BIOCYBERNETICS AND BIOMEDICAL ENGINEERING 41 (2021) 1039–1056

Available at www.sciencedirect.com

ScienceDirect

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

Original Research Article

Impact of noise on the performance of automatic systems for vocal fold lesions detection

Mario Madruga [a,b,*], *Yolanda Campos-Roca* [c], *Carlos J. Pérez* [a]

[a] *Departamento de Matemáticas, Universidad de Extremadura, Spain*
[b] *Facultad de Veterinaria, Avenida de la Universidad S/N, Cáceres, Spain*
[c] *Departamento de Tecnología de los Computadores y las Comunicaciones, Universidad de Extremadura, Spain*

ARTICLE INFO

Article history:
Received 22 December 2020
Received in revised form
25 June 2021
Accepted 4 July 2021
Available online 16 July 2021

Keywords:
Acoustic features
Computer aided diagnosis
Reinke's edema
Nodules
Noise robustness
Voice disorders

ABSTRACT

Automatic voice condition analysis systems have been developed to automatically discriminate pathological voices from healthy ones in the context of two disorders related to exudative lesions of Reinke's space: nodules and Reinke's edema. The systems are based on acoustic features, extracted from sustained vowel recordings. Reduced subsets of features have been obtained from a larger set by a feature selection algorithm based on Whale Optimization in combination with Support Vector Machine classification. Robustness of the proposed systems is assessed by adding noise of two different types (synthetic white noise and actual noise recorded in a clinical environment) to corrupt the speech signals. Two speech databases were used for this investigation: the Massachusetts Eye and Ear Infirmary (MEEI) database and a second one specifically collected in Hospital San Pedro de Alcántara (Cáceres, Spain) for the scope of this work (UEX-Voice database). The results show that the prediction performance of the detection systems appreciably decrease when moving from MEEI to a database recorded in more realistic conditions. For both pathologies, the prediction performance declines under noisy conditions, being the effect of white noise more pronounced than the effect of noise recorded in the clinical environment.

1. Introduction

Voice is a person's main communication tool and, therefore, the impact of voice disorders on quality of life can be substantial. For people involved in certain professions, such as teachers, singers, and many others, voice is also the main working tool and, as an immediate consequence, they are in a high risk of developing voice disorders due to excessive and/or incorrect use of their voices. Voice professionals are prone to suffer from organic diseases and will eventually need some kind of medical diagnosis and care [1]. Some of those voice disorders are exudative lesions of Reinke's space and are manifestations of different etiologic factors like voice abuse leading to nodules, or tobacco use linked to Reinke's Edema [2].

* Corresponding author at: Departamento de Matemáticas, Universidad de Extremadura, Spain.
E-mail addresses: mariome@unex.es (M. Madruga), ycampos@unex.es (Y. Campos-Roca), carper@unex.es (C.J. Pérez).
https://doi.org/10.1016/j.bbe.2021.07.001

The main methods used by otolaryngologists to diagnose laryngeal diseases are direct inspection of the larynx through the use of invasive techniques such as laryngoscopy and videostroboscopy [3], and/or evaluation of voice quality by hearing. The first group of diagnosis techniques causes discomfort to the patient and requires sophisticated equipment like endoscopic instruments or specialized video cameras, whereas the second group is subjective and strongly depends on the experience of the specialist [4].

In recent years, computer aided diagnosis (CAD) of voice disorders has attracted considerable scientific interest with the aim of providing an effective screening method for pathologies in an early stage. Using automatic voice condition analysis (AVCA) helps the physicians providing useful information in the differential diagnostic process [5]. These techniques usually consist of an acoustic feature extraction step followed by the application of machine learning algorithms under the assumption that voice quality is correlated with voice pathology [5]. Compared to the previously mentioned diagnosis methods, these techniques show the advantages that they are non-invasive, fast, objective, and low-cost. Also, acoustic analysis have been proved to be a sensitive, objective, and quantitative tool, being more accurate than perceptual assessment [6]. For example, they can be applied in preventive medicine to professionals at high risk of suffering from voice disorders [7]. Other contribution in the field of automatic detection of structural vocal-fold pathologies is [8], which offers experimental results of binary discrimination between normal and pathological voices, where the pathological voice class is composed of a variety of disorders. A recent scientific review on AVCA systems is provided by [9].

This paper focuses on laryngeal diseases, in particular, nodules and Reinke's edema. In this application context, the most usual vocal task is sustained phonation of /a/ vowels, where the speakers are asked to pronounce a vowel sound as steady as possible in terms of amplitude and fundamental frequency [10]. This vocal task has several advantages. First, it requires continuous motion of the vocal folds, which constitute the main structure involved in these pathologies. Also, this vocal task is quick and easy to perform and it is a common sound across different languages and accents. Sustained phonation of other vowel sounds or production of sentences have also been used [8].

Based on sustained vowel recordings, the studies in the literature consider many different characteristics of speech including perturbation measurements (such as jitter or shimmer), noise measures (such as harmonic-to-noise ratio (HNR) or glottal-to-noise excitation (GNE) ratio), Mel frequency cepstral coefficients (MFCCs), among others [5]. More recent studies show that nonlinear time series analysis methods may be more appropriate for pathological voices than classical measurements. Those methods, including Lyapunov exponents and correlation dimension, have been applied to classification of disordered voice samples [5].

There is also a variety of pattern recognition techniques based on supervised learning applied in this context in the scientific literature. Among the many different classification techniques that have been used, [5] highlights Support Vector Machines (SVM) and Gaussian Mixture Models as the most widely employed, although [9] compiles a much wider range of alternatives which have been used in this particular field. In general, when a large feature set size is used, the model becomes less comprehensible and there is a high risk of overfitting [11]. Therefore, for a reliable classification, it is important to use a small number of measurements, containing an optimal amount of information. Feature selection for classification is an active research area on its own, whose main objective is to reduce the dimension of the original feature set. Wrapper algorithms based on meta-heuristic optimization techniques allow to obtain a global optimum of the predictive accuracy achieved for a certain classification algorithm by using a simple and easy to implement concept. Among the different meta-heuristic approaches, the Whale Optimization Algorithm (WOA) is a recently proposed approach which mimics the hunting behavior of humpback whales. It was originally created as an optimization algorithm [12] and later adapted as a feature selection operator [13,14].

An important aspect to take into account regarding AVCA systems for speech disorder detection is robustness. When the recordings have been obtained under a controlled acoustic environment, the performance of these systems in real-life conditions remains unknown. A clear example is the Massachusetts Eye and Ear Infirmary (MEEI) database [15], whose recordings were taken in Kay Elemetrics and MEEI Voice and Speech Lab [16], being these conditions very difficult to reproduce in everyday situations. In order to be useful, it is required that these systems remain robust even when the recordings are captured in a non-controlled environment. Experiments have been carried on in order to assess different channels in remote disease monitoring [17,18]. Even mobile healthcare applications have been tested in controlled acoustical environments, like [19], which mentions that experiments are carried out in an as low as 30 dB background noise room. However, noise robustness is very seldom present in the scientific literature about automatic detection systems of organic voice disorders. [20] presents assumable noise levels of 25 dBA, 36 dBA, 30 dB, 40 dB and 50 dB for different studies, remarking that the maximum acceptable noise level was not investigated. [21] presents a study about the adverse effects of noise on voice quality measurement. This study focuses on fundamental frequency and perturbation measurements with no particular pathology addressed. [22] studies the numerical effects of noise on the computation of different acoustical features, although it does not test their classifying capabilities. [23] performs a preliminary study on the impact of noise on the automatic detection of a particular voice pathology: Reinke's edema. In the context of Parkinson's disease, [24] shows the impact of noise on an automatic detection system based on acoustic features. Finally, [25] proposes a technique to mitigate the possible differences in recording environments, characterized by different noise conditions.

The main goal of the present paper is to assess the negative effects of realistic noisy recording conditions on the outcome of an AVCA system for voice pathology detection. We have focused on two specific related diseases which are common vocal fold lesions, and their etiologies are related. However, we performed independent experiments with each disease in order to minimize the number of variables present in the study since the main goal is not building an automated

diagnostic system, but to check the potential effects of environmental noise on the outcomes of AVCA systems for vocal fold lesions detection.

We have built AVCA systems to discriminate pathological voices from healthy ones in the context of two structural organic speech pathologies: nodules and Reinke's edema. This work is a significant extension of the conference paper [23]. It introduces new case studies in a different pathology (vocal fold nodules) that allow to improve the generalization capability of the conclusions and to make a disease comparison. Also, it exposes a feature selection algorithm which has been designed, implemented, and tested for these applications. Specifically, the systems are built on reduced acoustic feature subsets, obtained by a feature selection algorithm based on Whale Optimization in combination with SVM classification. This algorithm has been implemented using parallel computing libraries and executed on a Beowulf cluster system. Two voice recording databases are employed: The first one is MEEI, recorded in the most favorable acoustic conditions; the second one is an own database, recorded in a more usual clinical environment. Also, system robustness is evaluated by adding two different types of noise (white Gaussian noise and actual clinical environment noise) to both databases and studying the impact on the discrimination capacity of each system.

2. Materials and methods

This section provides the main information on participants, collection and pre-processing of voice samples and noise recording. Also, the proposed feature extraction approach is summarized and the feature selection algorithm is explained.

2.1. *Participants*

MEEI database, commercialized by KayPentax Corp. [15], is one of the voice databases used for this work. This database, widely used for research in pathological voice classification, has been recorded under very strict acoustical and technical conditions (sound-proof booth, high-quality recording equipment, type of microphone, distance to the source...) [16]. It includes sustained /a/ recordings of 53 healthy and 657 pathological subjects, 19 and 25 of them suffering from vocal-fold nodules and Reinke's edemas, respectively.

Not all the voice samples in the MEEI database were recorded using the same technical parameters, being the healthy voices recorded at a sampling rate of 50 kHz, with a total length of 3 s, whereas the pathological voices were recorded at 25 kHz for one second. For the purpose of our experiments, all the waveforms were resampled when needed and trimmed so the whole database complies with the specifications of a sampling rate of 25 kHz and one second length.

An experiment has been conducted to collect a voice recording database (UEX-Voice) also based on sustained /a/ phonations. This database has been recorded in Hospital San Pedro de Alcántara (HSPdA), Cáceres, specifically, in an ordinary diagnostic room, with its door closed, providing only a certain isolation from the noisy aisles and waiting halls surrounding it.

