Biomedical Signal Processing and Control 89 (2024) 105706

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Precision medicine in ALS: Identification of new acoustic markers for dysarthria severity assessment

Raffaele Dubbioso [a,*,1], Myriam Spisto [b,1], Laura Verde [c], Valentina Virginia Iuzzolino [a], Gianmaria Senerchia [a], Giuseppe De Pietro [d], Ivanoe De Falco [d], Giovanna Sannino [d]

[a] *Department of Neurosciences, Reproductive Sciences and Odontostomatology, University of Naples "Federico II", Via Sergio Pansini, 5, Naples, 80131, Italy*
[b] *Department of Psychology of the University of Campania "Luigi Vanvitelli", Viale Ellittico, 31, Caserta, 81100, Italy*
[c] *Department of Mathematics and Physics of the University of Campania "Luigi Vanvitelli", Viale Abramo Lincoln, 5, Caserta, 81100, Italy*
[d] *National Research Council of Italy (CNR), Institute for High-Performance Computing and Networking (ICAR), Via Pietro Castellino, 111, Naples, 80131, Italy*

ARTICLE INFO

Keywords:
Precision medicine
Artificial intelligence
Bulbar functions
Dysarthria
Amyotrophic lateral sclerosis (ALS)
Classification

ABSTRACT

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease affecting motorneurons of the bulbar, cervical, thoracic, or lumbar segments. Bulbar presentation is a devastating characteristic that impairs patients' ability to communicate and is linked to shorter survival. Early acoustic manifestation of voice symptoms, such as dysarthria, is very variable, making its detection and classification challenging, both by human specialists and automatic systems. In this context, precision medicine, defined as "prevention and treatment strategies that take individual variability into account", has gained a great interest in the ALS community. Specifically, the use of innovative Artificial Intelligence techniques, such as Machine Learning, plays a pivotal role in finding specific patterns in the data set to help neurologists in clinical decision-making. Therefore, the main objective of this study was to find new markers, and new patterns, to promptly detect the possible presence of dysarthria and to correctly classify its severity. We have performed an acoustic analysis on different voice signals of various degrees of impairment acquired during outpatient visits at the ALS center of the "Federico II" University Hospital. From the collected signals, a new database containing different acoustic parameters was realized, on which several experiments were performed. The study led us to the discovery of markers that helped to develop a decision tree that separated healthy subjects from patients and, among patients, those with different severity of dysarthria. This model achieved good results in terms of dysarthria classification accuracy, 86.6%, which is excellent considering the small number of subjects in the data set.

1. Introduction

ALS is the most common adult-onset neurodegenerative disease of the motor neuron system that affects upper and lower motor neurons located in the cerebral cortex and spinal cord, respectively [1]. This dysfunction leads to progressive weakness of voluntary skeletal muscles involved in limb movement, swallowing (dysphagia), speaking (dysarthria), and respiratory function, with different clinical presentations.

The bulbar presentation, characterized by the progressive decline in swallowing and speech (from dysarthria to anarthria), is a devastating feature of the disease that leads to shorter survival (less than two years from diagnosis) and to reduced quality of life [1]. Indeed, dysarthria may occur in up to 25% of patients at disease onset and may develop in more than 80% during the disease course [2]. Due to the lack of a validated biomarker, dysarthria is usually assessed by clinical examination and using the first item of the ALS functional rating scale-revised (ALSFRS-R) [3], which represents so far the state-of-the-art for the evaluation of speech severity.

ALSFRS-R is based on subjective evaluation of patients' symptoms and therefore lacks reliability and sensitivity for detecting subclinical changes in the bulbar motor system and correctly classifying patients based on clinical severity. These features provide essential information about the pathophysiology of bulbar involvement and take into account the variability of its clinical presentation.

In the last years, precision medicine, defined as "prevention and treatment strategies that take individual variability into account" [4], has gained a great interest in the ALS community [5]. Beyond the personalizing of the treatment for every single patient, it is strongly

* Corresponding author.
E-mail address: raffaele.dubbioso@unina.it (R. Dubbioso).

[1] These authors share the first authorship.

https://doi.org/10.1016/j.bspc.2023.105706
Received 31 May 2023; Received in revised form 9 September 2023; Accepted 5 November 2023
Available online 16 November 2023

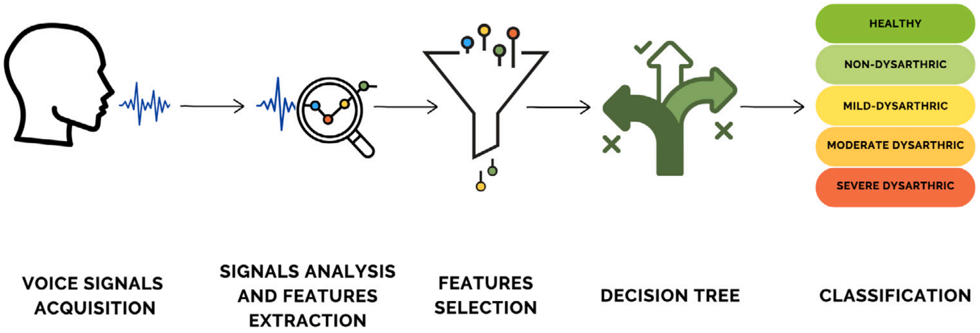

Fig. 1. Overview of the methodology implemented in this study to find new markers and new patterns for the dysarthria severity assessment.

important to the predictive component in the management of a disease, as precisely as possible, especially for neurodegenerative diseases, such as ALS.

Precision medicine aims to detect and treat perturbations in patients long before symptoms appear, thus optimizing the treatment for each patient and promptly acting in order to avoid the decline of the disease. This is possible through extensive biomarker testing, close monitoring, deep statistical analysis, and patient health coaching.

In the case of dysarthria, dedicated voice analysis devices have been used to add value to the diagnosis of ALS by detecting subclinical changes. Specifically, abnormalities of jitter, shimmer, articulatory rate, speaking rate, and pause rate have been described, and these features can be measured from remotely collected speech samples. More recently, two longitudinal studies [6,7] have assessed the sensitivity of remote speech analysis to detect and track early bulbar change over time [6], demonstrating its superiority compared to patient-reported measures [7].

Based on these premises, we herein present preliminary results of a clinical study whose main goal is to find new markers and new patterns to detect dysarthria promptly and correctly classify its severity in ALS patients. To achieve this goal, an acoustic voice analysis has been carried out over different speech signals acquired during outpatient visits at the ALS center of the "Federico II" University Hospital in Naples, Italy. From the collected signals, it has been realized a new database containing different acoustic parameters, on which several experiments have been performed. The study has led us to the discovery of new potential markers which contributed to developing a Decision Tree (DT) to separate correctly healthy subjects from patients and, among patients, those with different severity of dysarthria (i.e., Non-Dysarthric, Mild-Dysarthric, Moderate-Dysarthric, and Severe-Dysarthric).

