Speech Communication 152 (2023) 102952

Contents lists available at ScienceDirect

Speech Communication

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

# Multiple voice disorders in the same individual: Investigating handcrafted features, multi-label classification algorithms, and base-learners

Sylvio Barbon Junior [a,1], Rodrigo Capobianco Guido [b,*,1], Gabriel Jonas Aguiar [c,1], Everton José Santana [c,1], Mario Lemes Proença Junior [c,1], Hemant A. Patil [d,1]

[a] Department of Engineering and Architecture, University of Trieste, Piazzale Europa, 1 - 34127, Trieste FVG, Italy

[b] Instituto de Biociências, Letras e Ciências Exatas, Unesp - Univ Estadual Paulista (São Paulo State University), Rua Cristóvão Colombo 2265, Jd Nazareth, 15054-000, São José do Rio Preto SP, Brazil

[c] Computer Science Department, Londrina State University, Rodovia Celso Garcia Cid/PR 445, km 380, Campus Universitário, Zip code: 86057-970, Londrina PR, Brazil

[d] Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar 382007, India

ARTICLE INFO

*Keywords:*
Multiple voice disorders
Multi-label classification
Handcrafted feature extraction
Deep learning

ABSTRACT

Non-invasive acoustic analyses of voice disorders have been at the forefront of current biomedical research. Usual strategies, essentially based on machine learning (ML) algorithms, commonly classify a subject as being either healthy or pathologically-affected. Nevertheless, the latter state is not always a result of a sole laryngeal issue, i.e., multiple disorders might exist, demanding multi-label classification procedures for effective diagnoses. Consequently, the objective of this paper is to investigate the application of five multi-label classification methods based on problem transformation to play the role of base-learners, i.e., Label Powerset, Binary Relevance, Nested Stacking, Classifier Chains, and Dependent Binary Relevance with Random Forest (RF) and Support Vector Machine (SVM), in addition to a Deep Neural Network (DNN) from an algorithm adaptation method, to detect multiple voice disorders, i.e., Dysphonia, Laryngitis, Reinke's Edema, Vox Senilis, and Central Laryngeal Motion Disorder. Receiving as input three handcrafted features, i.e., signal energy (SE), zero-crossing rates (ZCRs), and signal entropy (SH), which allow for interpretable descriptors in terms of speech analysis, production, and perception, we observed that the DNN-based approach powered with SE-based feature vectors presented the best values of F1-score among the tested methods, i.e., 0.943, as the averaged value from all the balancing scenarios, under Saarbrücken Voice Database (SVD) and considering 20% of balancing rate with Synthetic Minority Over-sampling Technique (SMOTE). Finally, our findings of most false negatives for laryngitis may explain the reason why its detection is a serious issue in speech technology. The results we report provide an original contribution, allowing for the consistent detection of multiple speech pathologies and advancing the state-of-the-art in the field of handcrafted acoustic-based non-invasive diagnosis of voice disorders.

## 1. Introduction

The precise identification of different voice disorders persists as a challenge, requiring much dedication from health professionals. Usually, specific patterns perceived during vocal folds vibration, complemented by direct laryngeal examinations, are used to support the decisions (Casper and Leonard, 2011). Observing that the rate of successful diagnoses depends on the professionals' audition and commonly requires a subjective strategy, automated speech pathology detection (SPD) algorithms are certainly of paramount importance. Most of the current SPD algorithms are based on artificial intelligence methods, such as Support Vector Machines (SVMs), Decision Tree (DT), and Artificial Neural Networks (ANN), each of them representing a machine learning (ML)-based approach (AlRshoud et al., 2019; Lorenzo and Claudia, 2002; Ghasem et al., 2019; David, 2018; Belhaj et al., 2015; Verde et al., 2018a; Areiza-Laverde et al., 2018; Muhammad et al., 2012a).

In the above-cited scientific papers, voice disorders were classified by using ML models grounded on single-label classification procedures in which the outputs are mutually exclusive, i.e., each single pathology needs to be predicted. Nonetheless, the same subject could be

[*] Corresponding author.
*E-mail address:* guido@ieee.org (R.C. Guido).
[1] All the authors have contributed equally to this manuscript.

https://doi.org/10.1016/j.specom.2023.102952
Received 10 August 2021; Received in revised form 6 May 2023; Accepted 10 June 2023
Available online 20 June 2023
0167-6393/© 2023 Elsevier B.V. All rights reserved.

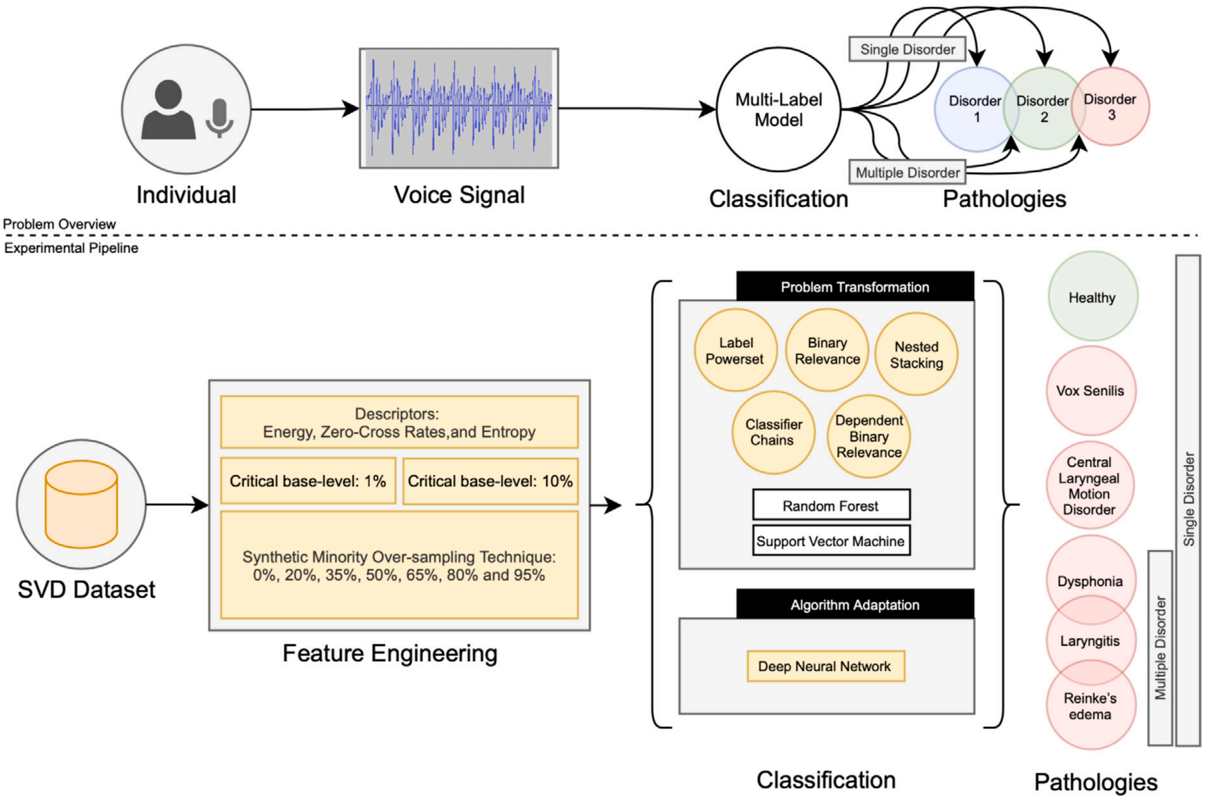


**Fig. 1.** Overview of problem and experimental pipeline to addressing multiple voice disorders in one individual using handcrafted features.

stricken by multiple voice pathologies concomitantly, demanding methods capable of addressing multiple classes of disorders. Particularly, multi-label classification (MLC) approaches, which have been applied in different research areas, such as molecular biology (Lin et al., 2013), clinical data (Zufferey et al., 2015; Wosiak et al., 2018), emotion, and sentiment analysis (Liu and Chen, 2015; Almeida et al., 2018), are concerned to associate instances with more than one conceptual class.

Consequently, this research paper explores the applications of MLC to identify a voice signal either as being healthy (HEA) or pathologically-affected with the following disorders: Dysphonia, Laryngitis, Reinke Edema, Vox Senilis, and Central Laryngeal Motion Disorder. Current literature shows that there are subjects affected by multiple pathologies at once, such as Dysphonia combined with Laryngitis, and Dysphonia combined with Reinke's edema, for instance. Consequently, we compared five problem-transformation methods for MLC, i.e., Label Powerset (LP), Binary Relevance (BR), Nested Stacking (NS), Classifier Chains (CC) and Dependent Binary Relevance (DBR), using Random Forest (RF) and Support Vector Machine (SVM) with three different kernels, i.e., Linear, Polynomial, and Radial Basis Functions, as base-learners, considering multiple pathologies in the same individual.

Additionally, by using the algorithm adaptation strategy, we took advantage of a Deep Neural Network (DNN) grounded on a Multi-Layer Perceptron to provide a fair comparison to problem-transformation approaches. Experiments were conducted with the original Saarbrücken voice dataset, containing 914 sample voices, complemented by augmented signals. Fig. 1 exposes an overview of problem and experimental pipeline. The most accurate result from problem-transformation corresponds to a value of accuracy of 93.76% when LP is combined with RF considering a balancing rate of 20% in the dataset $Set_1$, described ahead. Contrary to this, the DNN designed as a 5-Layer MLP achieved, under the same experimental setup, a value of accuracy of 91.60%. DNN outperformed the other methods when trained with a high number of augmented samples. The primary contribution of our research paper is the comparison of various multi-label classification approaches, as well as different feature engineering strategies, for the detection of multiple voice disorders in the same individual, particularly:

- exploring a real-life labelled dataset composed by multiple-disorders;
- evaluating the voice descriptors, and the importance of balancing and synthesizing samples.
- comparing problem transformation, and algorithm adaptation using different base-learners.

Notably, as detailed ahead in Section 3.1, the small number of voice samples available is one of the challenges that motivated our work. The precise identification of different voice disorders in the same individual is very tricky. Using a data-driven strategy, the dataset quality interferes directly with the results, requiring a robust collection of samples or alternative methods to support the creation of the automatic models. On one hand, we have to put the effort into creating a wide-range dataset, since we are in a scenario of a scarcity of data, dealing with a real-life dataset, labelled and composed with individuals with multiple disorders. This kind of dataset is **extremely rare and hard to be built** since the disorders affect one to another, which reduces the confidence of synthetic samples. On the other hand, we have several machine learning methods able to support the exploration of unbalanced and small datasets. We choose the latter option towards developing our research paving the way to discuss the performance of multi-label techniques in voice disorder identification, exploring Synthetic Minority Over-sampling Technique (SMOTE) in a multi-label context, handcrafted features and machine learning algorithms.

To allow for a better understanding of the concepts explained hereafter, the remainder of this paper is structured as follows: Section 2 reviews multi-label methods and comments on the related ML-based techniques applied to the detection of voice disorders, Section 3 details the proposed approach, as well as the metrics used for assessment, and, lastly, Sections 4–6 present the results, discussions, and conclusions, respectively.