All the recordings were taken using the same equipment: an AKG 520 head-worn condenser cardioid microphone attached to a TASCAM US322 sound card, being the recording software Audacity 2.0.5. The sampling rate was 44.1 kHz. Four phonations were recorded for each participant, of variable lengths depending on the capacity of each individual, so they were trimmed both at the beginning, ensuring no silence, and at the end, to obtain a uniform duration of one second. All the waveforms were downsampled to 25 kHz in order to match the sampling rate of MEEI database.

Fig. 1 shows the age distribution of the considered subjects with nodules and Reinke's edema from the MEEI and UEX-Voice databases. Summary statistics are provided in Table 1.

2.2. *Noise database*

A noise database has been specifically collected from the room where the research study took place. This room was placed in the external consultation area, on the second floor, of a hospital in a small town (population < 100.000). Background noise was recorded using the same equipment previously defined. The length of the recording was 11 min 50 s and included noise from different sources: multitalker babble, cell phone sounds, fluorescent lighting, door closing, and footsteps, among others. Since post-processing is made altering recording level, we are more interested in the nature of sound than in its power. The recordings made include a realistic representation of the variety of indoor noise sources that are present in the outpatient clinic area of any hospital during consultation hours. Furthermore, national and regional environmental noise laws are very strict in hospital surroundings. Anyway, the impact of external sources on the final recordings is negligible, as in free space the received noise power is inversely proportional to the square of the distance to the source, given that external sources are farther away than internal ones, and accounting for the attenuation due to building walls. For those reasons, considering that the voice samples are at most 3 s long, and that they are trimmed down to one second, these noise recordings provided enough variability to perform all the desired experiments.

The noise waveforms were recorded inside the empty diagnosis room with door and windows closed while noise level was being measured using a certified Brüel & Kjaer 2260 sound level meter, what allows us to assert the acoustical environment recreated when using these recordings. Three one-minute measurements showed an A-weighted mean L_{eq} of 34.17 dBA.

2.3. *Feature extraction*

A total of 94 features were extracted from each voice sample. These features have been previously used in scientific literature, either for voice disease detection, Parkinson's disease detection, or other biomedical signal analysis [5,9]. The extraction methods were coded in Python by direct implementation of the formal mathematical definition, by translating existing code from other authors, or by using available libraries of proven reliability from Python repositories. A comprehensible list is provided in Table 2 including short name,

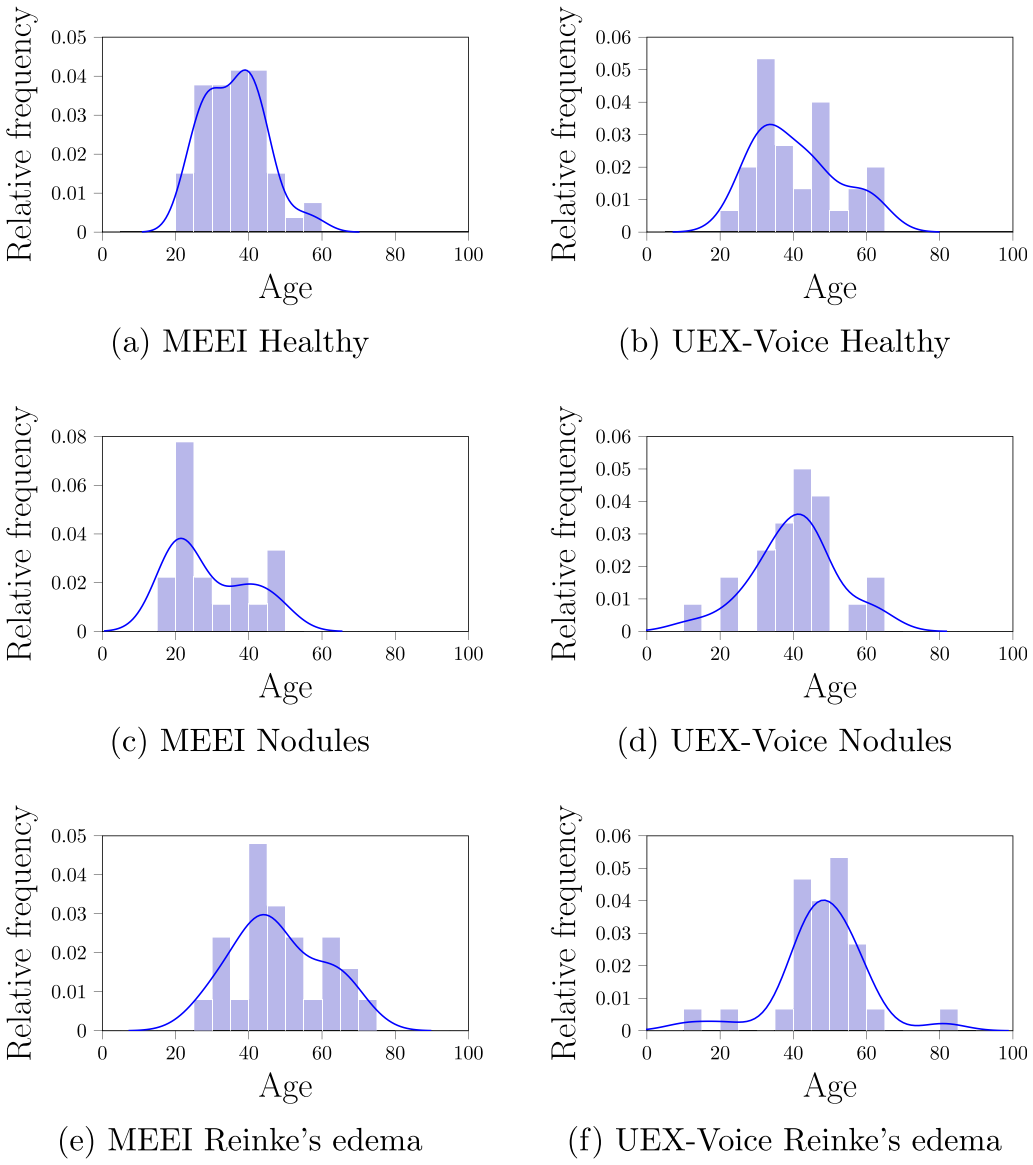


Fig. 1 – Age distribution of the subjects from MEEI and UEX-Voice databases.

Table 1 – Distribution of subjects by health status, sex and age.

		Sex		Age	
Database	Health status	Male	Female	Mean	Std. Dev.
MEEI	Healthy	21	32	36.00	8.29
	Nodules	1	17	29.11	10.45
	Reinke's edema	5	20	48.04	11.97
UEX-Voice	Healthy	4	26	40.76	11.18
	Nodules	1	23	40.41	11.33
	Reinke's edema	3	27	47.96	11.76

references to previous work, and variants taken into consideration.

Age and sex are two features inherent to each subject. Humans undergo several changes with aging that affect the voice production system. For example, changes in the larynx tend to alter the average fundamental frequency and to produce instability of vocal fold vibrations [34]. The impact is different for men than for women; in particular, fundamental frequency tends to increase in men and decrease in women due to some aging effects [35]. Also, women are more prone to suffer from organic voice diseases than men [36]. These and several other aspects related to the impact of age

Table 2 – Features extracted.

Short name	References	Full name and variants
		Linear
CPP	[26]	Cepstral peak prominence
GNE_X	[27]	Glottal-to-noise excitation ratio. Four different statistical features: mean, std, SNR_TKEO, SNR_SEO
GQ	[27]	Glottal quotient. Three statistics used: prc5_95, std cycle open, std cycle closed
HNR	[27]	Harmonic-to-noise ratio
JITTER_X	[27]	Jitter. Twenty-two different statistics used: abs_dif, diff_percent, PQ3_classical_Schoentgen, PQ3_classical_Baken, PQ3_generalised_Schoentgen, PQ5_classical_Schoentgen, PQ5_classical_Baken, PQ5_generalised_Schoentgen, PQ11_classical_Schoentgen, PQ11_classical_Baken, PQ11_generalised_Schoentgen, abs0th_perturb, DB, CV, TKEO_mean, TKEO_std, TKEO_prc5, TKEO_prc25, TKEO_prc75, TKEO_prc95, FM, range_5_95_perc
SHIMMER_X	[27]	Shimmer. Twenty-two different statistics used: abs_dif, diff_percent, PQ3_classical_Schoentgen, PQ3_classical_Baken, PQ3_generalised_Schoentgen, PQ5_classical_Schoentgen, PQ5_classical_Baken, PQ5_generalised_Schoentgen, PQ11_classical_Schoentgen, PQ11_classical_Baken, PQ11_generalised_Schoentgen, abs0th_perturb, DB, CV, TKEO_mean, TKEO_std, TKEO_prc5, TKEO_prc25, TKEO_prc75, TKEO_prc95, FM, range_5_95_perc
MFCC-X	[28]	Mel Frequency Cepstral Coefficient, 13 first coefficients MFCC0 - MFCC12
		Non-linear
D2	[10]	Correlation dimension
FMMI	[29]	First minimum in mutual information
FZCF	[29]	First zero of autocorrelation function
HURST	[10]	Hurst Exponent
MFSW	[30]	Multifractal spectrum width
ZCR	[31]	Zero crossing rate
		Entropies and complexities
PERMUTATION	[32]	Permutation entropy
PPE	[27]	Pitch period entropy
RPDE	[18]	Recurrence Period Density Entropy
SHANNON	[29]	Shannon entropy
LZ-X	[33]	Lempel–Ziv complexity. 16 features quantifying signal 2^1 to 2^{16} steps

and gender on speech have motivated the inclusion of these two features.

Many diseases affecting vocal production cause pitch-related alterations, specifically frequency or amplitude modulation, being sustained vocal analysis the most useful technique to apply [37]. Most studies till recent years focused their attention on acoustical features such as jitter or shimmer, which assume that voice production is a linear system. Though the definition of jitter seems very simple, i.e., the mean variation in the fundamental frequency of the phonation process, there is no method considered as standard for calculating such variation, mainly because the fundamental frequency calculation is not a trivial task. Most usual methods are provided by Multi-Dimensional Voice Program [15], the software tool provided by KayPentax with their database; and Praat suite. Other algorithms have been proposed, such as Sun's algorithm or SWIPE alternatives [38]. In our implementation jitter and shimmer were translated from MATLAB code given by [27]. We obtained 22 different measurements for both jitter and shimmer, each one corresponding to a different mathematical formulation.