An overview of the workflow of the methodology implemented in this study to find new markers and new patterns to classify dysarthria is illustrated in Fig. 1.

In summary, the strength of this study is not to simply make a classification but also to carry out an investigation on which are the most distinctive features for the dysarthria severity assessment of ALS patients. In fact, as far as we know, our work represents the first attempt to automatically extract explicit and easy-to-understand knowledge from ALS data that can be provided to physicians in the form of an interpretable classification pattern. In this way, clinicians can understand the motivations underlying the decisions made by the ML tool used. This represents a breakthrough in the ALS field, where, until now, other classifiers, such as Support Vector Machine (SVM) [8] or Deep Neural Networks (DNN) [9], have been used; in fact, the latter behave as "black boxes" meaning that they simply provide the classification without explaining the motivations underlying their decisions.

2. Related work

During the last decades, several automatic detection techniques have been proposed for the identification and classification of dysarthria in patients with ALS by using different acoustic features. In this subset of patients, alterations of values of these parameters can, in fact, indicate possible variations of the normal functioning of the pneumo-articulatory apparatus.

Speech intelligibility and speaking rate as indices of speech function to evaluate the bulbar function and disease severity in ALS patients were, for example, evaluated in [7]. Recordings of each subject's reading of sentences from the Speech intelligibility test (SIT) [10] were processed to estimate these measures. Speech samples obtained from picture descriptions were, instead, analyzed in [11]. In this case, Fundamental Frequency (F_0) range, durations of speech segments, and the number of pauses were assessed to analyze the natural speech in ALS patients. Statistical analyses were conducted in both studies to find the correlation between clinical and acoustic measures.

Recently, several studies have demonstrated the reliability of different ML approaches in assessing dysarthric disorder in patients suffering from ALS. In [12], for example, 70 acoustic parameters, including entropy and spectral measurements, Kurtosis, and rhythm variability, from the five vowels were estimated to detect possible voice alterations. Participants were divided into control subjects and ALS patients suffering from bulbar or nonbulbar dysfunction. The performance of several ML models to correctly classify voice samples was evaluated, obtaining the best results with Random Forest (RF) model when classifying bulbar vs. control participants (accuracy of 88.3%). While the SVM constitutes the best reliable ML model to distinguish bulbar and no-bulbar patients (accuracy of 91.0%).

Signals of five vowels were also processed in a previous study by these authors [13]. In this case, only jitter, shimmer, Harmonic to Noise Ratio (HNR), and pitch were estimated. An SVM model was used to classify bulbar and control participants, obtaining an accuracy equal to 95.8%. While the RF represents the best model to classify bulbar and no-bulbar patients, achieving an accuracy of 75.5%

Jitter, shimmer, HNR, and other acoustic parameters, such as Directional Perturbation Factor (DFP), Glottal-to Noise Excitation ratio (GNE), Mel-Frequency Cepstral Coefficients (MFCC), formants, F_0 contour-based parameters (Phonatory frequency range (PFR), Pitch period entropy (PPE), voice tremor, and harmonic structure of the vowels analysis) were, instead, considered to evaluate the speech progress of patients suffering from ALS in [14]. These parameters were extracted from only two vowels, /a/ and /i/, considered the vowels capable of providing the most relevant information suitable for a high-performance classifier. A Linear Discriminant Analysis (LDA) was used to distinguish healthy and subjects suffering from ALS, achieving an accuracy of 99.7% based on 32 features picked out by the Least Absolute Shrinkage and Selection Operator (LASSO) feature selection algorithm.

An LDA was also adopted to classify healthy or pathological subjects in [15]. Jitter, shimmer, and Pathological Vibrato Index (PVI) were estimated from vowel /a/ samples. An accuracy of 90.7 ± 1.7 was achieved considering only shimmer and PVI. While an accuracy of 91.6 ± 2.3 was obtained considering only jitter and PVI and classifying with the k-Nearest Neighbors (kNN). Vowels, fricatives, monosyllabic targets, and monologue were, instead, processed in [16]. MFCC were extracted. Accuracies equal to 80.2%, 89.3%, and 83.2% were, respectively, achieved considering only vowels and fricatives, monosyllabic or monologue tasks. In the first two cases, SVM was used, while a DNN was adopted when monologue was considered.

A sentence was, instead, processed in [17]. Features extracted from openSMILE (SMILE is an acronym for Speech and Music Interpretation by Large-space Extraction) [18] toolkit constitute inputs of SVM, achieving Area under the Curve (AUC) values equal to 0.99 ± 0.00 and 0.91 ± 0.15, respectively, for female and male subjects selecting appropriate features and classes.

Table 1 reports the different ML approaches outlined so far, also indicating, for each of them, the acoustic features, voice samples, and the best results obtained. The column "ML model" of Table 1 shows that, to perform classification on data related to ALS, most studies existing in the current literature only report the use of "black box" algorithms. This evidences the novelty of our approach, in which explicit knowledge is obtained and provided to physicians who can take advantage of it.

3. Data acquisition

To perform the identification of new potential acoustic markers for ALS dysarthria severity assessment, it was necessary to conduct a clinical study to collect data. This is because, for ALS, the lack of proper datasets is a critical issue. In fact, although several scientific initiatives have resulted in the growing availability of datasets that can be freely used, there is still a need for larger, more complete biomedical datasets that are more representative of specific populations and, especially, of different severity of diseases.

To overcome this problem, a clinical protocol has been designed and well formalized in a document approved by the Ethics Committee of the University of Naples "Federico II" (Protocol ID 100/17/ES01).

This clinical study aimed to collect data useful for creating healthcare models and tools to support the screening for detecting dysarthria (and/or other voice alterations) in patients with ALS. The experimental protocol clearly and explicitly describes the goals, the drawing, the methodology, the statistical considerations, and the overall organization of the performed study. For the drafting of the protocol, reference was made to the Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) 2013 standard [19], which includes several sections, administrative, introductory, methodological, ethics and disclosure, and appendix. Each section defines information, data, methodologies, and requirements, in form and structure, useful for conducting the study.

For the sake of brevity, only a brief description of the main sections of the protocol is given below.

3.1. Study design and settings

The study was conducted at the ALS center of the "Federico II" University Hospital of Naples in the period between 01/01/2022 and 12/31/2022. All subjects gave written informed consent before participation.

