# Automated Disease Detection by Machine Learning and Bio-Sound Analysis

Alberto Tena del Pozo

http://hdl.handle.net/10803/689256

Universitat de Lleida

# Automated Disease Detection by Machine Learning and Bio-Sound Analysis

by

Alberto Tena del Pozo

Submitted to the Department of Computer Science and Industrial Engineering and the Doctoral School of the University of Lleida in partial fulfillment of the requirements for the degree of

PhD Thesis in Engineering and Information Technology

at the

UNIVERSITY OF LLEIDA

March 2022

Thesis Supervisor . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Dr. Francesc Solsona
Academic Supervisor

Thesis Supervisor . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Dr. Francisco Clarià

**Universitat de Lleida**

# Automated Disease Detection by Machine Learning and Bio-Sound Analysis

by

Alberto Tena del Pozo

Submitted to the Department of Computer Science and Industrial Engineering and
the Doctoral School of the University of Lleida
on March 15, 2022, in partial fulfillment of the requirements for the degree of
PhD Thesis in Engineering and Information Technology

## Abstract

**Background:** This work investigated different methods based on machine learning bio-sounds analysis for the automatic identification of different conditions. Concretely, we conducted three studies to investigated the automatic identification of bulbar involvement in patients with amyotrophic lateral sclerosis (ALS) through voice analysis. Additionally, a study to detect COVID-19 positive cases through the automatic identification of COVID-19 coughs was performed.

The Northeast Amyotrophic Lateral Sclerosis Consortium (NEALS) bulbar subcommittee released a recent statement regarding the need for objective-based approaches to diagnose bulbar involvement in ALS patients. Bulbar involvement is a term used in ALS that refers to motor neuron impairment in the corticobulbar area of the brainstem which leads to a dysfunction of speech and swallowing. One of the earliest symptoms of bulbar involvement is voice deterioration, characterised by grossly defective articulation, extremely slow laborious speech, marked hypernasality and severe harshness. Bulbar involvement requires well-timed and carefully coordinated interventions. So, early detection is crucial to improving the quality of life and lengthening the life expectancy of those ALS patients who present this dysfunction. Recently, research efforts have focused on voice analysis to capture this dysfunction.

Analogously, easy detection of COVID-19 is a challenge. Quick biological tests do not give enough accuracy. Success in the fight against new outbreaks depends not only on the efficiency of the tests used, but also on the cost, time elapsed and the number of tests that can be done massively. Our proposal provides a solution to this challenge.

**Methods:** Three studies have been developed for the automated detection of bulbar involvement in patients with amyotrophic lateral sclerosis by machine learning and bio-sounds analysis.

The first study consisted of a methodology for diagnosing bulbar involvement efficiently through the acoustic parameters of uttered vowels in Spanish. The method focused on the extraction of features from the phonatory subsystem—jitter, shimmer, harmonics-to-noise ratio, and pitch—from the utterance of the five Spanish vowels. Then, we used various supervised classification algorithms, preceded by principal component analysis of the features obtained.

In the second study, we designed a new methodology for the automatic detection of bulbar involvement based on the phonatory subsystem and time-frequency characteristics. The methodology focused on providing a set of 50 phonatory subsystem and time-frequency features to detect this deficiency in males and females from the utterance of the five Spanish vowels. Then, multivariant analysis of variance was used to select the statistically significant features, and the most common supervised classifications models in clinical diagnosis were fitted to analyze their performance.

The third study consisted of providing a new methodology to automatically detect this dysfunction at early stages of the disease. The methodology focused on the creation of a voice fingerprint consisted of a pattern generated from the quasi-periodic components of a steady portion of the five Spanish vowels and from the computation of the five principal and independent components of this pattern. Then a set of statistically significant features were obtained and the most common supervised and semi-supervised classification models were implemented.

Additionally, a forth and last study was performed to design a freely available, quick and efficient methodology for the automatic detection of COVID-19 in raw audio files. The methodology was based on automated extraction of time-frequency cough features and selection of the more significant ones to be used to diagnose COVID-19 using a supervised machine-learning algorithm.