Besides jitter and shimmer, other spectrum and fundamental frequency related linear features have been studied. GQ was originally used to monitor Parkinson's disease [39], and shortly after for early diagnosis of pathological voice [40]. GQ takes into account the lengths of time the glottis is open and closed. CPP was proposed as a measure of breathiness and our version was coded following the definition given by [26]. HNR is intended to assess voice hoarseness and tries to estimate the relationship between purely harmonic to turbulent noise in voice production. MFCCs try to describe the spectral components and do not require a previous pitch estimation [28].

Nonlinear behaviors have been shown to play a role in the voice production process and, particularly, in the case of voice pathologies [5]. Therefore, assuming that voice diseases may induce a chaotic behavior in human voice production, nonlinear analysis has also been taken into consideration in the search for new accurate features [7]. RPDE considers the uncertainty in signal cycle estimates using both an embedded space and entropy, being related to fundamental frequency, nonlinear, and entropy measurements [18]. ZCR is not properly a nonlinear measurement, but it is useful in time series analysis [31], measuring the number of times the signal crosses zero level. D2 is an estimator of the correlation dimension, a measure of self-similarity of chaotic systems [10]. HURST and MFSW are closely related: HURST, also known as detrended fluctuation analysis, used in [10], measures a monofractal local fluctuation of the root-mean-square in a time series, whereas MFSW [30] analyzes the q-order Hurst

exponent, or multifractal fluctuation analysis, capable of distinguish fast and slow fluctuations. FMMI measures the time lag for which the signal adds a maximum of information about itself, or for which the information redundancy is minimal [29]. FZCF gives the input lag for which the autocorrelation function is minimal [29].

Another aspect that has been considered is the signal entropy, or the amount of information carried by the signal. Different approaches can be found in the scientific literature: SHANNON is a classical communication theory measurement of the information a signal carries [29]; PERMUTATION adds a perspective of symbolic dynamics, or the temporal order of the values in a series [32]; PPE quantifies the lack of control over pitch beyond natural vibrato and microtremor [27,40]. Finally, LZ measures the regularity or repetitiveness of a sequence [33].

2.4. *Feature selection and classification*

We built different systems for each database-disease combination. Those systems were created using clean samples from the databases, and their ability to handle additive noise was checked by inducing different types of noise at different SNR levels. The systems creation comprised two steps: feature selection and recording classification.

Given the number of features considered and datasets sizes, the risk of overfitting is a relevant issue, whichever classifier is used. To avoid this inconvenience, the following feature selection approach has been designed and implemented.

In general, features belonging to the same family, that is, those which share a common base algorithm, are highly correlated within the group [41]. This is shown in Fig. 2, which represents a heat map of the Pearson correlation coefficient for each pair of features. It can be observed that jitter, shimmer, and LZ features are highly correlated within their groups. Therefore, prior to any WOA related computation, the feature set was reduced to keep only one feature per group in the case of these three families.

The number of features considered after discarding the highly correlated ones is still high compared to the number of individuals included in each database, so further feature selection is performed. We used WOA [12], a bio-inspired evolutionary algorithm properly modified as a wrapper feature selection operator [14], which has recently started to be tested as a feature selection method [13]. It mimics the bubble-net feeding in the hunting behavior of the humpback whales. These whales hunt close to the surface by creating a net of bubbles where the prey is trapped. The algorithm mimics this behavior in two phases: one of them is called exploitation, when a whale herd tries to encircle a prey (solution or, in this case, set of features) in a spiral bubble-net attack; the other phase, called exploration, searches randomly for a new prey.

In each iteration, the algorithm selects a prey, a local optimum point. WOA selection algorithm relies on the fitness function from Eq. (1)

$$f = \alpha \times (1 - accuracy) + \beta \times \frac{number\ of\ selected\ features}{number\ of\ features} \quad (1)$$

based on the accuracy of a given classifier and the number of features selected to train such classifier. In this case, the objective is to maximize the accuracy, that is, minimize the error rate through the α parameter while minimizing the relative number of features using the β parameter, thus decreasing the risk of overfitting due to an excessive number of features involved. Both accuracy and relative number of features are in the range $[0, 1]$, and α and β also range $[0, 1]$ being $\beta = 1 - \alpha$.

Exploitation or prey encircling is done by taking the local optimum point obtained in the previous iteration, or a random point at the beginning of the execution, and then each search agent or "whale" describes a spiral around that point. To create such spiral the whale alters the optimum point, consisting of a feature set, and modifies it by adding or removing features, ensuring a lower euclidean distance to the optimal point in each iteration, so the new candidate obtained by each search agent is always closer to the local optimum point at each iteration.

At this point, as suggested by [42], we changed the updating mechanisms. Feature addition or subtraction in the solutions is performed using Eq. (2),

$$\vec{X}(t+1) = D' \cdot e^{bl} \cdot \cos(2\pi l) + \vec{X}^*(t), \quad (2)$$

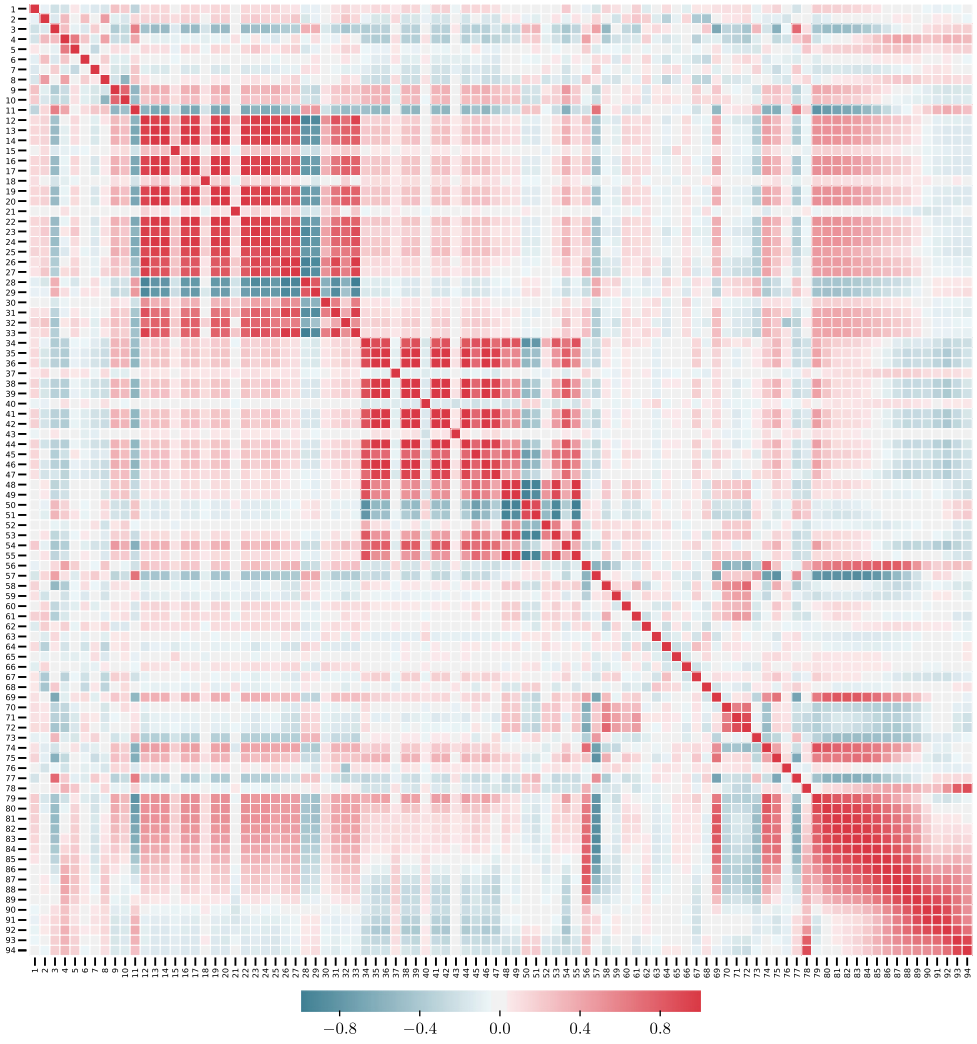
where b defines the spiral shape, l is a random number in $[-1, 1]$, D' is the euclidean distance to the best available solution and $\vec{X}^*$ is the best solution so far, as depicted in [12]. In this case, since the solutions space is discrete (a feature is either present or not) the updated leading position is transformed into a binary vector whose positions indicate whether the whale position in a given dimension or feature is above 0.5 or not (e.g. 4-dimensional solution [0.7, 0.3, 0.8, 0.9] would turn into [1, 0, 1, 1]).

In order to extend search to a wider portion of the solutions space, some of the whales will move randomly to another unrelated point in the space, what constitutes the exploration mechanism. Eventually, one or more whales will find a better solution than the temporary optimal one given by the last iteration, and then, all the search agents will turn to the best solution in terms of accuracy and feature number, and will start encircling it. The algorithm ends when it finds a solution with a fitness function lower than a given threshold, or when it has computed a maximum number of iterations.

The algorithm can be fine tuned by using tournament and roulette wheel selection mechanisms instead of a random operator to enhance the exploration phase, as well as crossover and mutation to optimize the exploitation phase [14]. In this case we have implemented the algorithm based on tournament selection as selection mechanism, and mutation as subset search.

Tournament selection randomly chooses two challengers within the search agents population and, according to a random number being greater than a given threshold, selects either the best or worst fitted candidate as new individual.

Mutation provides a tool for generating new possible solutions from actual solutions being considered in the current state. It randomly changes the state of some features from selected to un-selected or vice versa. The number of altered selections decreases as the algorithm reaches the hard limit of iterations, making it more prone to mutations at the beginning of the execution and more unlikely to mutate towards the end.