Subjects eligible for the study met the following inclusion criteria:

- Italian native speakers aged between 18 and 80 years;
- Subjects able to comply with the study visit schedule and other protocol requirements;
- Healthy subjects recruited among the patients' caregivers
- ALS patients with a "probable", "probable laboratory-supported", or "definite" diagnosis per the revised El Escorial criteria [20].

Instead, subjects meeting the following exclusion criteria were excluded:

- Individuals under the age of 18 and over the age of 80;
- Individuals with illnesses such as colds or upper respiratory infections;
- Individuals with other neurological disorders that may affect voice or speech (e.g., Alzheimer's disease, stroke, Parkinson's disease).
- Patients that scored 0 at the item 1 of ALSFRS-R, namely patients with loss of useful speech.

3.2. Procedures

The procedures performed for carrying out the study were divided into the following phases:

- Subject Recruitment: Promotion of the study to recruit participants;
- Information: during which the aims of the study are explained. At the end of this meeting, the interested parties are presented with the documents to be signed (information sheet and informed consent given in the appendix);
- Registration of selected participants;
- Medical examination of the enrolled subjects and compilation of the medical history form;
- Execution of tests: the participant carries out the tests defined for the data collection.

The acquisition activity was carefully prepared, setting up the environmental conditions that facilitate its success, such as a quiet environment in the vocalization recording phase necessary for the acoustic analysis, tools and procedures as simple and efficient as possible, and timely mechanisms for troubleshooting and recovering information that would otherwise be lost.

Specifically, tests were performed on subjects "in resting conditions" and in the absence of emotional and physiological stress and were carried out in a silent environment (<30 dB of background noise) and not too dry (humidity rate above 35%–40%).

Below are listed the tasks requested for each participant:

1. Recording of vocalization of at least five seconds of the vowels /a/, /e/, /i/, /o/, and /u/, one for each vowel, without interruption of loudness, positioning the microphone at a distance of approx. 20 cm from the lips, with an angle of about 45° [21,22];
2. Recording of the patient's voice while repeating the syllables /pa/, /ta/, /ka/ as fast as possible in a single breath in three different audio files [21,22];
3. Recording the patient's voice pronouncing the days of the week starting from Monday [21];
4. Recording of the patient's voice while reading the passage "The North Wind" [21];
5. Recording of the patient's voice while describing a picture for at least 60 s [23];
6. Saving the result obtained from the self-assessment questionnaire.

Audio recordings were acquired by the investigators using a Samsung Galaxy S8+ SM-G955F with Android version 9.0 operating system, on which a new customized version of our developed Vox4Health app was installed [24,25].

3.3. Enrolled participants

A total of 1067 voice signals were collected from 97 participants who have recorded 11 types of tasks: the vocalization of the vowels /a/, /e/, /i/, /o/ and /u/, the repetition of the syllables /pa/, /ta/, /ka/, the pronunciation of the days of the week in order starting from Monday, the reading of the passage "The North Wind", and the description of a picture.

Table 1
Overview of research studies on ALS classification by using ML models.

Reference	Year of publication	Voice samples	Acoustic parameters	ML model	Results
[12]	2023	Vowels /a/, /e/, /i/, /o/, /u/	Entropy measures	RF	Accuracy = 88.3%
			Variance of a signal Kurtosis Rhythm variability (mean and standard deviation)	*(when classifying bulbar vs. control participants)*	Sensitivity = 85.0% Specificity = 95.0%
			Mean frequency of the probability density function	SVM	Accuracy = 91.0%
			Average spectral energy	*(when classifying bulbar vs. non-bulbar participants)*	Sensitivity = 83.3% Specificity = 100.0%
[13]	2021	Vowels /a/, /e/, /i/, /o/, /u/	Jitter	SVM	Accuracy = 95.8%
			shimmer HNR pitch	*(when classifying bulbar vs. control participants)*	Sensitivity = 91.4% Specificity = 99.3%
				Neural Network (NN) *(when classifying no-bulbar vs. control participants)*	Accuracy = 92.5% Sensitivity = 90.3% Specificity = 96.4%
				RF *(when classifying bulbar vs. non-bulbar participants)*	Accuracy = 75.5% Sensitivity = 55.7% Specificity = 88.4%
				NN *(when classifying ALS vs. control participants)*	Accuracy = 92.2% Sensitivity = 90.8% Specificity = 95.6%
[14]	2021	Vowels /a/, /i/	jitter Shimmer HNR DFP GNE MFCC Formants F_0 contour based parameters	LDA	Accuracy = 99.7% Sensitivity = 99.3% Specificity = 99.9%
[15]	2019	Vowel /a/	Jitter Shimmer PVI	LDA *(obtained with shimmer and PVI)*	Averaged Recall = 89.5% Accuracy = 90.7 ± 1.7% Sensitivity = 86.7 ± 0.1% Specificity = 92.2 ± 2.3%
				kNN *(obtained with jitter and PVI)*	Averaged Recall = 86.9% Accuracy = 91.6 ± 2.3% Sensitivity = 76.3 ± 5.8% Specificity = 97.5 ± 1.7%
[16]	2019	Vowels /a/, /e/, /i/, /o/, /u/ fricatives /s/, /sh/, and /f/ monosyllabic targets /pa/,/ta/,/ka/ monologue	MFCC	SVM *for vowels and fricatives tasks*	Average accuracy = 80.2%
				SVM *for /pa/, /ta/, /ka/ tasks*	Average accuracy = 89.3%
				DNN *for monologue tasks*	Average accuracy = 83.2%
[17]	2019	sentence	OpenSMILE features	SVM *(achieved by using the 64 best features for symptomatic vs control female subjects)*	AUC = 0.99 ± 0.00
				(achieved by using the MFCC features for all ALS vs control male subjects)	AUC = 0.91 ± 0.15

There is a prevalence of pathological voices compared to healthy ones, the former numbering 74 (45 male and 29 female), the latter 23 (8 male and 15 female). In Table 2, details about the number of participants, distinguishing between healthy and pathological subjects, are provided. In particular, for each category, we have indicated the number and percentage of female and male subjects involved in the study, and for the patients, we have provided the numbers and the percentage for each severity class of dysarthria.

4. Database creation

4.1. Acoustic features

Speech is the result of the interaction of several vocal subsystems, such as phonation, resonance, articulation, and respiration. The purpose of the acoustic analysis is to assess the proper functioning of these subsystems, the alteration of which could be caused by a specific pathology such as dysarthria. The analysis of different speech tasks

Table 2
Description of the dataset: it is reported the number and percentage (with respect to the total of 97 subjects) of female and male participants involved in the study, and for the patients, it is also provided with the numbers and the percentage for each severity class of dysarthria. To determine the level of dysarthria of each patient, the ALSFRS-R scale for the clinical classification of dysarthria [3] has been used.