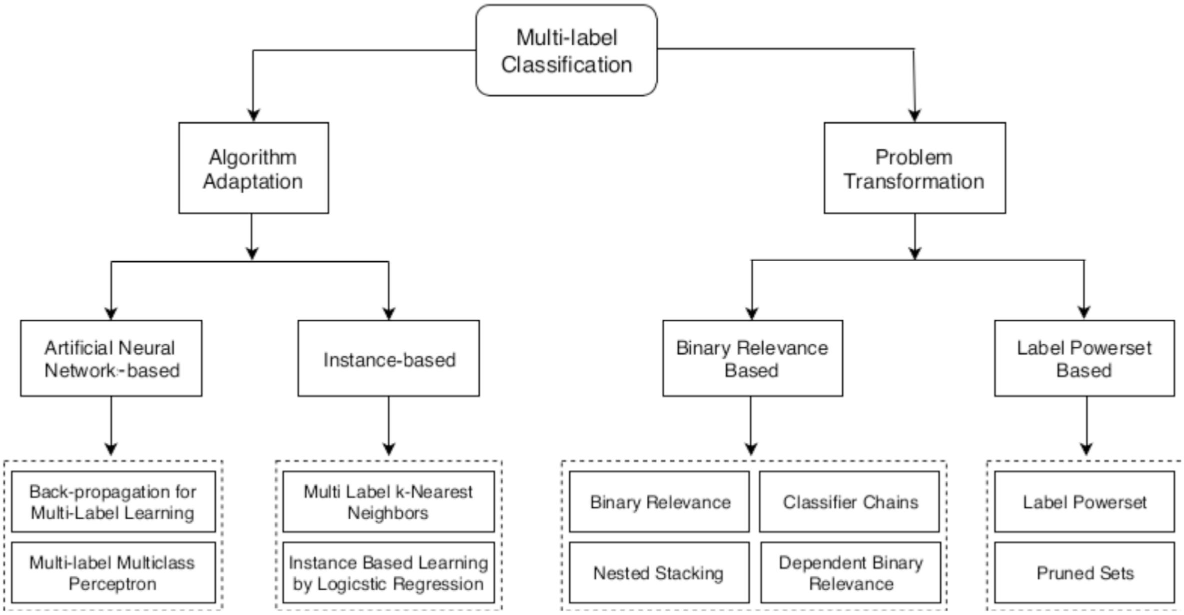

Multi-label Classification
Algorithm Adaptation
Problem Transformation
Artificial Neural Network-based
Instance-based
Binary Relevance Based
Label Powerset Based
Back-propagation for Multi-Label Learning
Multi-label Multiclass Perceptron
Multi Label k-Nearest Neighbors
Instance Based Learning by Logicstic Regression
Binary Relevance
Classifier Chains
Nested Stacking
Dependent Binary Relevance
Label Powerset
Pruned Sets

**Fig. 2.** Simplified taxonomy for MLC methods. After Tsoumakas et al. (2009).

## 2. Literature review

### 2.1. Multi-label classification

Let $L$ be a set of disjoint labels. In Single-Label Classification strategies, the main purpose is to learn from the samples that are related to a single label $l$ from $L$. On one hand, $|L|$=2 turns the problem to a binary classification problem. On the other hand, $|L|$> 2 characterizes a multi-class problem (Tsoumakas and Katakis, 2007). When each example is linked to a set of labels $Y \subseteq L$, then the learning task is defined as a multi-label classification. Fig. 2 exemplifies the taxonomy associated with MLC. MLC methods are split into two groups (Tsoumakas et al., 2009): (i) *Algorithm Adaptation*, and (ii) *Problem-Transformation*.

Algorithm Adaptation approach includes algorithms extending traditional single-label classifiers to deal with multi-label data directly, such as ANNs, Logistic Regression (LR) and k-Nearest Neighbours (k-NNs). Among algorithm adaptation strategies, those based on Multi-layer Neural Networks have been extended to cope with multi-label data. Modifications, such as adaptations in back-propagation learning algorithm (Zhang and Zhou, 2006; Wang et al., 2020; Lenc and Kral, 2016) or in weight updating strategies (Crammer and Singer, 2003) prove to be effective when using ANN for multi-label classification.

- **Deep Neural Network (DNN)**, the high predictive capacity provided by ANN paves the way for robust solutions, particularly, DNN. In Liu et al. (2017), for instance, a DNN was adopted to address the classification of documents in a text mining multi-label solution. It is important to mention that the successful application of DNN is strongly related to the scale of the largest datasets, i.e., a small dataset could not provide a proper training set for deep learning models.

Problem-Transformation approach consists of strategies which turn the original task into one or more algorithm-independent SLC routines, implying that any single-label classifier might be used as a base-learner. The base-learner is employed $k = 1, 2, 3, \ldots$ times for each possible label using all the subsets $D_i$ ($i = 1, 2, 3, \ldots,$ and $i \neq k$) for training and the subset $D_k$ for evaluation. Important examples are as follows.

- **Binary Relevance (BR)**, the simplest problem-transformation method. For each label, the method trains a binary classifier responsible for predicting whether the example contains that label or not (Tsoumakas and Vlahavas, 2007). This method reduces a given multi-label problem with $m$ labels to $m$ binary classification problems. More precisely, $m$ hypotheses $h_1, h_2, h_3, \ldots, h_m$ are induced, each of them being responsible for predicting the relevance of one label, using just $\mathcal{X}$ as the input space: $h_j : \mathcal{X} \rightarrow \{0, 1\}$. In the classification of a new unknown instance, the output is the union of labels classified with a positive indication. Therefore, since the labels are predicted independently, possible correlations between them are ignored.
- **Classifier Chains (CC)** (Read et al., 2011), which is grounded on the idea of partial conditioning. It means that, to predict the label $z_k$, the feature space is based on the original features, and also by the true label information from the previous labels $(z_1, z_2, \ldots, z_{k-1})$. Similarly to BR, this method trains $m$ binary classifiers for $m$ labels, although the feature space is different for each classifier. Particularly, CC tries to explore the correlation between the labels by using them as features. Since it is based on a chain of classifiers, the labels order influences on the global accuracy. Additionally, one label may be harder to predict than the others, implying that it should be placed at the end of the chain, while an easier-to-predict label should be at the beginning.
- **Nested Stacking (NS)** (Senge et al., 2013), which is a problem-transformation method similar to CC. During the training phase, a binary classifier learns each label by using the label from the previous classifier as a feature. The main difference between this method and CC is the label incorporated in the feature space. Whereas on CC the true label is used, NS uses the predicted label. Thus, the only classifier trained using the true label information is the first chain classifier. According to Senge et al. (2013), the predicted labels should be obtained through an internal out-of-sample method, such as $k$-fold cross-validation.
- **Dependent Binary Relevance (DBR)** (Montanes et al., 2014), which is based on BR. Essentially, DBR also trains one binary classifier for each label, however, the feature space is augmented with the labels that are not going to be predicted by the classifier being trained. Thus, to train a classifier $C_k$ in order to predict the label $y_k$, the feature space is composed by the original features and by the other labels $(y_1, \ldots, y_{k-1}, y_{k+1}, \ldots, y_m)$. To apply those classifiers to an unlabelled sample, multi-label strategies are required to create the labels to play the role of features, subsequently applying the DBR method.

- **Label Powerset (LP)**, which is an effective problem-transformation method. It takes every unique combination of labels in the multi-label training set and creates one label that represents this combination for a new SLC problem (Tsoumakas and Vlahavas, 2007). For a new instance, the single-label classifier predicts a metalabel, which is, indeed, a set of labels. The main drawback of LP method is the number of new labels it creates: considering $M$ unique combinations, it creates $2^M$ labels for the single-label classifier.

### 2.2. Related work on voice disorders sorting

The first scientific paper we reviewed, documented in Al-Naheri et al. (2017), concentrates on the development of a feature extraction method for detecting and classifying speech pathologies based on the analysis of different frequency bands. Besides entropy, the authors extracted the maximum peak amplitude from each frame of a voiced signal based on autocorrelation function. Particularly, distinct examples of the sustained vowel /a/ for both normal and pathologically-affected voices from three different databases in English, German, and Arabic were used. An SVM classifier was adopted, demonstrating significant difference between of both the types of signals. This is primarily due to the capability of SVMs to map lower dimensional feature space into higher dimensions. Moreover, due to Cover's theorem of separability, the classes that are nonlinearly separable in lower dimension feature space become linearly separable in higher dimensions (Cover, 1965).

The authors of paper Muhammad et al. (2012b) developed a feature extraction technique for automatic speech recognition (ASR) that combines a time–frequency analysis with a Gaussian Mixture Model (GMM) to distinguish speech pathologies. Data from 70 dysphonic subjects with six different voice disorders and 50 normal subjects were analysed. Mean values of accuracies of 97.48% and 99% were obtained for the text-independent and the text-dependent cases, respectively. Particularly, the authors verified that the proposed features outperformed the conventional Mel Frequency Cepstral Coefficients (MFCCs).

In paper Muhammad and Melhem (2014), the authors proposed an algorithm to classify pathologically-affected voices based on *Moving Pictures Expert Group 7* (MPEG)-7 audio low-level features. They specifically showed that MPEG-7 part-4 codes can accurately detect abnormalities. Their experiments were carried out on a subset of sustained vowels /a/, as in the word *dogma*, from MEEI speech corpus. For classification, they applied an SVM-based algorithm, achieving a value of accuracy of 99.99% with a standard deviation of 0.01%.

Additionally, the authors of paper Vikram and Umarani (2013) stated that MFCCs extracted from the phonemes /a/, /i/, and /u/ could be used as features in distinguishing normal from affected voices. Specifically, their system combined specific descriptors with a Gaussian Mixture Model-Universal Background Model (GMM-UBM) classifier. When 18 ordinary MFCCs served as input to the GMM-UBM classifier, the mean value of accuracy was 85.18%. Oppositely, when the wavelet-based MFCCs were used, 93.32% of the results were correct, indicating that the latter features enhance classification.

The authors of paper Akbari and Arjmandi (2014) developed a system to detect voice disorders using the Discrete Wavelet Packet Transformation (DWPT). They found that normal and pathologically-affected voices are well characterized on the basis of energy and entropy, extracted from a specific Wavelet-Packet tree with eight decomposition levels. The mean values of accuracy of 96.67% and 97.33% were obtained on the Kay Elemetrics database from entropy features, and wavelet packet-based energy, respectively.

In paper Hemmerling et al. (2016), the authors evaluated the usefulness of different methods for acoustic-based classification of laryngeal pathologies. First, a vector of 28 components was extracted from time-, frequency- and cepstral-domains when the sustained vowels /a/, /i/, and /u/, all at high, low, and normal pitches, were analysed. Ensuingly, they used Principal Component Analysis (PCA) to reduce the feature vector dimension and then choose the most effective acoustic features. When k-means clustering and Random Forest (RF) classifiers were adopted, full accuracy was obtained.

The authors of paper Martinez et al. (2012) presented a set of experiments on the detection of pathologically-affected voices over SVD corpus by using the MultiFocal toolkit for a discriminative calibration and fusion of features. A generative GMM trained with MFCCs, harmonics-to-noise ratio, normalized noise energy and glottal-to-noise excitation ratio was used as the classifier. Grouping different recordings from each speaker the proposal obtained performance over 90% accuracy.

The authors of paper Saeedi and Almasganj (2013) created a wavelet feature extraction method in which, instead of default filterbanks, dynamic bases were used and applied to extract features from the voice signals. Orthogonal wavelets were adjusted via lattice structure and, then, the best parameters were investigated through an iterative task based on a Genetic Algorithm (GA). An SVM was adopted to classify the signals, revealing that paralysis, nodules, polyps, edema, spasmodic dysphonia, and keratosis were completely catalogued.

The authors of paper Mekyska et al. (2015) presented a study of strategies for SPD focused on parametrization. They provided 92 widely used speech features and, additionally, a few which had not yet been tried. The significance of these descriptors was tested on three voice databases based on classification accuracy, sensitivity, and specificity. For the *Príncipe de Asturias (PdA) Hospital in Alcalá de Henares of Madrid* database, the authors made improvements in classification accuracy using a single-classifier approach. All the features introduced were identified by the *Mann–Whitney U-test* with at least one of the databases. Among the descriptors, those which especially quantify hoarseness and breathiness were the most prominent candidates to identify pathologically-affected voices.