1 AGE
2 GENDER
3 CPP
4 GNE mean
5 GNE std
6 GNE SNR TKEO
7 GEN NER SEO
8 GQ prc5-95
9 GQ std cycle open
10 GQ std cycle closed
11 HNR
12 JITTER abs diff
13 JITTER diff percent
14 JITTER PQ3 classical Schoentgen
15 JITTER PQ3 classical Baken
16 JITTER PQ3 generalised Schoentgen
17 JITTER PQ5 classical Schoentgen
18 JITTER PQ5 classical Baken
19 JITTER PQ5 generalised Schoentgen
20 JITTER PQ11 classical Schoentgen
21 JITTER PQ11 classical Baken
22 JITTER PQ11 generalised Schoentgen
23 JITTER abs 0th perturb
24 JITTER DB
25 JITTER CV
26 JITTER TKEO mean
27 JITTER TKEO std
28 JITTER TKEO prc5
29 JITTER TKEO prc25
30 JITTER TKEO prc75
31 JITTER TKEO prc95
32 JITTER FM
33 JITTER range 5 95 perc
34 SHIMMER abs diff
35 SHIMMER diff percent
36 SHIMMER PQ3 classical Schoentgen
37 SHIMMER PQ3 classical Baken
38 SHIMMER PQ3 generalised Schoentgen
39 SHIMMER PQ5 classical Schoentgen
40 SHIMMER PQ5 classical Baken
41 SHIMMER PQ5 generalised Schoentgen
42 SHIMMER PQ11 classical Schoentgen
43 SHIMMER PQ11 classical Baken
44 SHIMMER PQ11 generalised Schoentgen
45 SHIMMER abs 0th perturb
46 SHIMMER DB
47 SHIMMER CV
48 SHIMMER TKEO mean
49 SHIMMER TKEO std
50 SHIMMER TKEO prc5
51 SHIMMER TKEO prc25
52 SHIMMER TKEO prc75
53 SHIMMER TKEO prc95
54 SHIMMER FM
55 SHIMMER range 5 95 perc
56 MFCC 1
57 MFCC 2
58 MFCC 3
59 MFCC 4
60 MFCC 5
61 MFCC 6
62 MFCC 7
63 MFCC 8
64 MFCC 9
65 MFCC 10
66 MFCC 11
67 MFCC 12
68 MFCC 13
69 D2
70 FMMI
71 FZCF
72 HURST
73 MFSW
74 ZCR
75 PERMUTATION
76 PPE
77 RPDE
78 SHANNON
79 LZ 2
80 LZ 4
81 LZ 8
82 LZ 16
83 LZ 32
84 LZ 64
85 LZ 128
86 LZ 256
87 LZ 512
88 LZ 1024
89 LZ 2048
90 LZ 4096
91 LZ 8192
92 LZ 16384
93 LZ 32768
94 LZ 65536

Fig. 2 – Correlation heat map for all the extracted features.

The algorithm also makes use of crossover, where two candidate solutions are mixed, or "bred", in order to create a new candidate solution with mixed characteristics of the two original ones.

The overall procedure is shown as Algorithm 1. It begins initializing candidates in the search field. In every iteration, until it reaches the hard limit imposed or achieves a fitness function above a desired threshold, it performs the following actions. First, an update of the a, A, C, l, p parameters is performed for each whale. a decreases linearly from 2 to 0 as the number of iterations get closer to the hard limit; A and C define the whale position update along with a: A is a coefficients vector built using the value of a and a random vector in [0, 1] and C is built using the same random vector; l is a random number in $[-1, 1]$ which defines the spiral shape as seen in Eq. 2; and p is a random number in [0, 1] whose value determines whether the whale is going to encircle the best solution (exploit) or it is going to explore, and how. Then, if it chooses to explore, it either explores the solutions space by performing a tournament selection or mutation of the best solution, creating a new candidate by crossover. If it chooses to exploit the current best solution, the process is completed by encircling in a spiral shaped curve the best solution.

Algorithm 1. Whale Optimization Algorithm

```
leaderScore = ∞
candidates = random(searchagents, features)
while leaderScore < threshold doanditerations < maxIterations
do
  for all candidates do
    Update a, A, C, l, p following [14]
    if p > 0.5 then
      if |A| > 1 then
        Xrand ⇐ tournament selection
        RA ⇐ mutate(Xrand)
        RE ⇐ mutate(candidate)
        candidateLeftarrowcrossover(RA, RE)
      else
        D ⇐ mutate(leaderPosition)
        candidate ⇐ crossover(D, candidate)
      end if
    else
      Encircle LeaderPosition using Eq. (2)
    end if
  end for
  for all candidates do
    if fitness(candidate) < leaderScore then
      leaderScore ⇐ fitness(candidate)
      leaderPosition ⇐ candidate
    end if
  end for
  iterations ⇐ iterations + 1
end while
```

For the classifier, SVM is considered. Prior to any computation a grid search is performed to find the best parameters for each database, and only for the case without additional noise (called "clean" case) as we intend to show the effects of noise on classification accuracy. The search space includes the kernel function used, among the four implemented ones in Python scikit library (linear, poly, rbf, and sigmoid) as well as their specific parameters.

Given the database sizes, one single run of WOA algorithm could yield a feature set fitted to the training set and the initial random conditions used for that particular run, reaching a local optimal point not suitable for most work settings, thus the need of multiple runs in order to generalize performance. Stratified shuffle and split was performed, all the selected feature sets were collected, and the most repeated features were compiled as the optimal feature set for each database and condition.

3. Results

Experiments were carried out to check the performance and robustness of different detection systems for two databases of voice disorders: nodules and Reinke's edema. This section describes the experimental setting and the main results obtained.

3.1. *Experimental setting*

The experiments consisted of two steps: first, classification systems were built minimizing the number of features used in each case (each database and each disease); then, those features were used to classify the same subjects from the databases they were created from, with and without added noise in the voice recordings, under a stratified repeated random subsampling validation framework.

As there are two heterogeneous databases to work with, some previous steps were taken in order to ensure a reliable results comparison. Most of them, concerning technical recording characteristics like sampling rate or recording length are summarized in Section 2.1. However, UEX-Voice database consists in four recordings per participant in the experiment. In this case, considered features were extracted for each recording, and mean value for each one was used in the following experiments.

In order to minimize the feature set, for each voice database (without artificially-added noise) the collection of results was obtained as follows. 640 instances of the WOA algorithm were launched, each one consisting of 640 whales, and a maximum of 25000 iterations to find the optimum features set. Preliminary studies were carried out to get values for α and β that both yield good accuracy and low feature set size. The values chosen for this specific problem were $\alpha = 0.99, \beta = 0.01$. The execution provided 640 different sets, represented by binary vectors of length 35, where 1 represents the presence of a feature, and 0 the absence of the feature in the set. By adding all the vectors as if they were natural vectors we end up with a total appearances vector.

The most useful features were used to train a set of classifiers, one set per disease, using an increasing number of features. They were incrementally added in the most repeated

order obtained by the WOA algorithm, and classifier performance was computed until no improvement was found for at least three feature additions. At this point, the feature set yielding the local maximum was chosen, so it was possible to check the evolution of the classifier F1 score with respect to the number of features used. Validation of results was performed by stratified repeated random subsampling, by repeating this procedure 1000 times and averaging the results. Training and test sets were selected using a stratified shuffle and split schema, so in each repetition the ratio of healthy and pathological individuals remained constant and identical to the ratio present for the database and disease being considered in each experiment. 2/3 of randomly chosen subjects from the database were used as training set and 1/3 as testing set.

In order to check the robustness of these systems, two different sources of additive noise were used: artificially generated Gaussian white noise and an actual recording of noise taken inside HSPdA. Two different scenarios were considered within each case: taking a random sample within noise recording by randomly selecting a starting point from the noise vector for every single voice recording in each database, and adding both noise sample and voice recording, what inherently introduces more variability in the process; and taking a random sample by randomly selecting a starting point within the noise vector and adding this unique sample to every recording in the database.

This process was repeated from a signal to noise power ratio (SNR) ranging from 0 dB to 30 dB in steps of 6 dB. Since the UEX-Voice database recording conditions are known, we can assume that the noise level present in the recording session is proportional to the value provided in Section 2.2, although we have no means to quantify the voice signal power. On the other hand, we are considering that the MEEI database was recorded in such good acoustical and technical conditions that the noise contained in the recordings is negligible, and as such will not alter significantly the induced SNR. For each SNR level considered, the same training and test sets for each run of the classifier were considered, so we can avoid the variability that random sets would induce in the different classifiers, so differences in the results obtained are a consequence only of the induced noise in each case.

3.2. *Experimental results*

The next subsections summarize the results obtained for each disease, once applied the different levels of noise and trained the set of classifiers. For each database and disease, four experiments were considered, which relate to a particular combination of noise nature (white synthetic noise, or actual recorded noise) and randomness (same noise clip added to every sample in the database, or one randomly generated or selected clip per sample).

We have used confusion matrix analytics to measure the performance of the final classifiers. True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) were computed for each iteration, and accuracy, precision, recall, and F1 score (Eqs. (3)–(6)) were derived from them, and were later averaged:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3}$$

$$precision = \frac{TP}{TP + FP}, \tag{4}$$

$$recall = \frac{TP}{TP + FN}, \tag{5}$$

$$F1\ score = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \tag{6}$$

In each of the four plots in Figs. 3–6, the X-axis shows which features have been selected and, as we move to the right, we add the features to the subset being considered, so the curve represents at a given point the mean F1 score on Y-axis, obtained after stratified repeated random subsampling validation. The upper limit for the number of features has been selected taking in consideration the shape of the clean curve in each case as stated in Section 3.1.

Each plot shows seven curves, one obtained after training the classifiers using the original recordings, no noise added, called *clean*, and six graphs labeled after SNR levels ranging from 0 dB to 30 dB in steps of 6 dB.

3.2.1. *Nodules*

Figs. 3, 4 and Tables 3, 4 show the mean F1 scores using an increasing number of features for voices affected with nodules in MEEI and UEX-Voice, respectively. In the case without noise addition, the results show that the classification F1 scores decrease from 0.91 to 0.61 when moving from MEEI database (Fig. 3)) to UEX-Voice database (Fig. 4). In this case, the implemented procedure has allowed to identify a reduced feature subset (4 or 5 features, depending on the database) that allows to achieve a saturation behavior in the prediction performance. In the case of UEX-Voice database, these features are CPP and three MFCCs, that is, cepstral and spectral features. For MEEI the most useful features resulted to be: MFCC1, CPP, HURST, AGE, and MFSW, which is a mixture of features based on linear and non-linear analysis, and the age.