	Female	% respect to the total (97)	Male	% respect to the total (97)	Total
Patients	**29**	**30%**	**45**	**46%**	**74**
Level of Dysarthria					
Non-Dysarthric	17	18%	26	27%	43
Mild-Dysarthric	8	8%	15	15%	23
Moderate-Dysarthric	2	2%	2	2%	4
Severe-Dysarthric	2	2%	2	2%	4
Healthy Subjects	**15**	**15%**	**8**	**8%**	**23**
Total	**44**	**45%**	**53**	**55%**	**97**

Table 3
List of acoustic features extracted and used in this study.

Feature	Acronym	Reference
Degree of Vocal Arrests	DVA	[29]
Standard deviation of the Power Spectral Density	stdPSD	[27]
Maximum Phonation Time	MPT	[27]
Standard deviation of Fundamental frequency	stdF0	[27]
Jitter	Jitter	[30]
Shimmer	Shimmer	[30]
Harmonics-to-noise ratio	HNR	[30]
Proportion of Subharmonic Intervals	PSI	[27]
Proportion of Fundamental Frequency Tremor	PF0T	[31]
Proportion of Amplitude Tremor	PAT	[31]
Degree of hypernasality	EFn-M	[27]
Intermittent hypernasality	EFn-SD	[27]
Rhythm Acceleration	RA	[27]
Rhythm Instability	RI	[27]
Rate of Speech Timing	RST	[27]
Duration of Pause Intervals	DPI	[27]
Standard Deviation of Power	stdPWR	[27]
Voice onset time	VOT	[27]
Diadochokinetic Rate	DDKR	[27]
Diadochokinetic Irregularity	DDKI	[27]
Vowel Duration	VD	[27]
Net Speech Rate	NSR	[27]

Table 4
The acoustic features extracted for each speech signal.

Signals/Tasks	Acoustic features extracted
Vowels /a/, /e/, /i/, /o/, and /u/	DVA, stdPSD, MPT, stdF0, jitter, shimmer, HNR, PSI, PF0T, PAT, EFn-M, EFn-SD
Syllables /pa/, /ta/, /ka/	RA, RI
Days of the week	VOT, DDKR, DDKI, VD, stdPWR
Reading of the passage “The North Wind”	VOT, DDKR, DDKI, VD, stdPWR, RST, DPI, stdF0, NSR
Description of a picture	RST, DPI, stdPWR, stdF0

can highlight several speech aspects. Rhythm stability and acceleration can be assessed, for example, using the syllables /pa/, /ta/, and /ka/. Instead, individual words or sentences can be used to estimate the quality of articulation. At the same time, sustained vowels are adopted to measure phonatory characteristics. Finally, the reading of a specific text and the description of a picture on a given topic are used to assess the connected speech, highlighting the cooperation between all the subsystems of speech. In this study, several acoustic features are extracted from the collected speech signals by using the *Dysarthria Analyzer*, a software written in Matlab and tested in several clinical settings [26,27], after being appropriately filtered with a Butterworth band stop filter [28]. Table 3 reports all the features calculated for the voice signals acquired in this study, with their acronyms and the references to which it is possible to find more information for each of them.

4.2. Data set

The acoustic features, described in Section 4.1 and summarized in Table 3, were extracted from the different speech tasks collected during the clinical study. To estimate the most significant voice quality characteristics, specific features were computed for specific tasks. Table 4 details the acoustic features extracted for each speech signal.

Since we have recruited 97 subjects, and, for each subject, we have collected 11 voice signals, each of them has been processed and from which we have extracted a number of characteristics as just described before, we have obtained a data set in which each item is composed as follows:

- Subject ID;
- 5 (vowels) * 12 features + 3 (syllables) * 2 features, 1 (days of the week) * 5 features + 1 (reading) * 9 features + 1 (picture) * 4 features = 60+6+5+9+4 = 84 features;
- class (Healthy, Non-Dysarthric, Mild-Dysarthric, Moderate-Dysarthric, and Severe-Dysarthric). We used item 1 of the ALSFRS-R for the clinical classification of dysarthria [3]. ALSFRS-R includes 12 questions that can have a score of 0 to 4. A score of 0 on a question would indicate no function, while a score of 4 would indicate full function. Questions 1 to 3 are related to bulbar function (speech, salivation, and swallowing), questions 4 to 9 are related to limb function, and questions 10–12 are related to respiratory function. Specifically, in the speech item 1 no-dysarthric patients corresponded to normal speech process (score 4), mild-dysarthric indicated detectable speech disturbance (score 3), moderate corresponded to intelligible speech with repeating (score 2), severe-dysarthric indicated speech combined with nonvocal communication (score 1). Patients that scored 0, namely with loss of useful speech, were not included in the study.

5. Experiments

All the experiments reported in the current section have been carried out by using the Waikato Environment for Knowledge Analysis (WEKA) [32] tool, version 3.8.1. It contains a wide set of classifiers that can be subdivided into groups on the basis of the working mechanisms they use to perform classification. Given that our goal is not only to classify but also to achieve explicit information on the reasons for which a subject is assigned to a given class, we have only considered the groups that contain classifiers based either on explicit rules or on decision trees. Other classifiers, e.g., Artificial Neural Networks (ANNs) [33], SVM [8], or Radial Basis Functions (RBF) [34], although capable of good classification performance, behave as “black boxes”: they do not explain the motivations for their decisions; hence, they are not interesting for our purposes and will not be used throughout this paper.

Table 5 reports, for each such classifier, the class it belongs to, the complete name, the acronym used within the current paper, and the reference to its seminal paper.

Table 5
The classification algorithms used in the present paper.

Class	Algorithm	Acronym	Reference
Rules			
	JRip	JR	[35]
	One Rule	OR	[36]
	PART	PA	[37]
	Ridor	RI	[38]
Trees			
	J48	J4	[39]
	Random Forest	RF	[40]
	Random Tree	RT	[40]
	REPTree	RE	[41]

The decision has been made to carry out supervised learning by means of a division of the items into two sets, a training set and a testing one. The former is shown to a classifier so that this can perform learning on this data. The knowledge acquired is then tested on the items of the latter that have not been previously shown to the classifier. We have decided to use 66% of the items to perform training and the remaining 34% for testing. This results in 64 and 33 items, respectively. The decision has been taken to use for the training set the first 64 items in their sequential order of appearance in the data set.