In paper Ali et al. (2016), researchers developed an automatic speech pathology classification system with text-dependent running speech. According to them, three psychophysics conditions of hearing, i.e., critical band spectral estimation, equal loudness hearing curve, and the intensity loudness power law of hearing, were employed to compute the auditory spectrum. The all-pole models of the auditory spectra were evaluated in conjunction with a GMM classifier. Using Massachusetts Eye and Ear Infirmary (MEEI) database, a level of accuracy of 99.56% was obtained.

In paper Markaki and Stylianou (2011), the authors explored signals from acoustic and modulation frequency representations for detecting and discriminating voice pathologies. From that dimension-reduced representations, a feature selection based on the mutual information between voice classes and features was performed. For SPD, the system achieved a mean value of accuracy of 94.1%, considering a 95% confidence interval. Additionally, for pathology classification, the proposed system significantly outperformed the results obtained with cepstral-based features.

Specific techniques, such as those mentioned in Pranav and Sabarimalai (2017), use identification of glottal instants and electroglottography (EGG) parameters to identify multiple types of voice pathologies, including neurological, functional and laryngeal. Initially, the authors presented an adaptive variational mode decomposition-based method for suppressing low frequency artifacts and additive high frequency noise. Then, the algorithm built a novel EGG feature signal to determine the glottal closure and opening instants. Proceeding, the novel glottal instants were confirmed by computing the positive and negative zero-crossings around. Thus, the algorithm significantly outperformed existing systems for both noise-free and noisy-EGG inputs.

Interestingly, paper Sasou (2017) reports the automatic detection of speech pathologies to enable non-invasive and objective assessments based on the inspection of roughness, breathiness, asthenia, and strain. The proposed method adopted Higher-Order Local Auto-Correlation (HLAC) features, which were calculated from the excitation source signal obtained by an automatic topology-generated analysis. Additionally, the authors of paper Verde et al. (2018b) created a personalized methodology to estimate the fundamental frequency ($F_0$) of dysphonic voices. They found that the personalization supports two of the main factors that influence $F_0$, i.e., the subject's gender and age, allowing for a better distinction between normal and pathologically-affected voices. To evaluate their methodology, they carried out a set of tests and compared its classification ability to other algorithms documented. The results obtained showed that the authors' technique provided an acceptable value of accuracy of 77%.

In paper Lachhab et al. (2014), the authors proposed a simple and fast method to detect voices disordered due to esophageal constriction. A continuous speech recognition technique based on GMM and Hidden Markov Model Toolkit (HTK) platform was applied, in a speaker-dependent mode, over the French Pathological Speech Database. The acoustic vectors were linearly transformed by using Heteroscedastic Linear Discriminant Analysis (HLDA), which refitted them into a smaller space with good discriminative properties. The mean obtained value of accuracy of 63.59% was considered very promising, assuming that esophageal voices contain unnatural sounds difficult to understand.

The authors of paper Zhong et al. (2016) presented an intelligent approach for vocal folds damage detection based on Hidden Markov Models (HMMs). They showed that particularly-transformed pathologically-affected voice signals follow a Gaussian distribution and demonstrated that a type-2 fuzzy membership function (MF) is capable of finding them, stimulating the application of a nonlinear signal processing technique to handle the problem. The authors also observed that the Short-Time Fourier Transform (STFT) of a disordered voice usually fades at a rate that can be used as an identifier for SPD. Lastly, two fuzzy machines, a Bayesian technique and a linear classifier were used in conjunction with the phonemes /a/ and /i/ to distinguish normal from disordered voices. Simulation results showed that the type-2 fuzzy classifier outperforms the other strategies.

Importantly, paper Fonseca and Pereira (2008) presents a Least-squares SVM (LS-SVM) classifier using a radial basis function (RBF) kernel that led to an adequate larynx pathology classifier. A value of accuracy of 90% was attained considering two classes, i.e., normal voices and those collected from subjects with vocal fold nodules. In complement, the mean value of accuracy of 85% was reported in distinguishing normal voices from those affected by Reinke edema. Lastly, 8 in each 10 results were correctly labelled, considering nodules, and Reinke edema only.

The above-referenced scientific papers, for which an overview is provided in Table 1, essentially focus on vocal fold derived-information to characterize a subject's condition, i.e., healthy or pathologically-affected. Taking advantage of the previous findings, mainly those documented in Akbari and Arjmandi (2014) and Pranav and Sabarimalai (2017), we selected energy, zero-crossing rates, and entropy as the features to analyse five pathologies, considering that a subject might have more than one voice disorder at the same time, justifying the applications of MLC algorithms. We also compared the application of SVM and RF as base-learners, since those algorithms have a relevant accuracy for different problems. To the best of authors' knowledge, the experiments and results we report in this paper provide an original contribution to the SPD field.

## 3. Materials and methods

In this work, we selected five problem-transformation strategies and one algorithm adaptation method. The problem-transformation MLC methods, i.e., LP, BR, CC, NS, and DBR were chosen due to their notable performance in previous works (Zufferey et al., 2015; Wosiak et al., 2018) and implemented using R language[2] and the `utiml` package[3] (Rivolli and Carvalho, 2018) along with their default hyperparameter values. It is important to mention that RF and SVM were compared to support a discussion on accuracy specifically considering the application of ML algorithms. Our algorithm adaptation implementation was based on artificial neural networks (Wang et al., 2020; Lenc and Kral, 2016; Ji et al., 2020). The multi-layer perceptron network (MLP) was constructed using Keras (Chollet, 2018) for computational speed boost. We propose an MLP with five hidden layers ($n$-256-128-64-7), where $n$ is the size of $n$-dimensional feature vector.

We employed the binary cross-entropy as loss function, adaptive moment estimation (adam) as optimizer and rectified linear units (ReLU) as the classification function in our DNN to compute classification score. In addition, we applied batch normalization after ReLU for smoothing and improving the final predictive performance on test set. To provide a fair comparison between the results from problem-transformation and algorithm adaptation, we induce the models based on the same set of features, i.e., $Set_1$ and $Set_2$ described ahead, and evaluation strategy.

### 3.1. Voice disorders database

The experiments were carried out by using SVD database (Barry and Putzer, 2007), which was created by specialists from the *Institut für Fonetik of the Universität des Saarlandes*.[4] The dataset contains 914 voice signals diagnosed as healthy (HEA) or pathological. It is important to highlight that this dataset comprises data from subjects with single and multiple pathologies: Dysphonia (DYS), Laryngitis (LAR), Reinke's edema (RDE), Vox Senilis (VSE), Central Laryngeal Motion Disorder (CLMD), both Dysphonia and Laryngitis (DYS-LAR) and, also, both Laryngitis and Reinke's edema (LAR-RDE). Originally, the signals were sampled at 50000 samples per second, mono-channel, 16-bit, without compression. To the best of the authors' knowledge, this dataset is the only one in the literature containing multiple voice disorders in the same individuals.

Table 2 shows the way signals were distributed in the original and the augmented datasets, being the latter obtained upon the applications of SMOTE (Chawla et al., 2002). Originally, the healthy samples are the majority since many voice pathologies are rarely observed in daily life, compromising the creation of a balanced dataset. Thus, the prediction task suffers from a lack of balance (Haixiang et al., 2017). Particularly, SMOTE has overcome the imbalance dataset problem, as in Georgoulas et al. (2007), Lee et al. (2013), Krawczyk et al. (2015), Potharaju and Sreedevi (2016) and Saarela et al. (2019). We experimented balancing rates, i.e., the rates of synthetic samples, from 20% to 95% just to observe the impact on the prediction.

In our experiments, we used just the /a:/ vowels, using a 10-fold stratified cross-validation strategy. The stratified cross-validation method is a variant of cross-validation that ensures that the class distributions of the folds are similar to the class distributions of the entire data. The augmentation procedure was only conducted in the training folds to avoid overfitting and fairness in our experiment. This can help to ensure that the model is trained on a representative sample of the data and that its performance is not biased towards the majority class. The adoption of stratified cross-validation together with SMOTE data augmentation can be an effective approach for reducing bias in machine learning models, particularly in cases where the dataset is imbalanced, as shown in Batista et al. (2004). Particularly, by ensuring that the model is trained on a representative sample of the data and that the minority class is not underrepresented during training, we can improve the model generalization performance and reduce the risk of bias. In addition, it is important to mention that, although SVD contains more than two thousand voice samples, with more than half being

[2] https://www.r-project.org

[3] https://github.com/rivolli/utiml

[4] http://www.stimmdatenbank.coli.uni-saarland.de/

**Table 1**
Overview of the related work, where, in the third column, MVDSI means "multiple voice disorders in the same individual". Notably, no previous work handles MVDSI, as the proposed approach does.

| Authors and references | Main approaches and tools | MVDSI (yes/no) |
|---|---|---|
| Al-Naheri et al. (2017) | feature extraction; frequency bands; SVM | no |
| Muhammad et al. (2012b) | feature extraction; GMM; MFCC | no |
| Muhammad and Melhem (2014) | MPEG-7 features; SVM | no |
| Vikram and Umarani (2013) | MFCC; GMM-UBM | no |
| Akbari and Arjmandi (2014) | DWPT; energy; entropy | no |
| Hemmerling et al. (2016) | cepstrum; PCA; random forest; K-means | no |
| Martinez et al. (2012) | GMM; MFCC; glottal-to-noise ratio | no |
| Saeedi and Almasganj (2013) | wavelets; GA; SVM | no |
| Mekyska et al. (2015) | Mann–Whitney U-test; parametrization | no |
| Ali et al. (2016) | psychophysics; GMM | no |
| Markaki and Stylianou (2011) | modulation-related features | no |
| Pranav and Sabarimalai (2017) | glottal instants; EGG features | no |
| Sasou (2017) | HLAC; jitter; shimmer; neural nets | no |
| Verde et al. (2018b) | gender; age; fundamental frequency | no |
| Lachhab et al. (2014) | GMM; HLDA | no |
| Zhong et al. (2016) | HMM; fuzzy MF; STFT | no |
| Fonseca and Pereira (2008) | LS-SVM; RBF kernels | no |

**Table 2**
Dataset distribution of samples and classes without balancing and with several balancing rate. After Barry and Putzer (2007).

| Balancing rate | HEA | Pathology | | | | | | | Samples |
|---|---|---|---|---|---|---|---|---|---|
| | | DYS | LAR | RDE | VSE | CLMD | DYS-LAR | LAR-RDE | |
| 0% (Original) | 686 | 69 | 81 | 33 | 22 | 10 | 4 | 9 | 914 |
| 20% | 686 | 137 | 137 | 132 | 132 | 130 | 136 | 135 | 1625 |
| 35% | 686 | 207 | 162 | 231 | 220 | 240 | 240 | 234 | 2220 |
| 50% | 686 | 276 | 324 | 330 | 330 | 340 | 340 | 342 | 2968 |
| 65% | 686 | 414 | 405 | 429 | 440 | 440 | 444 | 441 | 3699 |
| 80% | 686 | 483 | 486 | 528 | 528 | 540 | 548 | 540 | 4339 |
| 95% | 686 | 621 | 648 | 627 | 638 | 650 | 648 | 648 | 5166 |

diagnosed as pathologically-affected. We used just 228 pathologically-affected voice samples (comprising single and multiple pathologies in the same individual) and 686 voice samples without pathology. It should be noted that each voice signal was collected from an individual.