F1 scores for MEEI database when adding 0 dB SNR noise are not even computable as we can not compute precision as well. This shows that the classifier marks every subject as healthy: Eq. (4) shows that, if there are neither TP nor FP (all the subjects classified as pathological), precision is not computable; also, by Eq. (5), recall equals 0. Looking at recall, which for binary classification shows the ability to detect pathological voices, for MEEI database we get values over 0.9 only for SNRs above 24 dB, and even then precision does not get over 0.9. In the case of UEX-Voice database F1 score, precision, and recall are lower, specially the latter.

The overall behavior results as expected, with higher SNR levels yielding better results, closer to the clean samples classifications. However, that behavior varies as we change the nature of the noise: Actual noise (subplots (a) and (b) in Figs. 3 and 4) tends to be less problematic when taking into consideration a few features, staying closer to the clean samples classifiers than synthetic noise (subplots (c) and (d) in Figs. 3 and 4).

3.2.2. *Reinke's edema*

Figs. 5, 6 and Tables 5, 6 show the discrimination results obtained in the case of Reinke's edema. The comparison between both databases in the case without additional noise

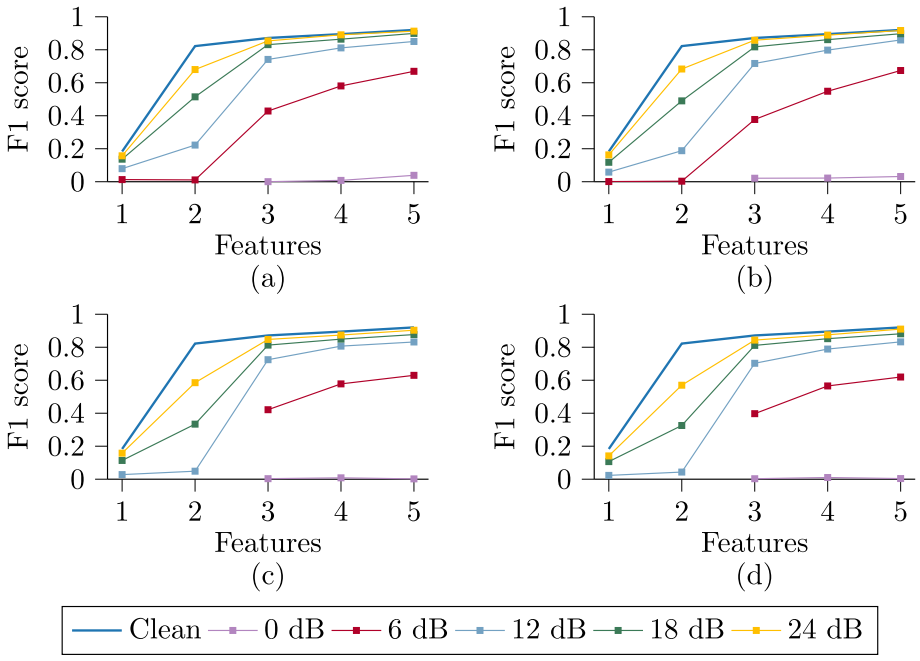


Fig. 3 – Mean F1 scores using cumulative features for nodules disease, MEEI database. SNRs ranging from 0 dB to 30 dB in steps of 6 dB. The features are: 1-MFCC1, 2-CPP, 3-HURST, 4-AGE, 5-MFSW. Noise characteristics: (a) realistic noise, fixed sample, (b) realistic noise, random sample, (c) white noise, fixed sample, (d) white noise, random sample.

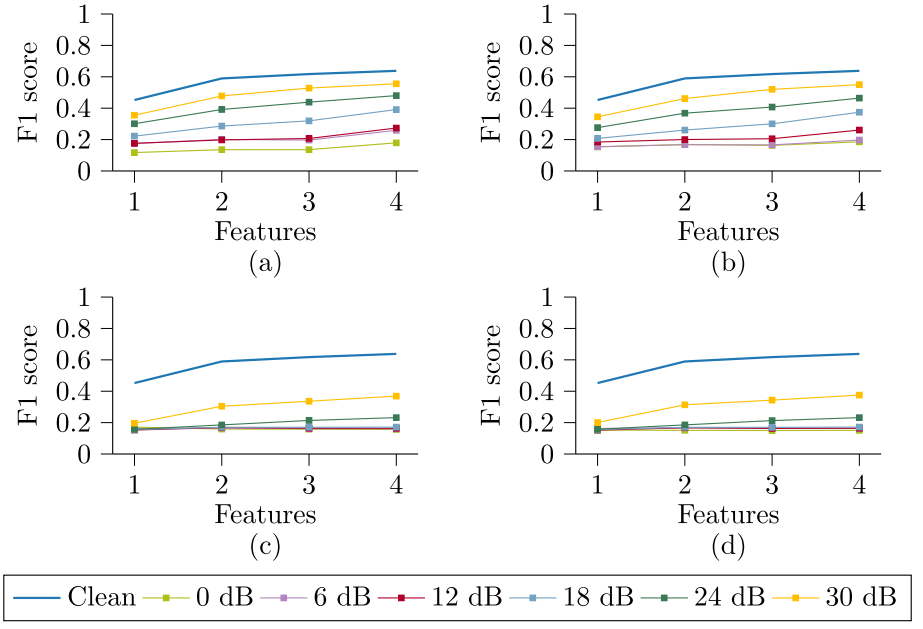


Fig. 4 – Mean F1 scores using cumulative features for nodules disease, UEX-Voice database. SNRs ranging from 0 dB to 30 dB in steps of 6 dB. The features are: 1-CPP, 2-MFCC7, 3-MFCC3, 4-MFCC2. Noise characteristics: (a) realistic noise, fixed sample, (b) realistic noise, random sample, (c) white noise, fixed sample, (d) white noise, random sample.

allows to extract similar conclusions than in the case of nodules. Again, the detection F1 score obtained with MEEI database is higher than in the case of UEX-Voice (0.98 versus 0.83). The number of features needed to reach a saturation behavior is 5 or 6, depending on the database. Cepstral and spectral features play again a relevant role, however an entropy feature is required in both feature subsets. In the case of MEEI, shimmer is also selected.

In the presence of additive noise, the detection performance decreases, and the impact is again higher in the case

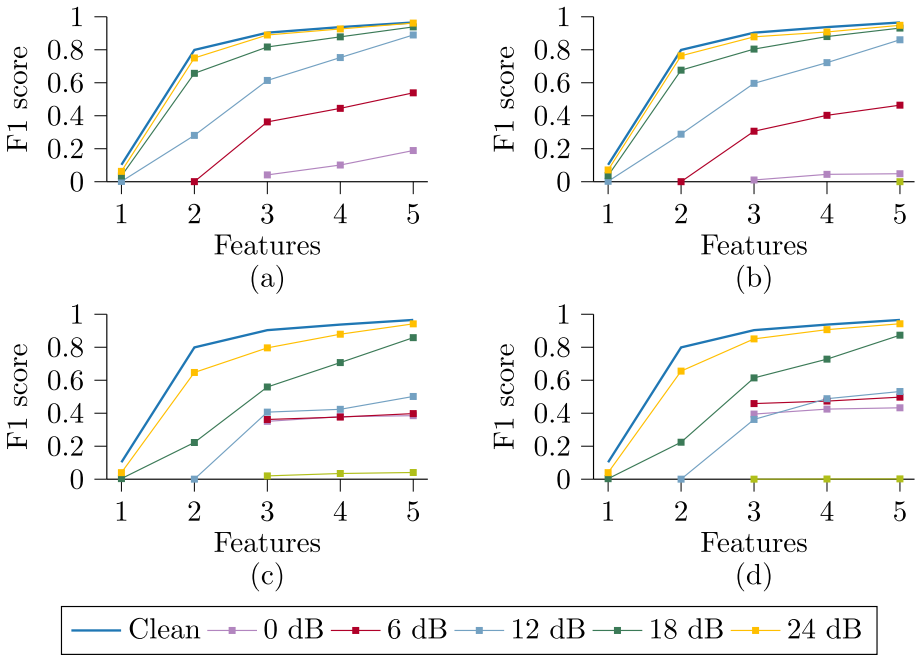


Fig. 5 – Mean F1 scores using cumulative features for Reinke's edema, MEEI database. SNRs ranging from 0 dB to 30 dB in steps of 6 dB. The features are: 1-MFCC1, 2-PERMUTATION, 3-Shimmer, 4-MFCC3, 5-CPP. Noise characteristics: (a) realistic noise, fixed sample, (b) realistic noise, random sample, (c) white noise, fixed sample, (d) white noise, random sample.

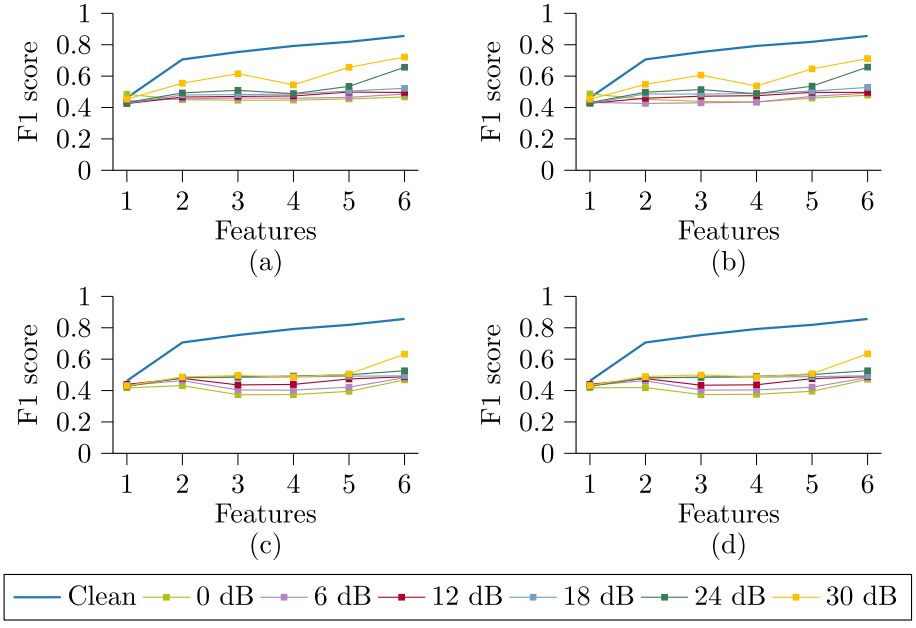


Fig. 6 – Mean F1 scores using cumulative features for Reinke's edema, UEX-Voice database. SNRs ranging from 0 dB to 30 dB in steps of 6 dB. The features are: 1-MFCC7, 2-CPP, 3-MFCC2, 4-SHANNON, 5-MFCC10, 6-MFCC4. Noise characteristics: (a) realistic noise, fixed sample, (b) realistic noise, random sample, (c) white noise, fixed sample, (d) white noise, random sample.

of synthetic white Gaussian noise than in the case of realistic noise. Also, as it happens in the experiment about nodules, the effect of noise addition is more pronounced on UEX-Voice database than in the case of MEEI.