As concerns the evaluation of the performance of each algorithm, given that the data set is highly unbalanced in terms of items belonging to the different classes, the choice of the accuracy metric would not be the most suitable one. In fact, accuracy would hide the fact that items of the minority classes are misclassified. In these cases, instead, it is preferable to use indicators such as the F_1 score, the Matthews Correlation Coefficient, or Cohen's Kappa coefficient. In this paper, we have chosen to utilize the widely used Kappa coefficient [42], denoted as κ. In classification, κ measures the agreement between the classification obtained and the truth values.

For two classes, called, respectively, the positive and the negative, κ is defined as:

$$\kappa = \frac{2 \cdot (TP \cdot TN - FN \cdot FP)}{(TP + FP) \cdot (FP + TN) + (TP + FN) \cdot (FN + TN)} \quad (1)$$

where TP are the true positives, FP are the false positives, TN are the true negatives, and FN are the false negatives.

When there are more than two classes, instead, the generalization of κ takes place through the computation by WEKA of the weighted κ ($w\kappa$). This is defined as the weighted average of the κ values computed for each class, where the weights are given by the percentages of the items belonging to that class with respect to the total number of items.

Both κ and $w\kappa$ normally achieve values in the range [0.0–1.0], where 1.0 represents perfect agreement between classification and truth values, while 0.0 represents no agreement, and the higher the value, the better the classification. In some particularly negative situations, nonetheless, the values can also be negative.

For each classification algorithm, no preliminary tuning phase of its parameters has been effected; rather, the default setting present in WEKA has been utilized throughout the experiments.

It should be remarked here that the execution of each of these algorithms depends on the use of an initial random seed: different seeds could yield different classifications, hence different performances. To get rid of this problem, in this paper, each algorithm is run 25 times, which provides 25 $w\kappa$ values, and the corresponding average and standard deviation are considered.

Within this general framework, several experiments have been performed, which are reported in the following subsections.

5.1. The original data set

The first experiment consisted in the use of the data set "as it is": each of the above classifiers has been used on it. The results are reported in Table 6. For each algorithm, the table reports the average of the 25 values of the $w\kappa$ score over the 25 runs, the standard deviation, the best value obtained, and the worst value achieved. For each statistical indicator, the best result is shown in bold.

The classification performance is quite poor for all the algorithms investigated, and in particular, OneRule, J48, RandomTree, and Reduced Error Pruning Tree (REPTree) show low classification ability. On average, the best-performing algorithm is Ridor, with an average value of 0.3520, whereas the best single-run performance has been obtained by JRip with 0.4807. Some algorithms obtain the same performance over the 25 runs independently of the random seed; for them, the related standard deviation is equal to 0.

This far-from-excellent classification performance should be expected, because the data set has some features that usually impact the classification quality:

- the data set only contains a small number of subjects. This is an intrinsic limitation; some techniques could help increase the number of items by adding some artificial ones. Unfortunately, none of them helped us to improve accuracy;
- the attributes are very numerous. In this case, *feature selection* can help;
- the classes are very unbalanced in terms of the number of items assigned to them. A first attempt has been the use of mechanisms such as the Synthetic Minority Oversampling TEchnique (SMOTE), unfortunately without any appreciable result. A further idea consists in a reduction in the number of classes that could merge the least populated ones into other more populated ones.

In the following two subsections, we will take into account feature selection and class reduction, respectively.

5.2. Feature selection

WEKA also contains ten algorithms that automatically perform feature selection. They are based on different ideas, ranging from correlation statistics to information gain to wrapper techniques.

The application of each of them on the original data set has provided a set of data set attributes that are considered the most significant for a good division into classes. These sets are different from one algorithm to another.

As a summary of this step, Table 7 reports the names of the feature selectors we have used, together with an acronym, a literature reference for each of them, and, most importantly, the number of features selected by each of them.

ClassifierAttributeEval and ClassifierSubsetEval selected all the data set attributes, meaning that they actually failed to select the most significant ones. No data set attribute was selected by all the eight remaining algorithms, and a total of 28 attributes were selected at least once. Table 8 shows these 28 attributes and the number of times each of them has been selected. In the table, for the sake of space, the columns show, respectively, the number of times attributes are selected, the number of attributes selected that number of times, and their names. Features are grouped in terms of the number of algorithms that select them.

As a general comment to the table, the attributes derived from the vowels, although being the vast majority of the data set, are seldom chosen by many algorithms, apart from some pertaining to the vowel /a/ and, to a lesser extent, to /i/; this is in accordance with what is reported in [14]. Particularly, all those related to the vowel /e/ are always discarded as not relevant. Many parameters related to the reading, instead, are very relevant.

Now, a decision must be made on the number of selected features that we have to use. On the one hand, this number should be as low as possible so as to only keep the most discriminant features, thereby favoring an as-crisp-as-possible separation among the classes. On the other hand, this number should be as high as possible to take into

Table 6
The results in terms of weighted Kappa coefficient ($w\kappa$) over the test set achieved on the original data set. The best results are shown in bold.

	JR	OR	PA	RI	J4	RF	RT	RE
Average	0.288	0.083	0.338	**0.352**	0.146	0.264	0.118	0.133
Std.dev	0.104	**0.000**	**0.000**	**0.000**	**0.000**	0.071	0.117	0.112
Best	**0.481**	0.083	0.338	0.352	0.146	0.386	0.332	0.377
Worst	0.092	0.083	0.338	**0.352**	0.146	0.121	−0.192	−0.082

Table 7
The feature selection algorithms used in the present paper.

Algorithm	Acronym	Reference	Selected
CfsSubsetEval	CfsSE	[43]	11
ClassifierAttributeEval	CAE	[32]	85
ClassifierSubsetEval	ClSE	[44]	85
CorrelationAttributeEval	CoAE	[32]	9
GainRatioAttributeEval	GRAE	[45]	19
InfoGainAttributeEval	IGAE	[45]	19
OneRAttributeEval	ORAE	[32]	14
ReliefFAttributeEval	RAE	[46]	8
SymmetricalUncertAttributeEval	SUAE	[32]	19
WrapperSubsetEval	WSE	[47]	8

Table 8
The features selected and the number of algorithms selecting them.

Number of times	Number of attributes	Attribute names
8	0	
7	3	PAT (A); DPI (picture); NSR (reading);
6	4	DDKR (reading); DPI (reading); VD (reading); RST (reading);
5	5	PF0T (A); stdPSD (I);PF0T (U); RST (picture); DDKI (reading);
4	5	stdF0 (A); PF0T(I); EFn_SD (I); MPT (I); EFn_M (U);
3	3	stdF0(I); stdF0(O); MPT (U);
2	0	
1	8	PSI (I); PF0T (O); PSI (U); RA (PA); stdF0 (picture); stdPWR(reading); DDKR(days); stdPWR(days)

account the outcome of as many feature selectors as possible, under the hypothesis that some interesting feature has been considered by a few selectors only.