### 3.2. Feature extraction for voice analysis

In the proposed approach, specific handcrafted features were extracted from the subjects' voices in order to describe their acoustic apparatus (Al-Nasheri et al., 2018; Shilaskar et al., 2017; Hegde et al., 2019). We extracted energy (SE), zero-crossing rates (ZCRs) and entropy (SH), which provide a joint time–frequency information map, as explained in papers Guido (2016a), Guido (2016b), and Guido (2018), respectively. Interestingly, as shown in paper Guido (2016a)-pp. 277–280, energy is one of the most elementary features used to describe the workload performed by speech-related biological entities: the lungs and the vocal organs. Moreover, hearing is the process of detecting energy (Zhau et al., 2001). In particular, due to the theory of hearing, and the mathematical model used to describe cochlea, this detected energy is processed as an amplitude and frequency modulation (AM-FM) signal representing the output of cochlear filterbank.

Complementary, energy measures have advantages over strictly spectral features because they are more robust to different kinds of transmission and recording variations (Guido, 2016a). Notably, in case of the algorithms designed in this paper, energy is relevant because it expresses the effort the speakers' lungs perform to utter, as a function of time. As demonstrated by the tests and results in paper Guido (2016a), signal energy can certainly be considered a useful feature for pattern recognition in speech and voice analysis, among others, even associated with modest classifiers (Guido, 2016a). Another interesting aspect is that, being simple, energy allows for the proposed features to be interpretable, what does not holds true in case of learned features (Guido, 2016a). Thus, energy consideration is extremely relevant for both speech production and perception.

Additionally, normalized ZCRs, as shown in paper Guido (2016b)-pp. 257–258, can be interpreted as specific neuronal structures, exhibiting a neural-like behaviour. ZCRs are of paramount importance because they carry information on formant frequencies, as shown in the study on integration, differentiation, and clipping speech signals (Licklider, 1948). As explained in paper Guido (2016b), ZCRs are extremely simple to be computed, with a linear order of time and space complexities. Naturally, they reveal spectral information on input data without an explicit conversion from time to frequency-domain, reducing computation time (Guido, 2016b).

Furthermore, relevant speech processing problems have frequently benefited from them. Years ago, transient signals were analysed to demonstrate they can be accurately found based on zero-crossings and, subsequently, applications related to the estimate of epochs in speech signals were performed, confirming that assumption. Word boundary detection, distinctions of voiced and unvoiced signals, speech recognition, speech pathology detection, and related applications have also benefited from ZCRs, as mentioned and demonstrated in paper Guido (2016b).

Lastly, as shown in paper Guido (2018)-pp. 165, entropies computed as described ahead show a close relationship with DNNs, possibly strengthening their statistical richness. Particularly, entropy obtained based on the proposed approach for feature extraction, has a flagrant potential, as also evidenced by the authors of different scientific papers such as Techakesari and Ford (2013), Xia and Xu (2012), Zarinbal et al. (2015), Zhang et al. (2016), Rallapalli and Alexander (2015). Similarly to the characterization of ZCRs as neurocomputing agents, entropy is shown to be the outcome of a specifically tuned deep neural network (DNN) that fuses important information (Guido, 2018), bringing a significant value to our experiments and allowing for more interpretable outcomes. Furthermore, applications on restricted-vocabulary speech recognition found in paper Guido (2018) reassure the efficacy of this feature.

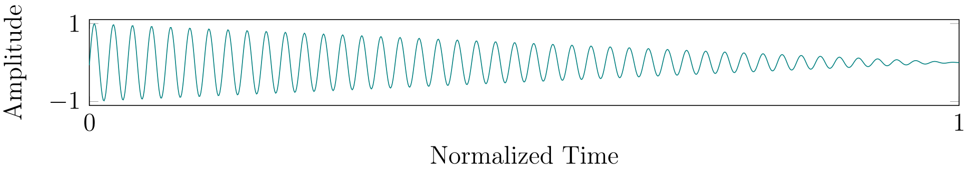


**Fig. 3.** Impulse response shape of the $i$th formant $h_i[n]$.

Consequently and as complemented ahead, these are the reasons why those features have been selected to work in conjunction in our experiment. Further details about those features follow.

### 3.2.1. Signal energy

If $s[\cdot]$ is the $M$-sample long digital signal under analysis, then SE, which is related to the potential to perform work (Salehi, 2015), is computed as in Eq. (1).

$$SE(s[\cdot]) = \sum_{i=0}^{M-1} (s_i)^2 \quad . \tag{1}$$

As exposed in Guido (2016a), i.e., the paper used to guide the application of SE in this experiment, a critical base-level ($0 < C < 100\%$) is required to conduct the analysis according to the method $A_3$ defined in that reference. In our experiments, we compared the effect of $C = 1\%$ and $C = 10\%$, i.e., a fine resolution and a wide resolution, producing 99 and 9 features, respectively.

### 3.2.2. Signal ZCRs

Different from SE, ZCR is related to the fundamental frequency ($F_0$) contained in $s[\cdot]$, being useful to analyse the spectral stability of phonation. It is computed as shown in Eq. (2), where $ZCR(s[\cdot]) \geqslant 0$, and $sign(y) = \begin{cases} 1, & \text{if } y \geq 0 \\ -1, & \text{if otherwise} \end{cases}$, and according to method $B_3$ defined in paper Guido (2016b) either using $C = 1\%$ or $C = 10\%$.

$$ZCR(s[\cdot]) = \frac{1}{2} \sum_{j=0}^{M-2} |sign(s_j) - sign(s_{j+1})| \quad . \tag{2}$$

Notably, ZCRs are particularly important for speech analysis and even speech perception. According to current literature, we can consider the input speech signal $s[\cdot]$ as the output of a linear time-invariant (LTI) system. Hence, if $p[n] = \sum_{k=-\infty}^{\infty} \delta[n-k]$ represents the impulse-train excitation source, and $h[\cdot]$ represents the vocal tract impulse response, then,

$$s[\cdot] = p[\cdot] * h[\cdot] \quad , \tag{3}$$

where $h[n] = h_1[n] * h_2[n] * ...$, being $h_i$, for $i = 1, 2, ....$, the impulse response of the vocal tract formant frequencies. Considering a second-order resonator filter in the Z-domain, then, $h_i[n]$ becomes

$$H_i[z] = \frac{b_{0i}}{a_{0i} + a_{1i}z^{-1} + a_{2i}z^{-2}} = \frac{b_{0i}}{(1 - p_{1i}z^{-1})(1 - p_{2i}z^{-1})} \quad ,$$

where $p_{1i} = \gamma_i e^{j\omega_{\gamma_i}}$ and $= p_{2i} = \bar{p_1}_i$ correspond to a complex conjugate pole-pair in the Z-plane. By using partial function expansion and the inverse Z-transform, we can observe that the impulse response of the $i$th formant is, for a stable resonator, like a damped sinusoid, as in Fig. 3. In particular,

$$h_i[n] = \frac{{\gamma_i}^n b_{oi}}{sin(\omega_{\gamma_i})} sin(\omega_{\gamma_i}(n+1))u(n+1) \quad ,$$

where $u(n+1)$ is the unit step function at point $(n+1)$.

Considering only the first formant model, we have, from Eq. (3) and according to Fig. 4,

$$s[n] = p[n] * h_1[n] \quad . \tag{4}$$

From Fig. 3, we can observe that zero-crossings in $h_i[n]$ dictate its period of oscillation and, hence, its reciprocal, i.e., $\frac{1}{ZCR}$, thus defining the $i$th formant frequency. Similarly, from Eq. (4) and Fig. 4, we can note that the same zero-crossings in $h_1[\cdot]$ are present in the output voiced speech waveform. Thus, zero-crossings in speech signals are related to the formant frequencies, even though other specific signal characteristics are severely damaged. Moreover, due to the experiments performed by Manfred Robert Schroeder (Schroeder, 1966), we know that human beings emit and perceive sounds by emitting spectral peaks, i.e., formants, not spectral valleys, i.e., anti-formants or valleys in the vocal system transfer function. Thus, ZCRs are important not only for speech analysis but also for speech production and perception.

Furthermore, according to the non-linear source-filter interaction theory (Quatieri, 2008)-pp. 153–161, whenever the glottis opens, there is a sudden change in −3 dB bandwidth ($B_1$) of the corresponding first formant ($F_1$), depending on the glottal geometry and according to aerodynamics and Bernoulli's principle. However, glottal geometry may change due to the damage caused by particular laryngeal pathologies, thus modifying $F_1$ and $B_1$ which causes, consequently, changes in the zero-crossings of $h_1[\cdot]$ and in the speech waveform. Therefore, ZCRs offer very important acoustic cues to detect fine details related to pathological issues in the vocal mechanism.

### 3.2.3. Signal entropy

Lastly, signal predictability is measured in this experiment by using SH, defined in Eq. (5) and in method $C_2$ of paper Guido (2018), where $p_i$ is the probability of the $i$th value in a set with $K$ signal values and $\beta = 2$ is the basis employed. We used a sliding window of length $L = \frac{C \cdot M}{50}$, adopting an overlapping rate of $V = 50\%$ which produces a feature vector of length $T = \left\lfloor \frac{100 \cdot M - L \cdot V}{(100-V) \cdot L} \right\rfloor$. Specifically, $SH$ is defined as:

$$SH(s[\cdot]) = -\sum_{i=0}^{K-1} p_i \cdot log_\beta(p_i) \quad . \tag{5}$$

Notably, there have been a considerable number of scientific papers dealing with entropy in the field of speech analysis. Particularly, as observed by the authors of paper Vinay and Bharathi (2019), maximum and minimum entropy values represent the presence of flat distribution of noise and clean speech, respectively. Entropy is also used by the authors of paper Babatsouli et al. (2016) to measure the "peakiness" of a speech spectral distribution, being a very useful feature for speech analysis and recognition. Specifically, a peaky spectrum with information on formant structure of voiced sounds is expected to present low entropy, whereas a flatter spectrum matching noisy regions of speech is expected to exhibit a higher entropy (Babatsouli et al., 2016). Entropy is observably guaranteed to be more sensitive than other linear measures for dysfluency identification, as observed by the authors of paper Misra (2004), where entropy is adopted to measure the degree of noisiness found in speech and audio signals.

Summarizing, we use and compare two different sets of feature vectors which were defined based on SE, ZCR, and SH. They are called $Set_1$ and $Set_2$, where the former and the latter correspond to feature extraction procedures using $C = 1\%$ and $C = 10\%$, respectively. We created these two different sets to provide a fair comparison between different problem-transformation and algorithm adaptation methods.

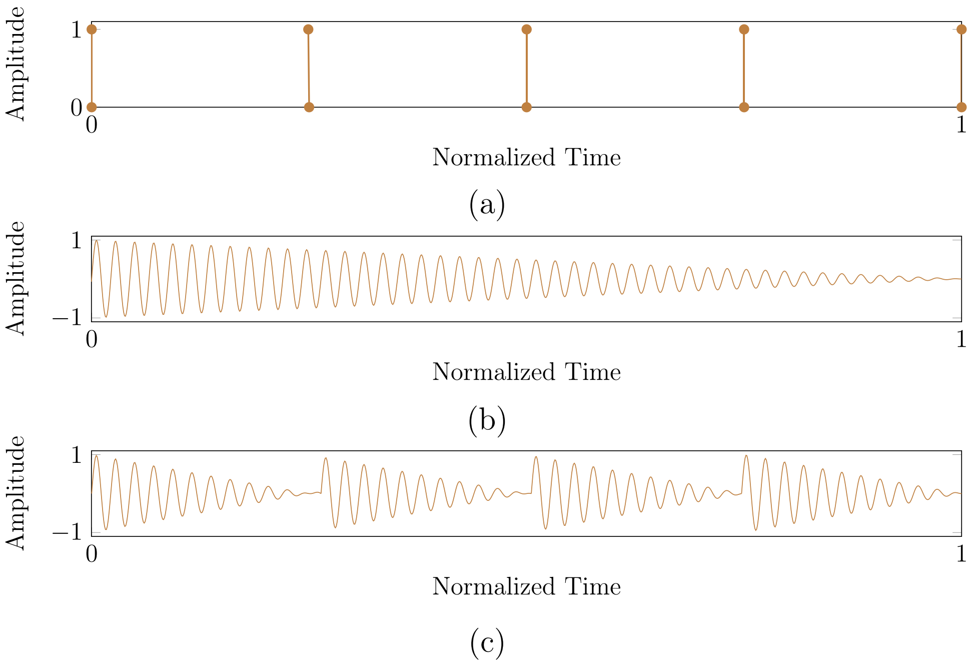


**Fig. 4.** (a): unit-sample impulse source $p[\cdot]$; (b): impulse response shape of the first formant $h_1[\cdot]$; (c): voiced speech signal $s[\cdot]$ resulting from the convolution of $p[\cdot]$ with $h_1[\cdot]$.