F1 scores for MEEI database along with accuracy show that the classifier is reliable for SNRs as low as 18 dB, where both values reach over 0.9 in the case of realistic noise. For UEX-Voice database, the minimum SNR to show acceptable perfor-

Table 3 – Accuracy, precision, recall and, F1 score values computed for MEEI database, nodules disease, SNRs ranging from 0 dB to 30 dB in 6 dB steps, using 5 features: MFCC1, CPP, HURST, AGE, and MFSW.

		Real fixed	Real random	White fixed	White random
Clean	Accuracy	0.95	0.95	0.95	0.95
	Precision	0.88	0.88	0.88	0.88
	Recall	0.95	0.95	0.95	0.95
	F1 score	0.91	0.91	0.91	0.91
30 dB	Accuracy	0.95	0.95	0.95	0.95
	Precision	0.87	0.88	0.86	0.87
	Recall	0.95	0.95	0.94	0.95
	F1 score	0.91	0.91	0.90	0.90
24 dB	Accuracy	0.95	0.94	0.94	0.94
	Precision	0.86	0.85	0.85	0.86
	Recall	0.94	0.92	0.91	0.91
	F1 score	0.90	0.88	0.88	0.89
18 dB	Accuracy	0.93	0.92	0.92	0.93
	Precision	0.85	0.81	0.87	0.88
	Recall	0.89	0.88	0.83	0.83
	F1 score	0.87	0.85	0.85	0.85
12 dB	Accuracy	0.88	0.87	0.88	0.88
	Precision	0.84	0.83	0.91	0.90
	Recall	0.68	0.64	0.59	0.58
	F1 score	0.75	0.72	0.71	0.70
6 dB	Accuracy	0.75	0.76	0.75	0.75
	Precision	0.85	0.94	0.98	0.98
	Recall	0.04	0.05	0.01	0.01
	F1 score	0.07	0.10	0.01	0.02
0 dB	Accuracy	0.74	0.74	0.74	0.74
	Precision	–	–	–	–
	Recall	0.00	0.00	0.00	0.00
	F1 score	–	–	–	–

mance is 24 dB, where F1 score drops from 0.83 to 0.71 and accuracy, precision and recall score show a similar degradation.

4. Discussion

The results involve two diseases, two different databases and four kinds of noise. This allows to perform a comparative analysis from different perspectives. In spite of the fact that we have studied the effects of noise addition in the performance of a classifier using F1 score as the reference metric, most studies use accuracy as the main performance indicator [9]. However, since we have also computed accuracy, and the best results are obtained using all the features selected in each case, we can compare our system performance with prior research in the field.

Comparing clean case performance allows us to analyze the differences from a database point of view, with MEEI database the detection accuracies reach 0.95, while the systems reach 0.71 for nodules and 0.84 for Reinke's edema with UEX-Voice database. This difference in performance between MEEI and a database obtained in more realistic conditions is in line with the scientific literature. Whereas most reported detection accuracies for MEEI data are in excess of 0.9, in [8] best accuracies of 0.784 and 0.762 were achieved after carrying out voice pathology detection experiments using the Hospital Universitario Príncipe de Asturias database (HUPA) and the Saarbrücken Voice Disorder database (SVD), respectively. [43] computed accuracy, recall, and other metrics when classifying recordings from MEEI (0.91–0.97 accuracy, 0.93–0.98 recall) and HUPA databases (0.68–0.82 accuracy, 0.77–0.85 recall). Moreover, [10] achieves 0.95/0.97 accuracy/recall for MEEI database using spectral-cepstral features, while the results with HUPA database using the same features only reach 0.78/0.74.

Some studies have taken into consideration noise corruption. For example, [24] studies environmental noise and white Gaussian noise effects on Parkinson's disease detection using a variety of vocal tasks including a phonation model based on sustained vowels. Both disease and noise are not directly comparable since Parkinson's disease is a neurological disease and different diseases require different analysis techniques which depend on the specific effects on voice [10]. Moreover, noise was recorded in 8 different scenarios. However, it shows that with clean training the accuracy for the phonation model drops from about 0.7 to 0.5 when SNR is equal to 0 dB, much like the results obtained here. Furthermore, some research has been made in order to alleviate the effects of different recording conditions on disease detection performance [25].

Table 4 – Accuracy, precision, recall and, F1 score values computed for UEX-Voice database, nodules disease, SNRs ranging from 0 dB to 30 dB in 6 dB steps, using 4 features: CPP, MFCC7, MFCC3, and MFCC2.

		Real fixed	Real random	White fixed	White random
Clean	Accuracy	0.71	0.71	0.71	0.71
	Precision	0.74	0.74	0.74	0.74
	Recall	0.52	0.52	0.52	0.52
	F1 score	0.61	0.61	0.61	0.61
30 dB	Accuracy	0.67	0.67	0.59	0.59
	Precision	0.74	0.73	0.59	0.59
	Recall	0.41	0.40	0.24	0.25
	F1 score	0.53	0.52	0.34	0.35
24 dB	Accuracy	0.64	0.63	0.53	0.53
	Precision	0.70	0.68	0.43	0.43
	Recall	0.33	0.32	0.14	0.14
	F1 score	0.45	0.44	0.21	0.21
18 dB	Accuracy	0.60	0.59	0.51	0.51
	Precision	0.63	0.59	0.35	0.35
	Recall	0.26	0.25	0.11	0.11
	F1 score	0.37	0.36	0.16	0.16
12 dB	Accuracy	0.55	0.54	0.51	0.50
	Precision	0.48	0.45	0.32	0.32
	Recall	0.16	0.17	0.10	0.10
	F1 score	0.24	0.24	0.15	0.15
6 dB	Accuracy	0.54	0.53	0.50	0.50
	Precision	0.46	0.40	0.30	0.30
	Recall	0.16	0.12	0.09	0.09
	F1 score	0.24	0.19	0.14	0.14
0 dB	Accuracy	0.54	0.53	0.49	0.49
	Precision	0.43	0.41	0.28	0.27
	Recall	0.11	0.12	0.08	0.08
	F1 score	0.17	0.18	0.13	0.13

On the other side, when noise is added with a low SNR, MEEI database gets much higher results for all metrics than UEX-Voice. Apart from the fact that MEEI database was collected in a more controlled acoustic environment, some authors have pointed out that this database contains no lightly pathological speakers [29], and that the normal and dysphonic voices present in the database are easily separable [44], which makes the classification task easy.

Given the proportions of healthy and pathological samples present in the databases, shown in Table 1, MEEI test set contains roughly 70% of healthy patients whereas in UEX-Voice 50% of test samples are healthy. Those ratios match the accuracies obtained for the worst SNR ratios for all the classifiers for both databases. Precision and recall values support the fact that the classifier is unable to distinguish pathological subjects and marks most of them as healthy. That explains the differences in the lower accuracy levels shown between Tables 3, 5 and 4, 6.

Considering the different kinds of noise it seems that realistic noise is less intrusive than white synthetic noise. This trend is specially pronounced for UEX-Voice database. A possible explanation for this is that the spectral compositions of both types of noise are different. White noise (Fig. 7a) is characterized by an even spectral power density, thus all the frequencies in the full bandwidth are interfered in the same way. However, realistic noise coming from several sources in the hospital environment concentrates most energy in a lower part of the spectrum. A spectrogram of an example of realistic noise segment is shown in Fig. 7b, where it is easy to see the spectral contributions of the noise sources taken in consideration, and how realistic noise most prominent frequencies lie in the lower half of spectrum, and even considering that bandwidth, frequencies below 4 kHz stand out.

Regarding noise randomness, the variability introduced by random noise sampling in all cases has little impact in the overall capacity of the resulting classifiers. Although some differences exist in the results, there is no consistency in any advantage of fixed over random sampling or vice versa, as we can see, for example, in Fig. 5, subplots (a) versus (b) or Fig. 5, subplots (c) versus (d).

The comparative analysis of the results from a disease perspective is more challenging. Vocal fold nodules are smooth, benign masses involving anterior or middle vocal folds and located superficially to the free edge of the fold. Reinke's edema (also known as polypoid degeneration) is characterized by an accumulation of fluid, usually in both vocal folds [45]. Since both pathologies share some histological characteristics, [2] even proposed to use the term "exudative lesions on the Reinke's space" to refer to nodules, polyps, and Reinke's edema. These histological characteristics affect the vibra-

Table 5 – Accuracy, precision, recall and, F1 score values computed for MEEI database, Reinke's edema disease, SNRs ranging from 0 dB to 30 dB in 6 dB steps, using 5 features: MFCC1, PERMUTATION, Shimmer, MFCC3, and CPP.