To make a decision, six data sets have been extracted from the original one, each of which contains all the features selected at least 7, 6, 5, 4, 3, and one time. Consequently, each of them consists of 3, 7 (3+4), 12 (3+4+5), 17 (3+4+5+5), 20 (3+4+5+5+3), and 28 (3+4+5+5+3+8) attributes. Then, the classifiers have been run on these reduced data sets. The presentation of all the results would take too much space here, yet we have noticed that there is no large difference between the results when three and seven features are used, while adding five further features does not contribute to improving the results.

Therefore, we have chosen to consider the seven features picked either six or seven times by the feature selectors. It is worth mentioning that five out of these seven make reference to the reading, one to the picture, and one to the vowel /a/.

With this reduced data set, we now repeat the experiments reported in Table 6.

The corresponding results are contained in Table 9.

Table 10 shows the percent variation in the performance of the different algorithms when passing from the original data set to the seven-feature reduced one. The last column shows the average values over the eight classifiers considered.

As it can be appreciated, most classifiers have improved their performance when using the reduced data set, whereas some others have worsened. Yet, the last column shows that, on average, the weighted Kappa score $w\kappa$ has improved by more than 8%, the standard deviation has decreased by almost 3%, the best solution has increased by almost 2%, and the worst solution has largely improved by more than 27%. All of these are positive features confirming that data set reduction has a positive effect on classification quality.

5.3. Class reduction

To suitably tackle this unbalanced multi-class data set, we have decided to proceed step by step: in each such step, we contrast one data set class against another one obtained by merging some of the other classes. This yields to perform a set of four binary classification tasks. Namely, the following steps have been taken:

- healthy versus non-healthy: in this latter, we merge all the classes related to the different severity of dysarthria, i.e., non-dysarthric, mild, moderate, and severe-dysarthric;
- non-dysarthric versus dysarthric: in this latter, mild, moderate, and severe-dysarthric are merged;
- mild dysarthric versus non-mild-dysarthric: in this latter, moderate and severe-dysarthric are merged;
- moderate-dysarthric versus severe-dysarthric.

For each of these steps, we have run the experiments in exactly the same way as described above.

For each such two-class task, we should report here the tables containing the results achieved by the various algorithms in terms of the weighted Kappa coefficient and the best classification rules obtained; unfortunately, the related experiments are too lengthy to be reported here. Therefore, hereinafter, we will only provide readers with the final results.

At the end of each step, the best classification rule obtained has been stored. Each of them is very simple, consisting of just one or at most two out of the seven selected features, and is highly discriminating. If we put them all together in the same order of the four effected steps, we can obtain the easy-to-understand classification tree reported in Algorithm 1.

Algorithm 1 The classification tree obtained

```
IF
  ((DPI (picture) $\leq$ 260.826923) AND (RST (reading) $\geq$ 401.817056))
    THEN
      class = healthy
    ELSE
      IF
        (DPI (reading) $\leq$ 236.5442765)
          THEN
            class = non-dysarthric
          ELSE
            IF
              (DPI (reading) $\leq$ 424.730254)
                THEN
                  class = mild-dysarthric
                ELSE
                  IF
                    (PAT (A) $\leq$ 6.715932)
                      THEN
                        class = moderate-dysarthric
                      ELSE
                        class = severe-dysarthric
```

If we apply this decision tree to the original data set, we obtain the confusion matrix shown in Table 11.

Table 9
The results in terms of weighted Kappa coefficient w_K over the test set achieved on the seven-feature data set. The best results are shown in bold.

	JR	OR	PA	RI	J4	RF	RT	RE
Average	0.243	−0.049	0.162	0.221	**0.392**	0.333	0.249	0.164
Std. dev	0.101	**0.000**	**0.000**	**0.000**	**0.000**	0.043	0.111	0.140
Best	0.416	−0.049	0.162	0.221	0.392	0.422	**0.555**	0.496
Worst	0.006	−0.049	0.162	0.221	**0.392**	0.245	0.092	−0.079

Table 10
Percent variation in the performance of the different algorithms when passing from the original data set to the seven–feature reduced one. The best results are shown in bold.

	JR	OR	PA	RI	J4	RF	RT	RE	Average
Average	−15.601	−158.944	−52.159	−37.273	**169.136**	26.119	111.657	23.131	8.258
Std.dev.	−2.979	0.000	0.000	0.000	0.000	**−38.981**	−5.423	25.535	−2.731
Best	−13.480	−158.944	−52.159	−37.273	**169.136**	9.216	67.078	31.687	1.908
Worst	−93.492	−158.944	−52.159	−37.273	**169.136**	102.819	−148.096	−3.659	−27.708

Table 11
The confusion matrix obtained using the decision tree reported in Algorithm 1. Here, class c1 represents the severe-dysarthric subjects, c2 the moderate-dysarthric ones, c3 the mild-dysarthric, c4 the non-dysartric, and class c5 contains the healthy subjects.

		Predicted Class				
		c1	***c2***	***c3***	***c4***	***c5***
Real Class	***c1***	4	0	0	0	0
	c2	0	4	0	0	0
	c3	0	0	23	0	0
	c4	1	1	4	37	0
	c5	0	0	1	6	16

This is an excellent situation where just 13 subjects are misclassified. Even more interestingly, almost all the items lie on the main diagonal, i.e., they are correctly assigned, and almost all the misclassified items belong to a neighboring class. This corresponds to a classification in which just 13 subjects out of the 97 are wrongly classified and to a weighted Kappa value w_K of 0.8060 (the accuracy being equal to 86.598%), which is excellent due to the data set limitations exposed in the early parts of this section. This value of w_K is much higher than those shown in Tables 6 and 9, where the best values obtained were equal to 0.4807 and 0.5547, respectively. This proves the goodness of the approach followed.

6. Discussion

In this preliminary work, we carried out a clinical study to collect new data useful to identify novel acoustic markers. These markers contributed to realize a decision tree that successfully stratified dysarthria severity, particularly for patients with more severe speech deficits.

As a general comment, the features derived from the vowels /a/ and, to a lesser extent, /i/ resulted in being very informative and in line with what is reported in the literature [14]. Notably, the vowel /a/ was the most sensitive parameter to discriminate ALS patients belonging to the most severe dysarthria categories. On the other hand, many parameters related to the reading and picture description, instead, were very relevant to separate healthy subjects from ALS patients and, among patients, those without dysarthria vs mild-dysarthria.