### 3.3. Base-learners selected

Since all MLC methods used in this work are problem-transformation strategies, an associated base-learner is required. RF (Breiman, 2001) and SVM (Vapnik, 1995) classifiers were selected because they are well known algorithms with relevant results in different problems. Furthermore, a relevant number of SPD algorithms has employed SVMs for building their classification models, as observed in sub- Section 2.2.

Particularly, the `randomForest` R package was used in our experiments, relying on default hyperparameters. Accordingly, based on R package `e1701`, three kernel functions, i.e., linear-kernel (L-SVM), polynomial-kernel (P-SVM), and radial-kernel (R-SVM), were associated with an SVM to assess the linearity of the problem being treated. All the implementation code used in the experiments is available online.[5]

### 3.4. Evaluation metrics

The predictive values of accuracy of the methods we propose were assessed by using a 10-fold cross validation strategy. Two different baselines were also adopted in the experimental setup: a model that always recommends the majority class, i.e., `Majority`. `Random`, was another baseline which represents a model that provides random recommendations.

Four of the most used multi-label classification example-based metrics were adopted: accuracy obtained as 1-Hamming loss, precision, recall, and F1-score. Given a dataset with $m$ instances, they were computed by using the following equations (Godbole and Sarawagi, 2004), Pereira et al. (2018):

$$accuracy = 1 - \frac{1}{m}\sum_{i=1}^{m}\frac{|Z_i \Delta Y_i|}{|L|} \quad , \tag{6}$$

$$precision = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \cap Z_i|}{|Z_i|} \quad , \tag{7}$$

$$recall = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \cap Z_i|}{|Y_i|} \quad , \tag{8}$$

$$F1\text{-}score = \frac{1}{m}\sum_{i=1}^{m}\frac{2|Y_i \cap Z_i|}{|Y_i| \cup |Z_i|} \quad , \tag{9}$$

where $Y_i$ represents the $i$th instance of the true set of labels, $Z_i$ represents $i$th instance of the predicted set of labels, and $\Delta$ represents the symmetric difference.

**Table 3**
Average accuracy (using RF and SVM) for each label considering the MLC problem transformation methods. The values in bold correspond to the best method for each label.

| Label | Method | | | | |
|---|---|---|---|---|---|
| | LP | BR | DBR | CC | NS |
| HEA | **71.10%** | 70.43% | 70.39% | 69.95% | 69.32% |
| CLMD | 96.35% | **96.88%** | 96.78% | 96.55% | 96.65% |
| DYS | **90.75%** | 90.07% | 89.43% | 90.12% | 89.81% |
| LAR | **86.76%** | 84.86% | 85.17% | 83.96% | 85.05% |
| RDE | **90.89%** | 90.19% | 89.72% | 87.88% | 89.53% |
| VSE | 94.91% | 95.48% | **95.63%** | 92.53% | 95.19% |
| Average | 88.46% | 86.31% | 87.85% | 86.83% | 87.59% |

## 4. Tests and results

The tests and results presented in this section are organized to support some comparisons and insights covering: (a) different values of accuracy from MLC methods and disorders prediction, (b) base-learners inductive values of accuracy and balancing and, complementarily, (c) related issues.

### 4.1. MLC predictive assessment for disorder prediction

All MLC methods achieved suitable predictive results. Table 3 presents the average values of accuracies for both base-learners, i.e., RF and SVM, embedded in the MLC methods, i.e., LP, BR, DBR, CC, and NS, over datasets $Set_1$ and $Set_2$.

This experiment took into account the results obtained for each disorder and all MLC problem-transformation methods, where an average

[5] http://www.uel.br/grupo-pesquisa/remid/?page_id=145

value of accuracy of 87.41% was obtained: 88.46% for LP, 86.31% for BR, 87.85% for DBR, 86.83% for CC, and 87.59% for NS. The highest value per label is highlighted in bold in Table 3. CLMD, DYS, RDE and VSE obtained a value of accuracy higher than 90% with LP. CC and NS, in average, have not achieved the best values of accuracy for any of the labels. The healthy condition, i.e., HEA, attained the worst accuracy among all the labels, disregarding the method.

The fact that healthy samples presented the lowest accuracy is related to the variability under healthy conditions that lead to misclassifications, intrinsically related to the false positive rate obtained. However, with a close look at the results by using F1-score as the main metric, particularly involving problem transformation methods such as Label Powerset (LP), as visible in Table 6, we note that the healthy samples were classified with higher performance than the disorders one. In this way, healthy voices demand suitable classification methods in a multi-label scenario. To face that challenge, we proposed more competitive methods, improving the detection of such a pattern. In addition, we emphasize that SMOTE, as we adopted, is capable of providing us with the expanded dataset we used, allowing for F1-score to be sufficiently convincing in view of the restricted data samples we have. Furthermore, we once again observe that we are working with a restricted dataset because **no other exists** with the characteristics we are exploring in this paper. Moreover, such data imbalance problem is a vexing issue in several other pattern classification problems especially dealing with the medical-domain data.

When comparing the disorders, CLMD achieved the best value of accuracy, close to 100%. Contrary to this, LAR was the most difficult pathology to be predicted, achieving, in the optimal case, 86.76% of accuracy. It is important to mention that the disorders with multiple labels, i.e., LAR-DYS and LAR-RDE, obtained the lowest values of accuracy, but all of them were superior to HEA. DYS and RDE achieved their best accuracy, i.e., 90.75% and 90.89%, respectively, when predicted by using LP. It is worth mentioning that, even with few data samples in the original dataset, the experiments exposed different patterns from these combinations of multiple disorders. Likewise, the predictive performance increased when using SMOTE to expand the original set of samples.

### 4.2. Machine learning inductive assessment and balancing improvements

In this Section, we discuss the predictive performance of ML algorithms: RF and SVM as base-learners for problem-transformation methods, and a DNN with five-layers as the algorithm adaptation method (Sorower, 2010).

F1-score for LP, considering both feature groups and varying the balancing rate (r), are presented in Table 4. To reassure the importance of balancing, we can observe the LP results: without balancing, there are minimal differences between the values of accuracy of the base-learners and Majority baseline but, after balancing, these differences increase considerably, reducing the F1-score of Majority baseline to 0.4218 and improving that of RF to 0.7926, even with the smallest rate of balancing, i.e., 20%. For the highest balancing rate, Majority baseline drops to 0.1326, performing worse than Random, whereas RF accuracy improves to 0.9262.

For problem transformation methods, the precision was smaller than the corresponding values of accuracy. Complementary, we observe that all the methods achieved their best results using RF as their base-learners. Following RF, L-SVM achieved the second best predictive value of accuracy. SVM with radial kernel was the third and then the polynomial kernel. Moreover, all the base-learners obtained better results than the baselines. Regarding all methods and base-learners, the association of LP with RF achieved the best results, with a value of accuracy of 93%, a value of precision and recall of 82% and an F1-score equals to 0.82.

For BR, DBR, CC, and NS, the same occurrence holds true: with the original dataset, i.e., without balancing, the label with more examples, i.e., Majority, presents approximately the same value of accuracy in comparison with ML algorithms. However, with a balanced dataset, the F1-score of RF was higher than that of Majority.

Our DNN model was composed of five dense layers with 99, 256, 128, 64, and 6 neurons, respectively. Each dense layer is followed by batch normalization, except the last one, i.e., the output layer. Relu was adopted as the activation function for all layers except the output layer, which uses sigmoid activation. The model has a total of 190342 trainable parameters which were learned by using Adam optimiser with a learning rate of 0.001 and binary cross entropy as the loss function. We adopted a batch size of 60 samples and trained the model for 1000 epochs. We also used Repeated K-Fold Cross Validation with 10 folds to evaluate the performance of our model.

Algorithm adaptation, i.e., the proposed DNN model, was capable of overcoming the problem transformation methods. Table 5 shows the obtained F1-score with several augmentation rates over both the datasets. However, our proposed method could not converge towards delivering predictions for all possible class combinations using the original dataset due to imbalance issues. F1-score boost over both datasets when the augmentation rate grew. Based on 20% of balancing, the DNN achieved an average F1-score of 0.906. Additionally, DNN was superior to all base-learners, disregarding the method, using all rates of balancing. As expected, F1-score rose following the balancing rate, achieving an average F1-score of 0.963 using both $Set_1$ and $Set_2$ when applying 95% of balancing rate.

The previous results highlight some drawbacks found when dealing with unbalanced datasets, such as the sensitivity of multi-label metrics to the label distribution in the dataset (Zufferey et al., 2015). Therefore, dataset balancing was required to support reliable and improved results. Based on SMOTE with a balancing rate of 20% in Majority, we were capable of reducing the unbalancing problem and providing classification improvements with MLC methods. Since this dataset needs the least amount of synthetic examples among those we experimented and is the closest to the original dataset, further analyses of this work were carried out and supported by that balancing rate.

### 4.3. Related issues

Historically, laryngitis is known to be a serious research issue for speech technology problems, in particular for speaker recognition (Doddington et al., 2000). Thus, it is interesting to note that our finding of most negative performance for laryngitis may explain this. Since our task includes the prediction of two states, it is important to know which label was the most difficult to detect. Particularly, when observing further results related to LP with 20% of balancing rate, we note that Laryngitis was the label presenting most false negatives, i.e., 115 for $Set_1$ and 127 for $Set_2$. Lastly, when comparing the critical base-levels of energy, i.e., $C = 1\%$ and $C = 10\%$, the former was clearly the best option, as shown by the results obtained with RF for all MLC methods in Table 6. Corroborating with RF, DNN obtained the highest performance with 0.916 and 0.897 of F1-score for $Set_1$ and $Set_2$, respectively. The average values of F1-score for $Set_1$ and $Set_2$ were 0.832 and 0.745, respectively. Besides, for each method, the average value of F1-score for $Set_1$ was higher than that obtained with $Set_2$.