		Real fixed	Real random	White fixed	White random
Clean	Accuracy	0.99	0.99	0.99	0.99
	Precision	1.00	1.00	1.00	1.00
	Recall	0.96	0.96	0.96	0.96
	F1 score	0.98	0.98	0.98	0.98
30 dB	Accuracy	0.98	0.98	0.98	0.97
	Precision	1.00	1.00	1.00	1.00
	Recall	0.94	0.93	0.92	0.91
	F1 score	0.97	0.96	0.96	0.96
24 dB	Accuracy	0.98	0.97	0.93	0.94
	Precision	1.00	1.00	1.00	1.00
	Recall	0.93	0.91	0.80	0.81
	F1 score	0.96	0.96	0.89	0.90
18 dB	Accuracy	0.94	0.95	0.84	0.84
	Precision	1.00	1.00	0.99	0.97
	Recall	0.83	0.84	0.50	0.52
	F1 score	0.91	0.91	0.66	0.67
12 dB	Accuracy	0.84	0.81	0.81	0.82
	Precision	1.00	1.00	1.00	1.00
	Recall	0.50	0.41	0.39	0.43
	F1 score	0.67	0.58	0.56	0.61
6 dB	Accuracy	0.78	0.76	0.77	0.81
	Precision	0.94	0.93	0.82	0.88
	Recall	0.32	0.26	0.37	0.46
	F1 score	0.47	0.41	0.51	0.60
0 dB	Accuracy	0.69	0.69	0.71	0.69
	Precision	0.87	0.72	0.88	0.77
	Recall	0.03	0.03	0.12	0.06
	F1 score	0.06	0.06	0.20	0.11

tory patterns of the vocal folds (increase in mass of the folds, reduction in the pliability of the overlying cover...), and may produce some common perceptual consequences, such as hoareseness and breathiness. Nevertheless, it can be observed that Reinke's edema is detected with higher accuracy and F1 score (0.99 and 0.98 respectively in the case of MEEI; 0.84 and 0.83 respectively in the case of UEX-Voice) than nodules (0.95 and 0.91 respectively in the case of MEEI; 0.71 and 0.61 respectively in the case of UEX-Voice), which may be the consequence of its inflammatory character producing a more severe impact on voice quality in comparison to a simple mass lesion.

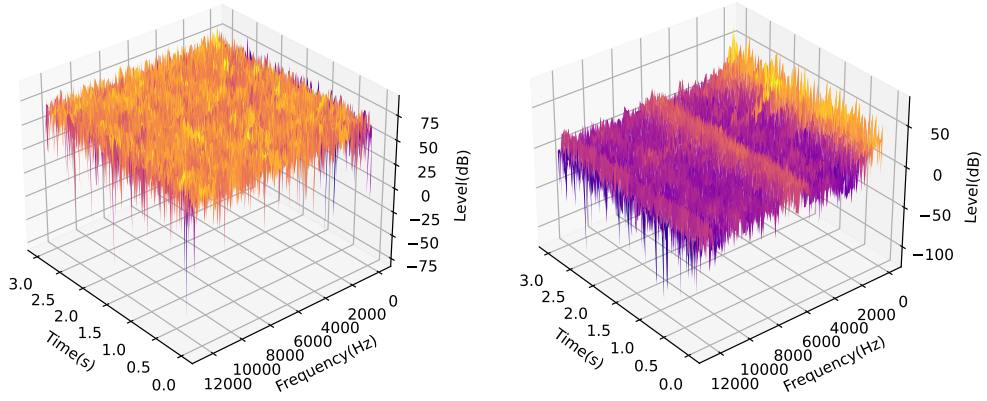
The overall structure of the system is in line with most of previous work, with the common steps of preprocessing, feature extraction, dimensionality reduction, machine learning training, and system evaluation [5]. Regarding dimensionality reduction, prior work in the field include techniques such principal component analysis, linear discriminant analysis, or minimum redundancy maximum relevance among others. The four experimental settings have led to four different feature subsets. The composition of these feature subsets is important as they may provide some clues not only on which features are more important when building a new detection system, but also which ones show a certain noise robustness.

Although the feature selection process is applied on the original databases, without noise addition, UEX-Voice is recorded in a more realistic acoustic environment, so it is possible to conclude that, if there are features that are selected using both databases, they may have a reliable discrimination potential across different databases under moderately controlled acoustic conditions. This is the case of CPP and MFCCs. They play a very important role, as CPP and at least one MFCC have been selected within the most useful features in the four cases, to the point that for nodules disease in UEX-Voice database all the selected features are MFCCs and CPP. Both share the advantage that, unlike traditional acoustic measures such as jitter or shimmer, they do not require a pitch estimation which may be difficult due to the absence of periodicity in severely dysphonic voices. This is in line with results obtained by [28], where it is shown that advanced multiband cepstral analysis might be useful in disease detection and even in disease discrimination, and [10] which shows the ability of spectral-cepstral features to classify disphonic voices based on a sustained vowel analysis.

The rest of selected features is heterogeneous among the four studied cases: Non-linear analysis features are found in Fig. 3 with HURST and MFSW, but no other case shows nonlinear features. Entropies make their appearance in both Reinke's edema cases, Figs. 5 and 6, with permutation and Shannon entropies, but not in nodules cases. From the classical perturbation measurements only shimmer is selected for MEEI Reinke's edema case, but not for UEX-Voice. The reason

Table 6 – Accuracy, precision, recall and, F1 score values computed for UEX-Voice database, Reinke's edema disease, SNRs ranging from 0 dB to 30 dB in 6 dB steps, using 6 features: MFCC7, CPP, MFCC2, SHANNON, MFCC10, and MFCC4.

		Real fixed	Real random	White fixed	White random
Clean	Accuracy	0.84	0.84	0.84	0.84
	Precision	0.84	0.84	0.84	0.84
	Recall	0.83	0.83	0.83	0.83
	F1 score	0.83	0.83	0.83	0.83
30 dB	Accuracy	0.76	0.76	0.68	0.69
	Precision	0.79	0.79	0.68	0.69
	Recall	0.71	0.70	0.68	0.68
	F1 score	0.75	0.74	0.68	0.69
24 dB	Accuracy	0.71	0.71	0.55	0.55
	Precision	0.72	0.71	0.55	0.55
	Recall	0.70	0.70	0.57	0.57
	F1 score	0.71	0.71	0.56	0.56
18 dB	Accuracy	0.56	0.56	0.47	0.47
	Precision	0.56	0.56	0.48	0.48
	Recall	0.58	0.57	0.51	0.51
	F1 score	0.57	0.57	0.49	0.49
12 dB	Accuracy	0.48	0.48	0.47	0.47
	Precision	0.48	0.48	0.47	0.47
	Recall	0.51	0.51	0.52	0.52
	F1 score	0.50	0.50	0.49	0.49
6 dB	Accuracy	0.48	0.48	0.48	0.48
	Precision	0.48	0.48	0.48	0.48
	Recall	0.51	0.51	0.52	0.52
	F1 score	0.50	0.50	0.50	0.50
0 dB	Accuracy	0.47	0.47	0.48	0.48
	Precision	0.47	0.47	0.48	0.48
	Recall	0.47	0.49	0.49	0.49
	F1 score	0.47	0.48	0.48	0.49



(a) White noise. (b) Realistic noise.

Fig. 7 – Spectrograms for white noise and an example of realistic noise segment from the hospital environment. Both noise recordings were used on the same voice recording.

might be that, as an amplitude perturbation measure, shimmer is very sensitive to noise, performing better in a more controlled acoustic environment.

On the classifier side, we chose SVM for its simplicity and execution speed since WOA feature selection is computationally expensive. Many alternatives have been used, line Hidden Markov Models, Gaussian Mixture Models, K-nearest neighbors or decision trees to name a few of them [9]. Most of the alternatives found in previous work uses that kind of algorithms, although in recent years artificial neural networks have gained popularity and we start to see studies using such techniques.

Deep learning methods have seldom been used in this specific application until recent times. [9] mentions artificial neural networks but only shows multilayer perceptron, which barely can be classified as a deep learning method. [46] presents 2 out of 45 studies using deep learning techniques, which date from 2019. The most plausible reason is database size. Looking at the numbers shown in Table 1, the number of samples is very low, and a small multilayer neural network comprises thousands of coefficients. DenseNet has been used on cepstrum features [47] with good results although it cites the low number of pathological samples as a limitation. Other classical deep learning approaches like VGG16 and CaffeNet have been used [48], with the particularity that those algorithms are used for image recognition and classification. Consequently, raw waveforms are feed into the network (in the form of spectrograms) since it will infer features, and transfer learning techniques (neural network partially trained with examples from other fields) are used to overcome the long training times and small dataset size limitations.

Research on robust pathology detectors has not been addressed until recent times. [49] performs experiments using four different databases, aiming at robustness against different recording conditions, but does not focus on specific differences between them. Little work has been done around noise robustness in voice quality assessment, so thorough comparisons can not be made, although this research is necessary. For example, [20,47] point towards differences in recording environment (e.g. background noise) as a limitation for different studies results comparison. [50] points out the difficulties to extrapolate the results obtained with different databases due their recording differences. However, [21] proposes a SNR level of 42 dB for perturbation measurements (jitter and shimmer) to be reliable, and estimates 30 dB as the lowest limit of SNR level for reliable usage of classifiers. This seems to match the results for MEEI database in Figs. 3 and 5, where the F1 score is almost identical for the clean and the 30 dB SNR cases, for all the numbers of features considered, specially when realistic noise is added.

Considering the impact of noise can benefit other research work focused on mobile health tools to detect vocal fold disorders. There is currently a high interest in the development of mobile-aided systems to manage a wide variety of diseases and, in particular, disorders affecting voice [17–19]. A critical aspect is to check if the approaches proposed for controlled conditions are robust or have to be modified when used in increasingly realistic environments.

5. Conclusion

The results of this paper highlight the importance of performing experiments on more realistic voice pathology databases, alternative to MEEI, since the achievable prediction accuracies are not expected to be comparable. The feature subsets obtained by feature selection with MEEI and with a more realistic database collected in the scope of this work emphasize the role of CPP and MFCCs as useful and robust features to discriminate pathological from healthy voices.

Also, the degrading impact of additive noise on AVCA systems based on acoustic features for detection of nodules and Reinke's edema has been demonstrated and quantified. Although the effect of real-world noise recorded in a clinical environment has been shown to be lower than that of white noise, the effect is sufficiently detrimental to motivate further research into noise-robust prediction systems.

In the future, it will be interesting to increase UEX-Voice database by including new organic pathologies. Also, exploring new techniques in the field like deep learning and looking for solutions to overcome the voice databases limitations are of research interest.