Examining the confusion matrix from a medical viewpoint, it was evident that the subjects with the three most compromised levels of dysarthria, i.e., those belonging to severe, moderate, and mild-dysarthric classes, were correctly classified. Regarding the non-dysarthric subjects, six were not correctly classified; yet, four of them were assigned to the neighboring class mild-dysarthric, and just two were assigned to the more severe classes. Finally, for the healthy control subjects, seven were classified as diseased, yet, six of them were seen as non-dysarthric, i.e., the less severe level, and just one was seen as mild-dysarthric.

Additionally, it should be remarked that we were able to obtain a decision tree that can be proposed to physicians: it is easy to understand and provides clear indications on which features to check to assess the degree of dysarthria severity.

Specifically, we created an interpretable classification pattern that allowed us to interpret the learned rules to gain insight into the problem. This approach differs from previous studies on the topic where the authors used traditional classifiers, such as SVMs or DNN, known to act as 'black boxes'. Furthermore, these studies did not clearly stratify patients according to the clinical severity of dysarthria, making their results poorly reliable and feasible.

Therefore, we believe that our decision tree might be very useful in a clinical setting since it gives objective measures that can be used to provide valuable information about disease progression, determine enrollment, stratify participants, and appropriately power a study.

Lastly, we acknowledge some limitations of our study. First, the total number of patients and the number in each group are small and not perfectly balanced. This issue is related to the longitudinal nature of our study, as patients were consecutively recruited in the outpatient clinic. In addition, ALS is considered a rare disease, and therefore our sample is perfectly in line with previous literature on the topic. Second, the sex and age distribution of healthy subjects does not match that of patients. However, this problem is also frequent in previous studies, as very often healthy subjects are recruited among caregivers.

7. Conclusions and outlook

In the last years, precision medicine has attracted great interest in the ALS community, especially toward drug development, as many failures in translation can be attributed in part to disease heterogeneity in humans.

In this paper, we developed an interpretable decision tree model capable of separating healthy and pathological subjects and distinguishing different severity of dysarthria in ALS patients. Interestingly, our model suggests for future studies the use of tasks such as reading or monologue to screen patients with mild dysarthria, whereas more conventional tests such as vocalization could be very informative for patients with severe dysarthria.

Therefore, easy-to-understand knowledge and interpretability of this model can constitute valid support to ALS clinicians, which is critical for determining the appropriate timing of interventions, providing counseling for patients, and evaluating functional changes during clinical trials.

In addition, this study paves the way for the future development of a longitudinal evaluation of the features contained in the proposed decision tree in order to get possible prognostic biomarkers. Future studies should also take into account more balanced healthy control and patient groups by increasing, for instance, the sample size through augmentation techniques or alternatively considering multicentre studies.

Finally, the inclusion of a neurological control population with dysarthria (i.e., Parkinson's disease, dementia, stroke) would be desirable to gain insight into the sensitivity and specificity of our speech analysis.

Fundings

The authors received no financial support for the research, authorship, and/or publication of this article.

CRediT authorship contribution statement

Raffaele Dubbioso: Conceptualization, Methodology, Resources, Data curation, Writing – review & editing, Supervision. **Myriam Spisto:** Conceptualization, Methodology, Data curation, Writing – review & editing. **Laura Verde:** Conceptualization, Software, Data curation, Writing – original draft, Writing – review & editing. **Valentina Virginia Iuzzolino:** Data curation, Writing – review & editing. **Gianmaria Senerchia:** Data curation, Writing – review & editing. **Giuseppe De Pietro:** Resources, Writing – review & editing. **Ivanoe De Falco:** Methodology, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Giovanna Sannino:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Supervision.

Declaration of competing interest

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.

References

[1] E.L. Feldman, S.A. Goutman, S. Petri, L. Mazzini, M.G. Savelieff, P.J. Shaw, G. Sobue, Amyotrophic lateral sclerosis, Lancet (2022).

[2] J.R. Duffy, Motor Speech Disorders E-Book: Substrates, Differential Diagnosis, and Management, Elsevier Health Sciences, 2019.

[3] J.M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, A. Nakanishi, B.A.S. Group, A. complete listing of the BDNF Study Group, et al., The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function, J. the Neurol. Sci. 169 (1–2) (1999) 13–21.

[4] F.S. Collins, H. Varmus, A new initiative on precision medicine, New England J. Med. 372 (9) (2015) 793–795.

[5] R. McFarlane, M. Galvin, M. Heverin, É. Mac Domhnaill, D. Murray, D. Meldrum, P. Bede, A. Bolger, L. Hederman, S. Impey, et al., PRECISION ALS—an integrated pan European patient data platform for ALS, in: Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, Taylor & Francis, 2023, pp. 1–5.

[6] G.M. Stegmann, S. Hahn, J. Liss, J. Shefner, S. Rutkove, K. Shelton, C.J. Duncan, V. Berisha, Early detection and tracking of bulbar changes in ALS via frequent and remote speech analysis, NPJ Digit. Med. 3 (1) (2020) 132.

[7] M. Eshghi, Y. Yunusova, K.P. Connaghan, B.J. Perry, M.F. Maffei, J.D. Berry, L. Zinman, S. Kalra, L. Korngut, A. Genge, et al., Rate of speech decline in individuals with amyotrophic lateral sclerosis, Sci. Rep. 12 (1) (2022) 15713.

[8] Z.-Q. Zeng, H.-B. Yu, H.-R. Xu, Y.-Q. Xie, J. Gao, Fast training support vector machines using parallel sequential minimal optimization, in: 2008 3rd International Conference on Intelligent System and Knowledge Engineering, Vol. 1, IEEE, 2008, pp. 997–1001.

[9] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Netw. 61 (2015) 85–117.

[10] M. Dorsey, K. Yorkston, D. Beukelman, M. Hakel, Speech Intelligibility Test for Windows, Institute for Rehabilitation Science and Engineering at Madonna, 2007.

[11] N. Nevler, S. Ash, C. McMillan, L. Elman, L. McCluskey, D.J. Irwin, S. Cho, M. Liberman, M. Grossman, Automated analysis of natural speech in amyotrophic lateral sclerosis spectrum disorders, Neurology 95 (12) (2020) e1629–e1639.

[12] A. Tena, F. Clarià, F. Solsona, M. Povedano, Voiceprint and machine learning models for early detection of bulbar dysfunction in ALS, Comput. Methods Programs Biomed. 229 (2023) 107309.

[13] A. Tena, F. Claria, F. Solsona, E. Meister, M. Povedano, et al., Detection of bulbar involvement in patients with amyotrophic lateral sclerosis by machine learning voice analysis: diagnostic decision support development study, JMIR Med. Inf. 9 (3) (2021) e21331.