A relevant evaluation among the possible strategies is to consider the number of produced models. Different from the single-label classification that generates models in the same number of targets, when addressing multi-label classification, some methods could increase the number of models, requiring more computational resources and time to train the solution. In Mastelini et al. (2019), the authors proposed a metric named Counting of Trained Regression Models (CTRM) for multi-output problems. In our domain, by counting models for classification problems, we can compare the solutions in terms of the number models trained. Using this perspective, LP and DNN generate the same number of single-label classification. LP increases the number of classes considering all possible label combinations (i.e. in our domain, from 5

**Table 4**
F1-scores for all base-learners, methods and each balancing rate (r) of SMOTE for both the datasets, i.e., $Set_1$ and $Set_2$.

| Method | Dataset | Classifier | | | | | |
|---|---|---|---|---|---|---|---|
| | | RF | L-SVM | P-SVM | R-SVM | Majority | Random |
| LP | Original | 0.7430 | 0.7503 | 0.7481 | 0.7503 | 0.7503 | 0.1530 |
| | 20% (r) | 0.7926 | 0.5805 | 0.4799 | 0.5796 | 0.4218 | 0.1854 |
| | 35% (r) | 0.8388 | 0.5900 | 0.4312 | 0.5869 | 0.3087 | 0.1988 |
| | 50% (r) | 0.8779 | 0.5755 | 0.4129 | 0.5958 | 0.2309 | 0.2141 |
| | 65% (r) | 0.9005 | 0.6000 | 0.4608 | 0.6253 | 0.1852 | 0.2169 |
| | 80% (r) | 0.9162 | 0.6229 | 0.5033 | 0.6607 | 0.1577 | 0.2202 |
| | 95% (r) | **0.9262** | 0.6472 | 0.5577 | 0.6852 | 0.1326 | 0.2203 |
| BR | Original | 0.7406 | 0.7503 | 0.7488 | 0.7479 | 0.7503 | 0.2258 |
| | 20% (r) | 0.7513 | 0.5538 | 0.4774 | 0.5851 | 0.0799 | 0.2571 |
| | 35% (r) | 0.8123 | 0.5154 | 0.4446 | 0.6010 | 0.1079 | 0.2647 |
| | 50% (r) | 0.8536 | 0.5043 | 0.4265 | 0.6046 | 0.1143 | 0.2764 |
| | 65% (r) | 0.8837 | 0.5161 | 0.4489 | 0.6391 | 0.1187 | 0.2781 |
| | 80% (r) | 0.9008 | 0.5371 | 0.4761 | 0.6746 | 0.1241 | 0.2796 |
| | 95% (r) | **0.9124** | 0.5415 | 0.4940 | 0.6968 | 0.1255 | 0.2776 |
| DBR | Original | 0.7421 | 0.7497 | 0.7485 | 0.7412 | 0.7503 | 0.2286 |
| | 20% (r) | 0.7585 | 0.5573 | 0.4822 | 0.5855 | 0.0799 | 0.2705 |
| | 35% (r) | 0.8161 | 0.5341 | 0.4553 | 0.6137 | 0.1079 | 0.2737 |
| | 50% (r) | 0.8528 | 0.5390 | 0.4680 | 0.6220 | 0.1143 | 0.2719 |
| | 65% (r) | 0.8850 | 0.5501 | 0.4957 | 0.6506 | 0.1187 | 0.2791 |
| | 80% (r) | 0.9011 | 0.5710 | 0.5260 | 0.6884 | 0.1241 | 0.2761 |
| | 95% (r) | **0.9133** | 0.5808 | 0.5411 | 0.7036 | 0.1255 | 0.2816 |
| CC | Original | 0.7377 | 0.7501 | 0.7475 | 0.7404 | 0.7503 | 0.2418 |
| | 20% (r) | 0.7440 | 0.5550 | 0.4357 | 0.5390 | 0.0799 | 0.2640 |
| | 35% (r) | 0.8028 | 0.5164 | 0.4212 | 0.5644 | 0.1079 | 0.2668 |
| | 50% (r) | 0.8346 | 0.4968 | 0.4161 | 0.5720 | 0.1143 | 0.2664 |
| | 65% (r) | 0.8684 | 0.5111 | 0.4412 | 0.6086 | 0.1187 | 0.2738 |
| | 80% (r) | 0.8867 | 0.5371 | 0.4668 | 0.6440 | 0.1241 | 0.2825 |
| | 95% (r) | **0.9019** | 0.5540 | 0.4860 | 0.6672 | 0.1255 | 0.2761 |
| NS | Original | 0.7390 | 0.7503 | 0.7497 | 0.7464 | 0.7503 | 0.2911 |
| | 20% (r) | 0.7431 | 0.5717 | 0.4787 | 0.5749 | 0.0799 | 0.2566 |
| | 35% (r) | 0.7971 | 0.5238 | 0.4462 | 0.5904 | 0.1079 | 0.2391 |
| | 50% (r) | 0.8352 | 0.5033 | 0.4433 | 0.5928 | 0.1143 | 0.2360 |
| | 65% (r) | 0.8667 | 0.5214 | 0.4600 | 0.6233 | 0.1187 | 0.2314 |
| | 80% (r) | 0.8927 | 0.5373 | 0.4947 | 0.6590 | 0.1241 | 0.2292 |
| | 95% (r) | **0.9016** | 0.5451 | 0.5133 | 0.6833 | 0.1255 | 0.2271 |

**Table 5**
F1-scores for DNN and each balancing rate (r) of SMOTE for both the datasets, i.e., $Set_1$ and $Set_2$.

| Method | Dataset | Feature set | | |
|---|---|---|---|---|
| | | $Set_1$ | $Set_2$ | Average |
| DNN | 20% (r) | | 0.897 | 0.906 |
| | 35% (r) | 0.947 | 0.926 | 0.936 |
| | 50% (r) | 0.958 | 0.938 | 0.948 |
| | 65% (r) | 0.962 | 0.945 | 0.953 |
| | 80% (r) | 0.956 | 0.952 | 0.954 |
| | 95% (r) | **0.972** | **0.955** | **0.963** |

to 7 classes) and DNN adapts its structure. BR, DBR, and CC increased the number of classifiers considering the number of labels. In this study, these three methods generated five different models. Finally, NS can create several layers of models to take advantage of predictions made by the previous layers. In our experiments, we used two layers, reaching a total of 10 models, where five are in the first layer and five are in the second one. Using the count of trained classification models, we can state that LP and DNN were the most concise.

Our results revealed the DNN as the most predictive method demanding a single model to tackle the classification problem. It is important to mention the DNN tuning poses an additional effort towards adapting the architecture and hyperparameters when fitting the model. However, the authors of paper de Carvalho and Freitas (2009) highlighted that this strategy may present a better performance in difficult real-world problems than the problem transformation methods.

## 5. Discussion

Principal components analysis (PCA) is a well-known method that produces a sequence of best linear approximations to a provided feature vector. This method can provide patterns in data to exploit their similarities and differences. Thus, four PCAs were calculated to support a general overview of the features sets $Set_1$ and $Set_2$. Two scenarios were built over all classes, as shown in Figs. 5(a) and 5(b). Accordingly, other two PCAs were computed to expose the behaviour of single and multiple diseases focusing on DYS, LAR, RDE, DYS-LAR, and LAR-RDE patterns, as in Figs. 5(c) and 5(d).

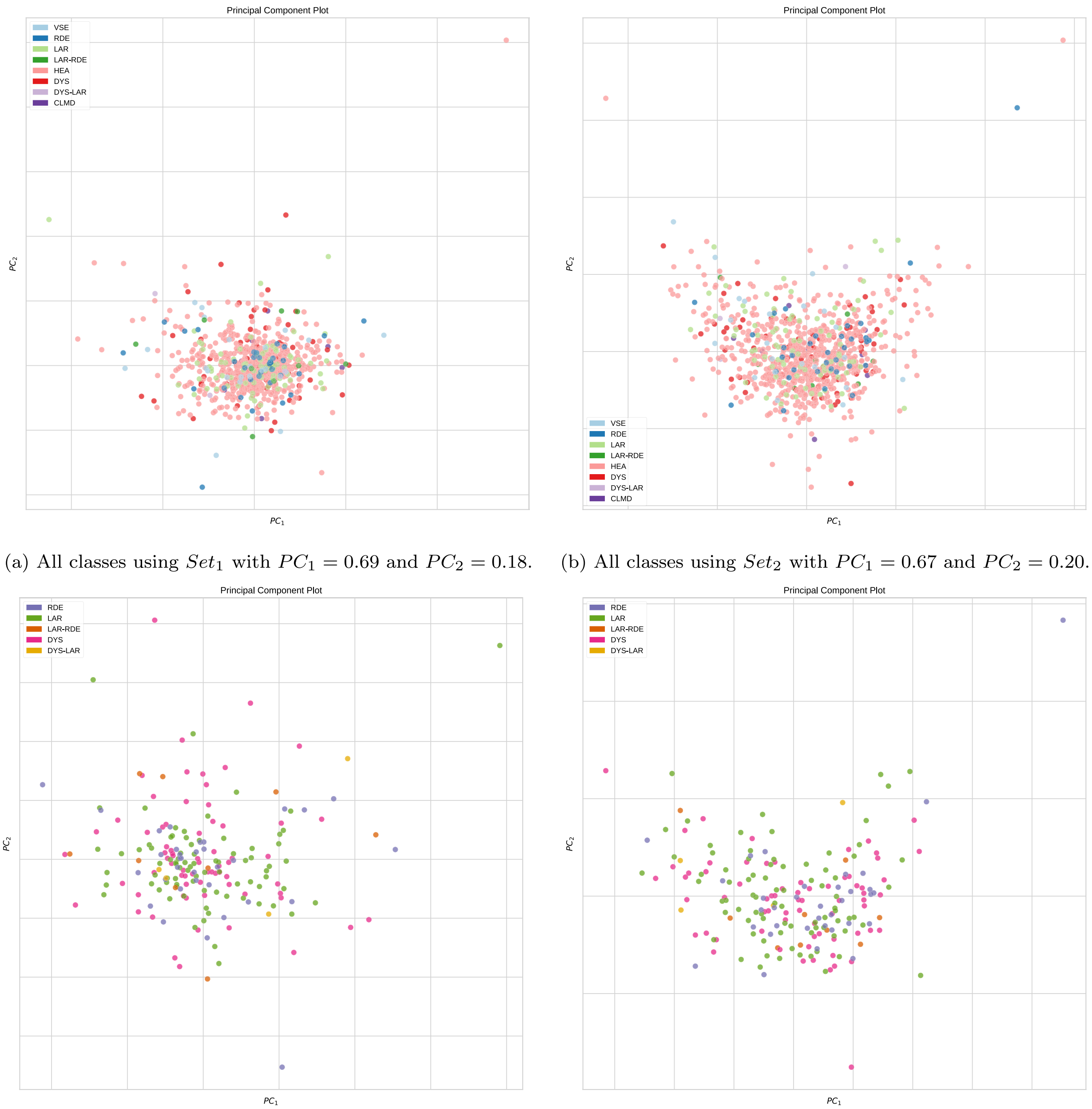
Using two different PCAs from all classes, it was possible to explain 87% of variance for both feature sets. Observing the projection, it was not possible to note dense regions or particular patterns, since all classes presented spread samples over all projected space. This behaviour leads us to employ the non-linear modelling provided by the selected machine learning algorithms.

Analysing the selected classes grounded on multiple pathologies, as in Figs. 5(c) and 5(d), we can state that there is no dense or particular region of sample concentration, regardless of feature sets. In addition, it is possible to highlight that the corresponding distribution corroborates the usage of augmentation techniques, since it was possible to keep the main characteristics of initial dataset without creating distribution of minority classes over dense spots. Different from the computation of all classes, these scenarios allow for a high variance, superior to 95%.