CRediT authorship contribution statement

Mario Madruga: Conceptualization, Software, Validation, Data curation, Writing - original draft, Visualization. **Yolanda Campos-Roca:** Conceptualization, Methodology, Resources, Writing - review & editing. **Carlos J. Pérez:** Conceptualization, Methodology, Validation, Formal analysis, Resources, Writing - review & editing, Supervision.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

The authors would like to thank Dr. Moreno for his medical advising, Sandra Paniagua and Esther de la O. for their work recording the UEX-Voice database in HSPdA and all the voluntary individuals, patients and healthy. They also would like to thank Prof. Gómez (from the Acoustics Laboratory) for making the sound level meter available and the Advanced Scientific Computation at the University of Extremadura to provide access to computation facilities.

This research has been supported by project MTM2017-86875-C3-2-R (Ministerio de Ciencia, Innovación y Universidades), projects IB16054, GR18108 and GR18055 (Junta de Extremadura/European Regional Development Funds, EU), and FPU18/03274 grant (Ministerio de Ciencia, Innovación y Universidades).

REFERENCES

[1] Rufo M, Martín J, Pérez C, Paniagua S. A Bayesian decision analysis approach to assess voice disorder risks by using acoustic features. Biometr J 2019;61(3):503–13.

[2] Hantzakos A, Remacle M, Dikkers FG, Degols JC, Delos M, Friedrich G, Giovanni A, Rasmussen N. Exudative lesions of Reinke's space: a terminology proposal. Eur Arch Otorhinolaryngol 2009;266(6):869.

[3] Echternach M, Döllinger M, Köberlein M, Kuranova L, Gellrich D, Kainz M. Vocal fold oscillation pattern changes related to loudness in patients with vocal fold mass lesions. J Otolaryngol Head Neck Surg 2020;49(1):1–9.

[4] Sataloff RT. Clinical assessment of voice. Plural publishing; 2017.
[5] Gómez-García J, Moro-Velázquez L, Godino-Llorente J. On the design of automatic voice condition analysis systems. Part I: Review of concepts and an insight to the state of the art. Biomed Sig Process Control 2019;51:181–99.
[6] Kowalska-Taczanowska R, Friedman A, Koziorowski D. Parkinson's disease or atypical parkinsonism? The importance of acoustic voice analysis in differential diagnosis of speech disorders. Brain Behav 2020;10(8) e01700.
[7] Paniagua M, Pérez C, Calle-Alonso F, Salazar C. An acoustic-signal-based preventive program for university lecturers' vocal health. J Voice 2018;34(1):88–99.
[8] Kadiri SR, Alku P. Analysis and detection of pathological voice using glottal source features. IEEE J Select Top Sig Process 2019;14(2):367–79.
[9] Hegde S, Shetty S, Rai S, Dodderi T. A survey on machine learning approaches for automatic detection of voice disorders. J Voice 2019;33(6):947–58.
[10] Orozco-Arroyave JR, Belalcazar-Bolanos EA, Arias-Londoño JD, Vargas-Bonilla JF, Skodda S, Rusz J, Daqrouq K, Hönig F, Nöth E. Characterization methods for the detection of multiple voice disorders: neurological, functional, and laryngeal diseases. IEEE J Biomed Health Inf 2015;19 (6):1820–8.
[11] J. Tang, S. Alelyani, and H. Liu, Feature selection for classification: A review, Data classification: Algorithms and Applications, 2014:37–64.
[12] Mirjalili S, Lewis A. The whale optimization algorithm. Adv Eng Softw 2016;95:51–67.
[13] Canayaz M, Demir M. Feature selection with the whale optimization algorithm and artificial neural network. In: 2017 International Artificial Intelligence and Data Processing Symposium (IDAP).
[14] Mafarja M, Mirjalili S. Whale optimization approaches for wrapper feature selection. Appl Softw Comput 2018;62:441–53.
[15] Massachusetts Eye and Ear Infirmary, Voice disorders database, Version 1.03 (cd-rom), Lincoln Park, NJ: Kay Elemetrics Corporation; 1994.
[16] Travieso CM, Alonso JB, Orozco-Arroyave JR, Solé-Casals J, Gallego-Jutglà E. Automatic detection of laryngeal pathologies in running speech based on the HMM transformation of the nonlinear dynamics. Int Conf Nonlinear Speech Process 2013.
[17] Arias-Vergara T, Vásquez-Correa J, Orozco-Arroyave J, Nöth E. Speaker models for monitoring Parkinson's disease progression considering different communication channels and acoustic conditions. Speech Commun 2018;101:11–25.
[18] Tsanas A, Little M, Ramig L. Remote assessment of parkinson's disease symptom severity using the simulated cellular mobile telephone Network. IEEE Access. 2021.
[19] Cesari U, De Pietro G, Marciano E, Niri C, Sannino G, Verde L. Voice disorder detection via an m-Health system: Design and results of a clinical study to evaluate Vox4Health. BioMed Res Int 2018;2018.
[20] Saggio G, Costantini G. Worldwide healthy adult voice baseline parameters: a comprehensive review. J Voice 2020.
[21] Deliyski DD, Shaw HS, Evans MK. Adverse effects of environmental noise on acoustic voice quality measurements. J Voice 2005;19(1):15–28.
[22] van der Woerd B, Wu M, Parsa V, Doyle P, Fung K. Evaluation of Acoustic Analyses of Voice in Nonoptimized Conditions. J Speech Language Hearing Res 2020;63(12):3991–9.
[23] Madruga M, Campos-Roca Y, Pérez CJ. Robustness assessment of automatic Reinke's edema diagnosis systems. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
[24] Vásquez-Correa JC, Serra J, Orozco-Arroyave JR, Vargas-Bonilla JF, Nöth E. Effect of acoustic conditions on algorithms to detect Parkinson's disease from speech. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2017.
[25] Madruga M, Campos-Roca Y, Pérez C. Multicondition training for noise-robust detection of benign vocal fold lesions from recorded speech. IEEE Access. 2020.
[26] Fraile R, Godino-Llorente JI. Cepstral peak prominence: A comprehensive analysis. Biomed Signal Process Control 2014;14:42–54.
[27] Tsanas A. Acoustic analysis toolkit for biomedical speech signal processing: concepts and algorithms. Models Anal Vocal Emiss Biomed Appl 2013;2:37–40.
[28] Alves M, Silva G, Bispo B, Dajer M, Rodrigues P. Voice disorders detection through multiband cepstral features of sustained vowel. J Voice 2021.
[29] Henríquez P, Alonso JB, Ferrer MA, Travieso CM, Godino-Llorente JI, Díaz-de María F. Characterization of healthy and pathological voice through measures based on nonlinear dynamics. IEEE Trans Audio Speech Language Process 2009;17 (6):1186–1195.
[30] Ihlen EAF. Introduction to multifractal detrended fluctuation analysis in matlab. Front Physiol 2012;3:141.
[31] Islam R, Tarique M, Abdel-Raheem E. A survey on signal processing based pathological voice detection techniques. IEEE Access 2020;8:66749–76.
[32] Riedl M, Müller A, Wessel N. Practical considerations of permutation entropy. Eur Phys J Spec Top 2013;222(2):249–62.
[33] Orozco JR, Vargas JF, Alonso JB, Ferrer MA, Travieso CM, Henríquez P. Voice pathology detection in continuous speech using nonlinear dynamics. In: 2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA).
[34] Behrman A. Speech and voice science. Plural publishing; 2017.
[35] Hixon TJ, Weismer G, Hoit JD. Preclinical speech science: Anatomy, physiology, acoustics, and perception. Plural Publishing; 2018.
[36] Van Houtte E, Van Lierde K, D'haeseleer E, Claeys S. The prevalence of laryngeal pathology in a treatment-seeking population with dysphonia. Laryngoscope 2010;120(2):306–312.
[37] Brückl M, Ghio A, Alain, Viallet F. Measurement of tremor in the voices of speakers with Parkinson's disease. Proc Comput Sci 2018;128:47–54.
[38] Illner V, Sovka P, Rusz J. Validation of freely-available pitch detection algorithms across various noise levels in assessing speech captured by smartphone in Parkinson's disease. Biomed Signal Process Control 2020;58 101831.
[39] Tsanas A. Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning, Ph.D. dissertation, Oxford University, UK, 2012.
[40] Tsanas A, Gómez-Vilda P. Novel robust decision support tool assisting early diagnosis of pathological voices using acoustic analysis of sustained vowels, in Multidisciplinary Conference of Users of Voice, Speech and Singing (JVHC 13); 2013.
[41] Despotovic V, Skovranek T, Schommer C. Speech based estimation of Parkinson's disease using Gaussian processes and automatic relevance determination. Neurocomputing 2020;401:173–81.
[42] Luan F, Cai Z, Wu S, Liu S, He Y. Optimizing the low-carbon flexible job shop scheduling problem with discrete whale optimization algorithm. Mathematics 2019;7(8):688.

[43] Arias-Londoño J, Godino-Llorente J, Sáenz-Lechón N, Osma-Ruiz V, Víctor, Castellanos-Domínguez G. An improved method for voice pathology detection by means of a HMM-based feature space transformation. Pattern Recognit 2010;43 (9):3100–12.
[44] Daoudi K, Bertrac B. On classification between normal and pathological voices using the MEEI-KayPentax database: Issues and consequences. In: Fifteenth Annual Conference of the International Speech Communication Association.
[45] Sataloff RT, Chowdhury F, Portnoy JE, Hawkshaw MJ, Joglekar S. Surgical techniques in otolaryngology-head & Neck Surgery: Laryngeal Surgery. JP Medical Ltd 2013.
[46] Syed S, Rashid M, Hussain S. Meta-analysis of voice disorders databases and applied machine learning techniques. Math Biosci Eng: MBE 2020;17(6):7958–79.
[47] Fang S, Tsao Y, Hsiao M, Chen J, Lai Y, Lin F, Wang C. Detection of pathological voice using cepstrum vectors: A deep learning approach. J Voice 2019;33(5):634–41.
[48] Alhussein M, Muhammad G. Voice pathology detection using deep learning on mobile healthcare framework. IEEE Access 2018;6:41034–41.
[49] Harar P, Galaz Z, Alonso-Hernandez Jesus B, Mekyska J, Burget R, Smekal Z. Towards robust voice pathology detection. Neural Comput Appl 2018:1–11.
[50] Karan B, Sahu S, Mahto K. Parkinson disease prediction using intrinsic mode function based features from speech signal. Biocybern Biomed Eng 2020;40(1):249–64.