[14] M. Vashkevich, Y. Rushkevich, Classification of ALS patients based on acoustic analysis of sustained vowel phonations, Biomed. Signal Process. Control 65 (2021) 102350.

[15] M. Vashkevich, A. Petrovsky, Y. Rushkevich, Bulbar ALS detection based on analysis of voice perturbation and vibrato, in: 2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications, SPA, IEEE, 2019, pp. 267–272.

[16] B. Suhas, D. Patel, N.R. Koluguri, Y. Belur, P. Reddy, A. Nalini, R. Yadav, D. Gope, P.K. Ghosh, Comparison of speech tasks and recording devices for voice based automatic classification of healthy subjects and patients with amyotrophic lateral sclerosis, in: INTERSPEECH, 2019, pp. 4564–4568.

[17] S.E. Gutz, J. Wang, Y. Yunusova, J.R. Green, Early identification of speech changes due to amyotrophic lateral sclerosis using machine classification, in: Interspeech, 2019, pp. 604–608.

[18] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM International Conference on Multimedia, 2010, pp. 1459–1462.

[19] A.-W. Chan, J.M. Tetzlaff, D.G. Altman, A. Laupacis, P.C. Gøtzsche, K. Krleža-Jerić, A. Hróbjartsson, H. Mann, K. Dickersin, J.A. Berlin, et al., SPIRIT 2013 statement: Defining standard protocol items for clinical trials, Ann. Intern. Med. 158 (3) (2013) 200–207.

[20] B.R. Brooks, R.G. Miller, M. Swash, T.L. Munsat, El escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis, Amyotroph. Lateral Scler. Other Motor Neuron Disord. 1 (5) (2000) 293–299.

[21] J. Rusz, J. Hlavnička, M. Novotnỳ, T. Tykalová, A. Pelletier, J. Montplaisir, J.-F. Gagnon, P. Dušek, A. Galbiati, S. Marelli, et al., Speech biomarkers in rapid eye movement sleep behavior disorder and Parkinson disease, Ann. Neurol. 90 (1) (2021) 62–75.

[22] A. Cantagallo, F. La Porta, L. Abenante, A. Bergonzoni, C. Giannone, S. D Altri, G. Ghiselli, La valutazione della disartria: il Profilo Robertson ed il Questionario di autovalutazione, Acta Phoniatrica Lat. 28 (1/2) (2006) 246.

[23] R. Capasso, G. Miceli, Esame Neuropsicologico Per L'Afasia: ENPA, Vol. 4, Springer Science & Business Media, 2001.

[24] U. Cesari, G. De Pietro, E. Marciano, C. Niri, G. Sannino, L. Verde, Voice disorder detection via an m-Health system: Design and results of a clinical study to evaluate Vox4Health, BioMed Res. Int. 2018 (2018).

[25] L. Verde, G. De Pietro, G. Sannino, Vox4Health: Preliminary results of a pilot study for the evaluation of a mobile voice screening application, in: Ambient Intelligence-Software and Applications–7th International Symposium on Ambient Intelligence, ISAmI 2016, Springer, 2016, pp. 131–140.

[26] J. Hlavnička, The Dysarthria analyzer, 2019, URL http://www.dysan.cz/.

[27] J. Hlavnička, Automated Analysis of Speech Disorders in Neurodegenerative Diseases (Ph.D. thesis), Czech Technical University, 2019.

[28] P. Podder, M.M. Hasan, M.R. Islam, M. Sayeed, Design and implementation of Butterworth, Chebyshev-I and elliptic filter for speech signal analysis, 2020, arXiv preprint arXiv:2002.03130.

[29] M.J.V. García, I. Cobeta, G. Martín, H. Alonso-Navarro, F.J. Jimenez-Jimenez, Acoustic analysis of voice in Huntington's disease patients, J. Voice 25 (2) (2011) 208–217.

[30] J.P. Lewis, Fast template matching, in: Vision Interface, Vol. 95, No. 120123, Quebec City, QC, Canada, 1995, pp. 15–19.

[31] J. Hlavnicka, T. Tykalov'a, O. Ulmanova, P. Dusek, D. Horakova, E. Ruzicka, J. Klempir, J. Rusz, Characterizing vocal tremor in progressive neurological diseases via automated acoustic analyses, Clin. Neurophysiol. 131 (5) (2020) 1155–1165.

[32] S.R. Garner, et al., Weka: The waikato environment for knowledge analysis, in: Proceedings of the New Zealand Computer Science Research Students Conference, Vol. 1995, Citeseer, 1995, pp. 57–64.

[33] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (6088) (1986) 533–536.

[34] D.S. Broomhead, D. Lowe, Radial Basis Functions, Multi-Variable Functional Interpolation and Adaptive Networks, Tech. rep., Royal Signals and Radar Establishment Malvern (United Kingdom), 1988.

[35] W.W. Cohen, Fast effective rule induction, in: Machine Learning Proceedings 1995, Elsevier, 1995, pp. 115–123.

[36] R.C. Holte, Very simple classification rules perform well on most commonly used datasets, Mach. Learn. 11 (1993) 63–90.

[37] E. Frank, I.H. Witten, Generating Accurate Rule Sets Without Global Optimization, University of Waikato, Department of Computer Science, 1998.

[38] P. Compton, R. Jansen, Knowledge in context: A strategy for expert system maintenance, in: AI'88: 2nd Australian Joint Artificial Intelligence Conference Adelaide, Australia, November 15–18, 1988 Proceedings, Vol. 2, Springer, 1990, pp. 292–306.

[39] J.R. Quinlan, C4. 5: Programs for Machine Learning, Elsevier, 2014.

[40] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[41] L.A. Breslow, D.W. Aha, et al., Simplifying decision trees: A survey, Knowl. Eng. Rev. 12 (1) (1997) 1–40.

[42] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1) (1960) 37–46.

[43] M.A. Hall, Correlation-based feature subset selection for machine learning, 1998, Thesis submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy at the University of Waikato.

[44] M. Dash, H. Liu, H. Motoda, Consistency based feature selection, in: Knowledge Discovery and Data Mining. Current Issues and New Applications: 4th Pacific-Asia Conference, PAKDD 2000 Kyoto, Japan, April 18–20, 2000 Proceedings, Vol. 4, Springer, 2000, pp. 98–109.

[45] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106.

[46] M. Robnik-Šikonja, I. Kononenko, et al., An adaptation of Relief for attribute estimation in regression, in: Machine Learning: Proceedings of the Fourteenth International Conference, Vol. 5, ICML'97, Citeseer, 1997, pp. 296–304.

[47] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1–2) (1997) 273–324.