As observed in papers Orozco-Arroyave et al. (2015), Gómez-García et al. (2019), Amami and Smiti (2017), Ankıshan (2019), Arji et al. (2019), Cummins et al. (2018) and those cited in Section 2, ML algorithms have been extensively adopted to solve traditional problems in

**Table 6**
F1-scores of every MLC problem transformation method using RF as base learner and DNN for each label and for both the sets of features.

| Label | LP | | BR | | DBR | | CC | | NS | | DNN | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $Set_1$ | $Set_2$ | $Set_1$ | $Set_2$ | $Set_1$ | $Set_2$ | $Set_1$ | $Set_2$ | $Set_1$ | $Set_2$ | $Set_1$ | $Set_2$ |
| HEA | 0.828 | 0.763 | 0.794 | 0.730 | 0.802 | 0.732 | 0.798 | 0.711 | 0.797 | 0.713 | 0.858 | 0.818 |
| CLMD | 0.962 | 0.801 | 0.950 | 0.778 | 0.943 | 0.763 | 0.945 | 0.705 | 0.949 | 0.666 | 0.982 | 0.971 |
| DYS | 0.810 | 0.766 | 0.776 | 0.735 | 0.785 | 0.742 | 0.754 | 0.708 | 0.737 | 0.700 | 0.905 | 0.856 |
| LAR | 0.810 | 0.784 | 0.778 | 0.712 | 0.778 | 0.732 | 0.789 | 0.710 | 0.793 | 0.722 | 0.893 | 0.861 |
| RDE | 0.868 | 0.760 | 0.820 | 0.717 | 0.825 | 0.740 | 0.829 | 0.706 | 0.830 | 0.711 | 0.936 | 0.927 |
| VSE | 0.857 | 0.798 | 0.861 | 0.786 | 0.872 | 0.779 | 0.853 | 0.779 | 0.866 | 0.751 | **0.919** | **0.934** |
| Avg | 0.856 | 0.779 | 0.830 | 0.743 | 0.834 | 0.748 | 0.828 | 0.720 | 0.829 | 0.711 | 0.916 | 0.897 |



(a) All classes using $Set_1$ with $PC_1 = 0.69$ and $PC_2 = 0.18$.

(b) All classes using $Set_2$ with $PC_1 = 0.67$ and $PC_2 = 0.20$.

(c) Selected classes from $Set_1$ with $PC_1 = 0.84$ and $PC_2 = 0.11$.

(d) Selected classes from $Set_2$ with $PC_1 = 0.85$ and $PC_2 = 0.11$.

**Fig. 5.** Bidimensional Principal Component spaces obtained from four different scenarios using the original dataset: (a) all classes with $Set_1$, (b) all classes with $Set_2$, (c) selected pathologies with multiple cardinalities (DYS, LAR, RDE, DYS-LAR, and LAR-RDE) from $Set_1$ and (d) selected pathologies with multiple cardinalities (DYS, LAR, RDE, DYS-LAR, and LAR-RDE) from $Set_2$.

medicine, where, particularly for voice disorders sorting, the existing classification strategies are commonly used to identify the subject's condition non-invasively. From all those papers, we can learn that promising results depend on the discriminative capacity of the selected features. Thus, many features have been proposed and intensively experimented to describe temporal, spectral or time–frequency characteristics from voice data.

In this paper, we employed the features suggested in previous pieces of research, as documented in papers Al-Nasheri et al. (2018), Shilaskar et al. (2017), Hegde et al. (2019) and Guido (2016a). Consequently, our feature vectors were designed to achieve suitable results, where some detection scenarios, such as IoT, m-Health, and big data environments demand either a reduced usage of resources or the processing of a massive amount of voice data. For those cases, features as proposed in paper Ankıshan (2019) work similarly to those we adopted: based on Fibonacci space representation, they reduce data requirements and produce meaningful information for the classification tasks.

By exploring speech pathology as a binary problem (Muhammad et al., 2012b; Vikram and Umarani, 2013; Akbari and Arjmandi, 2014; Hemmerling et al., 2016; Martinez et al., 2012; Mekyska et al., 2015; Ali et al., 2016; Markaki and Stylianou, 2011; Sasou, 2017; Verde et al., 2018b; Lachhab et al., 2014; Zhong et al., 2016; Amami and Smiti, 2017; Ankıshan, 2019; Cummins et al., 2018), as a multi-class strategy (Gómez-García et al., 2019) or both (Arji et al., 2019), current literature focuses on bringing insights involving signal quality and accuracy improvement. Accordingly, the straightforward adoption of a machine learning model trained with suitable features capable of describing noisy information, as in this paper and in papers Orozco-Arroyave et al. (2015) and Gómez-García et al. (2019), avoids the overhead of heavy algorithms and their hyperparameters tuning. Particularly, the existence of multiple disorders had been neglected until the proposal found in paper Orozco-Arroyave et al. (2015), where the authors study three different sources for pathologically-affected voices: *laryngeal*, *functional*, and *neurological*, reporting accuracies within the range 81% ∼ 98% depending on the pronounced vowel and idiom.

SVM, which is one of the classifiers we tested, was also used in papers Al-Naheri et al. (2017), Muhammad and Melhem (2014), Saeedi and Almasganj (2013), Pranav and Sabarimalai (2017), Fonseca and Pereira (2008), and particularly in paper Amami and Smiti (2017) to handle noise features when discriminating normal from pathologically-affected voices. The authors of that paper reported important achievements, however, restricted to a binary classification problem assessed over a modest dataset. Differently, the combination of LP and RF assessed over SVD dataset for multi-label classification, which provided the best accuracies according to our experiments, were not reported in previous pieces of work. Complementarily, it is very important to highlight that the combination of SE, ZCRs, and SH, obtained based on methods $A_3$, $B_3$, and $C_2$ which were defined in papers Guido (2016a), Guido (2016b), and Guido (2018), respectively, with the base-learners and classifiers tested in this paper **had never been reported in the literature**. Thus, to our best knowledge, our results provide an original contribution, advancing the state-of-the-art.

Finally, it is important to note we focused on evaluating a vast type of MLC algorithms from problem-transformation and algorithm adaptation. DNN superiority was obtained considering the usage of synthetic samples to balance the training set and handcrafted features. This result was important to validate our proposal of handcrafted features and find the best combination of descriptors. However, it is relevant to discuss the usage of handcrafted features in a deep learning solution. The abstraction capacity provided by DNN when handling signals directly is well-known, requiring massive data resources to fit its models. We applied DNN on handcrafted features considering the scarcity of a real-life dataset and examples to support studies regarding pathologically-affected voices, even more when addressing multiple voice disorders in one individual. In addition, our most interesting finding of most false negative for laryngitis may have more deeper relation with speaker recognition tasks (Doddington et al., 2000), remaining an open research problem. Thus, DNN capacity to process raw data directly was not employed in this work to match our dataset size and to provide a fair comparison among all MLC methods and also, besides to study the handcrafted features.

## 6. Conclusions

In this paper, we investigated handcrafted features and ML methods for assessing multiple incidences of voice disorders in the same subject. More precisely, multi-label classification methods were successfully employed to identify subjects with healthy or pathologically-affected voices, i.e., Laryngitis, Dysphonia, Reinke Edema, Vox Senilis, Central Laryngeal Motion Disorder, both Laryngitis and Dysphonia, or both Laryngitis and Reinke Edema. The results have showed that all MLC methods were statistically superior to `Random` and `Majority`. The most complex prediction was related to the disorders that occur at the same time, however, all the disorders have superior predictive performance when compared to healthy subjects.

Particularly, the DNN-based approach presented the best values of F1-score among the tested methods, i.e., 0.943 as the averaged value from all balancing scenarios, justifying label dependencies. Further comparisons revealed that the critical level of energy $C = 1\%$ used to compute feature vectors composed by SE, ZCR, and SH is the best option, i.e., a more refined analysis is relevant. Last but not least, recent feature learning strategies, such as those based on auto-encoders or deep learning, addressing the raw samples directly, were not considered in this paper because one of our intentions was the possibility to interpret the features, allowing for a clearer understanding of the problem.

Notably, one of the limitations readers might find in this paper is related to the cross-validation procedures applied to a high-balanced rate dataset. Indeed, the 13 multi-labelled samples were hundred times oversampled, causing a low variance in the dataset and, thus, degrading the statistical significance of the accuracies we reported. Nevertheless, as explained above, we adopted the only existing dataset to perform the experiments. In addition, we believe that SMOTE, as we adopted, allows for F1-score to be sufficiently convincing in view of the restricted original data.

As a future work, we suggest applying MLC to a database that presents the co-occurrence of additional voice pathologies, especially the complex ones. Lastly, we would like to emphasize that we are not proposing a single method, but rather exploring a range of multi-label techniques from different families; we are leveraging handcrafted features to provide a robust model. After an extensive literature review, we were unable to find other particular methods that specifically address our scenario and could be used as an additional competitor in our experimental study. Thus, additional comparisons would not be meaningful and would distort the main objective of our paper. By exploring a range of multi-label techniques, we are sure we have already provided a comprehensive assessment of their effectiveness for the detection of multiple voice disorders: this is our main goal and, thus, we believe that our current results will enable us to identify the most promising techniques for future research and clinical applications.

## Acronyms

**BR** Binary Relevance.

**CC** Classifier Chains.

**CLMD** Central Laryngeal Motion Disorder.

**DBR** Dependent Binary Relevance.

**DNN** Deep Neural Network.

**DYS** Dysphonia.

**LAR** Laryngitis.

**LP** Label Powerset.

**MLC** Multi-label Classification.

**NS** Nested Stacking.

**RDE** Reinke Edema.

**RF** Random Forest.

**SLC** Single-label Classification.

**SVM** Support Vector Machine.

**VSE** Vox Senilis.

## Declaration of competing interest

ALL the authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Funding

- “Conselho Nacional de Desenvolvimento Científico e Tecnológico” (CNPq) in Brazil under grant of projects 420562/2018-4, 310668/2019-0 and 303854/2022-7;
- “Fundação de Amparo à Pesquisa do Estado de São Paulo” (Fapesp) in Brazil under the grant of project 2021/12407-4;
- “Conselho de Aperfeiçoamento de Pessoal do Ensino Superior” (CAPES) in Brazil under finance code 001.

## References

Akbari, A., Arjmandi, M.K., 2014. An efficient voice pathology classification scheme based on applying multi-layer linear discriminant analysis to wavelet packet-based features. Biomed. Signal Process. Control 10, 209–223.

Al-Naheri, A., et al., 2017. Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions. IEEE Access 6, 6969–6974.

Al-Nasheri, A., et al., 2018. Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions. IEEE Access 6, 6961–6974.

Ali, Z., et al., 2016. Automatic voice pathology detection with running speech by using estimation of auditory spectrum and cepstral coefficients based on the all-pole model. J. Voice 30 (6), 757.e7–757.e19.

Almeida, A.M., et al., 2018. Applying multi-label techniques in emotion identification of short texts. Neurocomputing 320, 35–46.

AlRshoud, M., et al., 2019. Implementation of voice pathology detection system using feature selection. Comput. Methods Programs Biomed. 171, 9.

Amami, R., Smiti, A., 2017. An incremental method combining density clustering and support vector machines for voice pathology detection. Comput. Electr. Eng. 57, 257–265.

Ankıshan, H., 2019. Classification of acoustic signals with new feature: Fibonacci space (FSp). Biomed. Signal Process. Control 48, 221–233.

Areiza-Laverde, H.J., Castro-Ospina, A.E., Peluffo-Ordonez, D.H., 2018. Voice pathology detection using artificial neural networks and support vector machines powered by a multicriteria optimization algorithm. In: International Workshop on Experimental and Efficient Algorithms, L’Aquila, Italy. pp. 148–159.

Arji, G., et al., 2019. A systematic literature review and classification of knowledge discovery in traditional medicine. Comput. Methods Programs Biomed. 168, 39–57.

Babatsouli, E., et al., 2016. Entropy as a measure of mixedupness of realizations in child speech. Poznan Stud. Contemp. Linguistics 4 (52), 605–627.

Barry, W., Putzer, M., 2007. Saarbrücken Voice Database. Institute of Phonetics, Universitat des Saarlandes, http://www.stimmdatenbank.coli.uni-saarland.de.

Batista, Gustavo E.A.P.A., Prati, Ronaldo C., Monard, Maria Carolina, 2004. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor. Newsl. 6 (1), 20–29.

Belhaj, A., Bouzid, A., Ellouze, N., 2015. Edema and nodule pathological voice identification by SVM classifier on speech signal. Comput. Softw. 10 (5), 495.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Casper, J.K., Leonard, R., 2011. Understanding Voice Problems: A Physiological Perspective for Diagnosis and Treatment, fourth ed. Lippincott Williams & Wilkins Press, Baltimore, USA.

