perkell, P.C. Guiod, S.L. Goldman, Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice, J. Speech Hear. Res. 38 (6) (1995) 1212–1223.

[11] P. Barche, K. Gurugubelli, A.K. Vuppala, Towards automatic assessment of voice disorders: A clinical approach, in: Interspeech 2020, ISCA, 2020, pp. 2537–2541, http://dx.doi.org/10.21437/Interspeech.2020-2160, URL: https://www.isca-speech.org/archive/interspeech_2020/barche20_interspeech.html.

[12] D. Kent Ray, Hearing and believing, Am. J. Speech-Lang. Pathol. 5 (3) (1996) 7–23, Publisher: American Speech-Language-Hearing Association.

[13] J.-W. Lee, H.-G. Kang, J.-Y. Choi, Y.-I. Son, An investigation of vocal tract characteristics for acoustic discrimination of pathological voices, BioMed. Res. Int. 2013 (2013) 758731.

[14] D. Martínez, E. Lleida, A. Ortega, A. Miguel, J. Villalba, Voice pathology detection on the Saarbrücken voice database with calibration and fusion of scores using MultiFocal toolkit, in: D. Torre Toledano, A. Ortega Giménez, A. Teixeira, J. González Rodríguez, L. Hernández Gómez, R. San Segundo Hernández, D. Ramos Castro (Eds.), Advances in Speech and Language Technologies for Iberian Languages, Vol. 328, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 99–109, http://dx.doi.org/10.1007/978-3-642-35292-8_11, URL: http://link.springer.com/10.1007/978-3-642-35292-8_11. Series Title: Communications in Computer and Information Science.

[15] I.M.M. El Emary, M. Fezari, F. Amara, Towards developing a voice pathologies detection system, J. Commun. Technol. Electron. 59 (11) (2014) 1280–1288.

[16] Ö. Eskidere, A. Gürhanlı, Voice disorder classification based on multitaper mel frequency cepstral coefficients features, Comput. Math. Methods Med. 2015 (2015) 1–12.

[17] N. Souissi, A. Cherif, Dimensionality reduction for voice disorders identification system based on mel frequency cepstral coefficients and support vector machine, in: 2015 7th International Conference on Modelling, Identification and Control, ICMIC, IEEE, Sousse, Tunisia, 2015, pp. 1–6, http://dx.doi.org/10.1109/ICMIC.2015.7409479, URL: http://ieeexplore.ieee.org/document/7409479/.

[18] M.S. Hossain, G. Muhammad, Healthcare big data voice pathology assessment framework, IEEE Access 4 (2016) 7806–7815.

[19] Z. Ali, M. Alsulaiman, G. Muhammad, I. Elamvazuthi, A. Al-nasheri, T.A. Mesallam, M. Farahat, K.H. Malki, Intra- and inter-database study for arabic, english, and german databases: Do conventional speech features detect voice pathology? J. Voice 31 (3) (2017) 386.e1–386.e8.

[20] D. Hemmerling, A. Skalski, J. Gajda, Voice data mining for laryngeal pathology assessment, Comput. Biol. Med. 69 (2016) 270–276.

[21] M. Dahmani, M. Guerti, Vocal folds pathologies classification using Naïve Bayes networks, in: 2017 6th International Conference on Systems and Control, ICSC, IEEE, Batna, Algeria, 2017, pp. 426–432, http://dx.doi.org/10.1109/ICoSC.2017.7958686, URL: http://ieeexplore.ieee.org/document/7958686/.

[22] P. Harar, Z. Galaz, J.B. Alonso-Hernandez, J. Mekyska, R. Burget, Z. Smekal, Towards robust voice pathology detection: Investigation of supervised deep learning, gradient boosting, and anomaly detection approaches across four databases, Neural Comput. Appl. 32 (20) (2020) 15747–15757.

[23] L. Verde, G. De Pietro, G. Sannino, Voice disorder identification by using machine learning techniques, IEEE Access 6 (2018) 16246–16255.

[24] G. Muhammad, M. Alsulaiman, Z. Ali, T.A. Mesallam, M. Farahat, K.H. Malki, A. Al-nasheri, M.A. Bencherif, Voice pathology detection using interlaced derivative pattern on glottal source excitation, Biomed. Signal Process. Control 31 (2017) 156–164.

[25] S.R. Kadiri, P. Alku, Analysis and detection of pathological voice using glottal source features, IEEE J. Sel. Top. Sign. Proces. 14 (2) (2020) 367–379.

[26] K. Wu, D. Zhang, G. Lu, Z. Guo, Joint learning for voice based disease detection, Pattern Recognit. 87 (2019) 130–139.

[27] M. Huckvale, C. Buciuleac, Automated detection of voice disorder in the Saarbrücken voice database: Effects of pathology subset and audio materials, in: Interspeech 2021, ISCA, 2021, pp. 1399–1403, http://dx.doi.org/10.21437/Interspeech.2021-1507, URL: https://www.isca-speech.org/archive/interspeech_2021/huckvale21_interspeech.html.

[28] H. Wu, J. Soraghan, A. Lowit, G. Di-Caterina, A deep learning method for pathological voice detection using convolutional deep belief networks, in: Interspeech 2018, ISCA, 2018, pp. 446–450, http://dx.doi.org/10.21437/Interspeech.2018-1351, URL: https://www.isca-speech.org/archive/interspeech_2018/wu18b_interspeech.html.