Chawla, n.V., et al., 2002. Smote: Synthetic minority over-sampling technique. J. Artificial Intelligence Res. 16, 321–357.

Chollet, F., 2018. Keras: The python deep learning library. ascl. Jun:ascl-1806.

Cover, T.M., 1965. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Trans. Electron. Comput. (3), 326–334.

Crammer, K., Singer, Y., 2003. A family of additive online algorithms for category ranking. J. Mach. Learn. Res. 3, 1025–1058.

Cummins, N., Baird, A., Schuller, B.W., 2018. Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. Methods 151, 41–54.

David, M.A., 2018. Diadochokinesis-based expert system considering articulatory features of plosive consonants for early detection of Parkinson’s disease. Comput. Methods Programs Biomed. 154, 89–97.

de Carvalho, A.C., Freitas, A.A., 2009. A tutorial on multi-label classification techniques. Found. Comput. Intell. (5), 177–195.

Doddington, G.R., et al., 2000. The NIST speaker recognition evaluation: Overview, methodology, systems, results, perspective. Speech Commun. (31), 225–254.

Fonseca, E.S., Pereira, J.C., 2008. Normal versus pathological voice signals: Using wavelet analysis and support vector machines. IEEE Eng. Med. Biol. Mag. 28 (5), 44–48.

Georgoulas, G., et al., 2007. Novel approach for fetal heart rate classification introducing grammatical evolution. Biomed. Signal Process. Control 2 (2), 69–79.

Ghasem, S.B., et al., 2019. Diagnosis of autism spectrum disorder based on complex network features. Comput. Methods Programs Biomed. 177, 277–283.

Godbole, S., Sarawagi, S., 2004. Discriminative methods for multi-labeled classification. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Sidney, Australia. pp. 22–30.

Gómez-García, J.A., Moro-Velázquez, L., Godino-Llorente, J.I., 2019. 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, 181–199.

Guido, R.C., 2016a. A tutorial on signal energy and its applications. Neurocomputing 179, 264–282.

Guido, R.C., 2016b. ZCR-aided neurocomputing: A study with applications. Knowl.-Based Syst. 105, 248–269.

Guido, R.C., 2018. A tutorial-review on entropy-based handcrafted feature extraction for information fusion. Inf. Fusion 41, 161–175.

Haixiang, G., et al., 2017. Learning from class-imbalanced data: Review of methods and applications. Expert Syst. Appl. 73, 220–239.

Hegde, S., et al., 2019. A survey on machine learning approaches for automatic detection of voice disorders. J. Voice 33 (6), 947.e11–947.e33.

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

Ji, M., et al., 2020. Multi-label learning for crop leaf diseases recognition and severity estimation based on convolutional neural networks. Soft Comput. 24, 15327–15340.

Krawczyk, B., Schaefer, G., Wozniak, M., 2015. A hybrid cost-sensitive ensemble for imbalanced breast thermogram classification. Artif. Intell. Med. 65 (3), 219–227.

Lachhab, O., et al., 2014. Improving the recognition of pathological voice using the discriminant HLDA transformation. In: 3rd IEEE International Coloquium in Information Science and Technology. CIST, pp. 370–373.

Lee, B.J., et al., 2013. Prediction of body mass index status from voice signals based on machine learning for automated medical applications. Artif. Intell. Med. 58 (1), 51–61.

Lenc, L., Kral, P., 2016. Deep neural networks for czech multi-label document classification. In: International Conference on Intelligent Text Processing and Computational Linguistics. Springer, Cham, pp. 460–471.

Licklider, J.C.R., 1948. The influence of interaural phase relations upon the masking of speech by white noise. J. Acoust. Soc. Am. 20 (150), 150–159.

Lin, W.-Z., et al., 2013. Iloc-animal: A multi-label learning classifier for predicting subcellular localization of animal proteins. Mol. Biosyst. 9 (4), 634–644.

Liu, S.M., Chen, J.-H., 2015. A multi-label classification based approach for sentiment classification. Expert Syst. Appl. 42 (3), 1083–1093.

Liu, J., et al., 2017. Deep learning for extreme multi-label text classification. In: Proc. of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Tokyo, Japan. pp. 115–124.

Lorenzo, M., Claudia, M., 2002. Software corrections of vocal disorders. Comput. Methods Programs Biomed. 68 (2), 135–145.

Markaki, M., Stylianou, Y., 2011. Voice pathology detection and discrimination based on modulation spectral features. IEEE Trans. Audio, Speech, Lang. Process. 19 (7), 1938–1948.

Martinez, D., et al., 2012. Voice pathology detection on the Saarbrücken voice database with calibration and fusion of scores using multifocal toolkit. In: Advances in Speech and Language Technologies for Iberian Languages. Springer, Berlin, Germany, pp. 99–109.

Mastelini, S.M., et al., 2019. Multi-output tree chaining: An interpretative modelling and lightweight multi-target approach. J. Signal Process. Syst. 91 (2), 191–215.

Mekyska, J., et al., 2015. Robust and complex approach of pathological speech signal analysis. Neurocomputing 167, 94–111.

Misra, H., 2004. Spectral entropy based feature for robust ASR. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, QB, Canada. p. 193.

Montanes, E., et al., 2014. Dependent binary relevance models for multi-label classification. Pattern Recognit. 47 (3), 1494–1508.

Muhammad, G., Melhem, M., 2014. Pathological voice detection and binary classification using MPEG-7 audio features. Biomed. Signal Process. Control 11, 1–9.

Muhammad, G., et al., 2012a. Multidirectional regression (MDR)-based features for automatic voice disorder detection. J. Voice 26 (6), 817.e19–817.e27.

Muhammad, G., et al., 2012b. Multidirectional regression (MDR)-based features for automatic voice disorder detection. J. Voice 26 (6), 817e19–27.

Orozco-Arroyave, J.R., et al., 2015. Characterization methods for the detection of multiple voice disorders: Neurological, functional, and laryngeal diseases. IEEE J. Biomed. Health Inf. 19 (6), 1820–1828.

Pereira, R.B., et al., 2018. Correlation analysis of performance measures for multi-label classification. Inf. Process. Manage. 54 (3), 359–369.

Potharaju, S.P., Sreedevi, M., 2016. An improved prediction of kidney disease using smote. Indian J. Sci. Technol. 9 (31), 1–7.

Pranav, S.D., Sabarimalai, M., 2017. Effective glottal instant detection and electroglottographic parameter extraction for automated voice pathology assessment. IEEE J. Biomed. Health Inf. 22 (2), 398–408.

Quatieri, T.F., 2008. Discrete-Time Speech Signal Processing: Principles and Practice. Pearson.

Rallapalli, V.H., Alexander, J.M., 2015. Neural-scaled entropy predicts the effects of nonlinear frequency compression on speech perception. J. Acoust. Soc. Am. 138 (5), 3061–3072.

Read, J., et al., 2011. Classifier chains for multi-label classification. Mach. Learn. 85 (3), 333.

Rivolli, A., Carvalho, A.C., 2018. The utiml package: Multi-label classification in R. R J. 10 (2), 24–37.

Saarela, M., Ryynanen, O.P., Ayramo, S., 2019. Predicting hospital associated disability from imbalanced data using supervised learning. Artif. Intell. Med. 95, 88–95.

Saeedi, N.E., Almasganj, F., 2013. Wavelet adaptation for automatic voice disorder sorting. Comput. Biol. Med. 43 (6), 699–704.

Salehi, P., 2015. Using patient's speech signal for vocal ford disorders detection based on lifting scheme. In: IEEE 2nd International Conference on Knowledge-Based Engineering and Innovation. KBEI, Tehran, Iran, pp. 561–568.

Sasou, A., 2017. Automatic identification of pathological voice quality based on the GRBAS categorization. In: Asia-Pacific and Information Processing Association Annual Summit and Conference. APSIPA ASC, Malaysia, pp. 1243–1247.

Schroeder, M.R., 1966. Vocoders: Analysis and synthesis of speech. Proc. IEEE 54 (5), 720–734.

Senge, R., et al., 2013. Rectifying classifier chains for multi-label classification. In: Proceedings Workshop LWA, Lernen-Wissensentdeckung-Adaptivitat, Bamberg, Germany. pp. 151–158.

Shilaskar, S., Ghatol, A., Chatur, P., 2017. Medical decision support system for extremely imbalanced datasets. Inform. Sci. 384, 205–219.

Sorower, M.S., 2010. A Literature Survey on Algorithms for Multi-Label Learning, Vol. 18, no. 1. Oregon State University, Corvallis, p. 25.

Techakesari, O., Ford, J.J., 2013. Relative entropy rate based model selection for linear hybrid system filters of uncertain nonlinear systems. Signal Process. 93, 12–22.

Tsoumakas, G., Katakis, I., 2007. Multi-label classification: An overview. Int. J. Data Warehous. Min. (IJDWM) 3 (3), 1–13.

Tsoumakas, G., Katakis, I., Vlahavas, I., 2009. Mining multi-label data. In: Data Mining and Knowledge Discovery Handbook. Springer, New York, USA, pp. 667–685.

Tsoumakas, G., Vlahavas, I., 2007. Random k-labelsets: An ensemble method for multilabel classification. In: European Conference on Machine Learning. Springer, Warsaw, Poland, pp. 406–417.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York.

Verde, L., De Pietro, G., Sannino, G., 2018a. Voice disorder identification by using machine learning techniques. IEEE Access 6, 16246–16255.

Verde, L., Pietro, G., Sannino, G., 2018b. A methodology for voice classification based on the personalized fundamental frequency estimation. Biomed. Signal Process. Control 42, 134–144.

Vikram, C.M., Umarani, K., 2013. Phoneme independent pathological voice detection using wavelet bases, MFCCs and GMM-SVM hybrid classifier. In: International Conference on Advances in Computing, Communications and Informatics. ICACCI, Chengdu, China, pp. 153–156.

Vinay, N.A., Bharathi, S.H., 2019. Dysfluency recognition by using spectral entropy features. Int. J. Eng. Adv. Technol. (IJEAT) 6 (8), 517–520.

Wang, S., Bi, S., Zhang, Y.J., 2020. Locational detection of false data injection attack in smart grid: A multi-label classification approach. IEEE Internet Things J. 7 (9), 8218–8227.

Wosiak, A., Glinka, K., Zakrzewska, D., 2018. Multi-label classification methods for improving comorbidities identification. Comput. Biol. Med. 100, 279–288.

Xia, M., Xu, Z., 2012. Entropy/cross entropy-based group decision making under intuitionistic fuzzy environment. Inform. Fusion 13 (1), 31–47.

Zarinbal, M., Zarandia, M.H.F., Turksen, I.B., 2015. Relative entropy collaborative fuzzy clustering method. Pattern Recognit. 48 (3), 933–940.

Zhang, X., Mei, C., Chen, D., Li, J., 2016. Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy. Pattern Recognit. 56, 1–15.

Zhang, M.L., Zhou, Z.H., 2006. Multi-label neural networks with applications to functional genomics and text categorization. IEEE Trans. Knowl. Data Eng. 18 (10), 1338–1351.

Zhau, G., Hansen, J.H.L., Kaiser, J.F., 2001. Non-linear feature based classification of speech under stress. IEEE Trans. Speech Audio Process. (9), 201–216.

Zhong, Z., et al., 2016. Nonlinear signal processing for vocal folds damage detection based on heterogeneous sensor network. Signal Process. 126, 125–133.

Zufferey, D., et al., 2015. Performance comparison of multi-label learning algorithms on clinical data for chronic diseases. Comput. Biol. Med. 65, 34–43.