[29] M. Alhussein, G. Muhammad, Automatic voice pathology monitoring using parallel deep models for smart healthcare, IEEE Access 7 (2019) 46474–46479.

[30] M.A. Mohammed, K.H. Abdulkareem, S.A. Mostafa, M. Khanapi Abd Ghani, M.S. Maashi, B. Garcia-Zapirain, I. Oleagordia, H. Alhakami, F.T. AL-Dhief, Voice pathology detection and classification using convolutional neural network model, Appl. Sci. 10 (11) (2020) 3723.

[31] S.A. Syed, M. Rashid, S. Hussain, H. Zahid, Comparative analysis of CNN and RNN for voice pathology detection, BioMed. Res. Int. 2021 (2021) 1–8.

[32] V. Guedes, F. Teixeira, A. Oliveira, J. Fernandes, L. Silva, A. Junior, J.P. Teixeira, Transfer learning with AudioSet to voice pathologies identification in continuous speech, Procedia Comput. Sci. 164 (2019) 662–669.

[33] M.S. Hossain, G. Muhammad, A. Alamri, Smart healthcare monitoring: a voice pathology detection paradigm for smart cities, Multimedia Syst. 25 (5) (2019) 565–575.

[34] G. Muhammad, M. Alhussein, Convergence of artificial intelligence and internet of things in smart healthcare: A case study of voice pathology detection, IEEE Access 9 (2021) 89198–89209.

[35] L. Geng, Y. Liang, H. Shan, Z. Xiao, W. Wang, M. Wei, Pathological voice detection and classification based on multimodal transmission network, J. Voice (2022) S0892199722003708.

[36] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.

[37] S. Hegde, S. Shetty, S. Rai, T. Dodderi, A survey on machine learning approaches for automatic detection of voice disorders, J. Voice 33 (6) (2019) 947.e11–947.e33.

[38] B. Woldert-Jokisz, Saarbruecken Voice Database, Institut für Phonetik, Universität des Saarlandes, 2007.

[39] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J.K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, K. Evanini, The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language, in: Interspeech 2016, ISCA, 2016, pp. 2001–2005, http://dx.doi.org/10.21437/Interspeech.2016-129, URL: https://www.isca-speech.org/archive/interspeech_2016/schuller16_interspeech.html.

[40] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: The munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM International Conference on Multimedia, MM '10, Association for Computing Machinery, New York, NY, USA, 2010, pp. 1459–1462, http://dx.doi.org/10.1145/1873951.1874246.

[41] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA, 2016, pp. 785–794, http://dx.doi.org/10.1145/2939672.2939785, URL: https://dl.acm.org/doi/10.1145/2939672.2939785.

[42] H.R. Vaezi Joze, A. Shaban, M.L. Iuzzolino, K. Koishida, MMTM: Multimodal transfer module for CNN fusion, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, Seattle, WA, USA, 2020, pp. 13286–13296, http://dx.doi.org/10.1109/CVPR42600.2020.01330, URL: https://ieeexplore.ieee.org/document/9156844/.

[43] M. Huckvale, Z. Liu, C. Buciuleac, Automated voice pathology discrimination from audio recordings benefits from phonetic analysis of continuous speech, Biomed. Signal Process. Control 86 (2023) 105201.

[44] A. Rosenberg, Classifying skewed data: importance weighting to optimize average recall, in: Interspeech 2012, ISCA, 2012, pp. 2242–2245, http://dx.doi.org/10.21437/Interspeech.2012-131, URL: https://www.isca-speech.org/archive/interspeech_2012/rosenberg12_interspeech.html.

[45] R.T. Sataloff, Professional Voice: The Science and Art of Clinical Care, fourth ed., Plural Publishing Inc, San Diego, CA, 2017.

[46] F. Weninger, F. Eyben, B.W. Schuller, M. Mortillaro, K.R. Scherer, On the acoustics of emotion in audio: What speech, music, and sound have in common, Front. Psychol. 4 (2013).

[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, Scikit-learn: Machine learning in python, Mach. Learn. Python (2011).

[48] N.P. Narendra, P. Alku, Glottal source information for pathological voice detection, IEEE Access 8 (2020) 67745–67755.

[49] F. Javanmardi, S.R. Kadiri, M. Kodali, P. Alku, Comparing 1-dimensional and 2-dimensional spectral feature representations in voice pathology detection using machine learning and deep learning classifiers, in: Interspeech 2022, ISCA, 2022, pp. 2173–2177, http://dx.doi.org/10.21437/Interspeech.2022-10420, URL: https://www.isca-speech.org/archive/interspeech_2022/javanmardi22_interspeech.html.

[50] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, M.P. Lungren, Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines, NPJ Digit. Med. 3 (1) (2020) 136.

[51] B.C. Ross, Mutual information between discrete and continuous data sets, PLoS ONE 9 (2) (2014) e87357.

[52] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, Las Vegas, NV, USA, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90, URL: http://ieeexplore.ieee.org/document/7780459/.

[53] M. Tan, Q.V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, 2019, pp. 6105–6114, URL: http://proceedings.mlr.press/v97/tan19a.html.

[54] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, Honolulu, HI, 2017, pp. 2261–2269, http://dx.doi.org/10.1109/CVPR.2017.243, URL: https://ieeexplore.ieee.org/document/8099726/.

[55] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015, URL: http://arxiv.org/abs/1409.1556.