**Results:** In the first study, support vector machines performed better (Accuracy 95.8%) than the models analyzed in the related work. We also show how the model can improve human diagnosis, which can often misdiagnose bulbar involvement.

In the second study, we obtained a set of statistically significant features for males and females to capture this dysfunction. To date, the Accuracy obtained (98.01% for females and 96.10% for males both obtained with random forest), outperformed the models of our first study and those models found in the literature.

In the third study, random forest obtained the best accuracy (93.5%) when compared controls and ALS patients with bulbar involvement and support vector machines obtained 91.0% of *Accuracy* with 100.0% of *Specificity* when comparing directly ALS patients with and without bulbar involvement. Our model provided alternative annotation of bulbar and no bulbar subjects by means of semi-supervised machine-learning algorithms that improved even more the performance of our proposal.

In the fourth study, random forest has performed better to detect COVID-19 positive coughs than the other models analyzed. An Accuracy close to 90% was obtained.

**Conclusions:** The results obtained are very encouraging and demonstrate the efficiency and applicability of the machine learning bio-sounds analysis for the automated detection of certain conditions. It may be an appropriate tool to help in the diagnosis of ALS by multidisciplinary clinical teams, in particular to improve the diagnosis of bulbar involvement. It could also be useful to help for an early response to further COVID-19 outbreaks or other pandemics that may arise in the future.

The first study show how the model can improve human diagnosis, which can often misdiagnose bulbar involvement.

Adding time-frequency features to more classical phonatory-subsystem features increase the prediction capabilities of the machine learning models to detect bulbar involvement. Studying men and women separately has given additional success.

The results obtained to improve the annotation of ALS patients in whom bulbar involvement was not detected yet by using current subjective approaches are very encouraging and demonstrate the efficiency and applicability of the methodology presented. It may be an appropriate tool for screening bulbar involvement in early stages of the disease.

Finally, the fourth study demonstrates the feasibility of the automatic diagnose of COVID-19 from coughs, and its applicability to detecting new outbreaks.

# Resum

**Introducció:** En aquesta recerca hem investigat diferents mètodes d'aprenentatge automàtic basats en l'anàlisi de bio-sons per a la identificació automàtica de diferents malalties. Concretament, vam realitzar tres estudis per investigar la identificació automàtica de l'afectació bulbar en pacients amb esclerosi lateral amiotrófica (ELA) mitjançant anàlisis de veu. Adicionalment, es va realitzar un estudi per detectar casos positius de COVID-19 mitjançant la identificació automàtica de tos COVID-19.

El subcomitè bulbar del Northeast Amyotrophic Lateral Sclerosis Consortium (NEALS) va publicar una declaració recent sobre la necessitat d'enfocaments basats en paràmetres objectius per diagnosticar l'afectació bulbar en pacients amb ELA. L'afectació bulbar és un terme que s'utilitza en l'ELA que fa referència al deteriorament de les neurones motores a la zona corticobulbar del tronc cerebral que provoca una disfunció de la parla i de la deglució. Un dels primers símptomes d'afectació bulbar és el deteriorament de la veu, caracteritzat per una articulació greument defectuosa, parla laboriosa i extremadament lenta, hipernasalitat marcada i aspror severa. L'afectació bulbar requereix intervencions ben coordinades i ben temporitzades. Per tant, la detecció precoç és crucial per millorar la qualitat de vida i allargar l'esperança de vida dels pacients amb ELA. Recentment, els esforços de recerca s'han centrat en l'anàlisi de la veu per captar aquesta disfunció.

De manera anàloga, la fàcil detecció de la COVID-19 és un repte. Les proves biológiques no són prou precises. L'èxit en la lluita contra nous brots depèn no només de l'eficiència de les proves utilitzades, sinó també del cost, del temps transcorregut i del nombre de proves que es poden fer massivament. La nostra proposta dona una solució a aquest repte.

**Mètodes:** S'han desenvolupat tres estudis per a la detecció automatitzada de l'afectació bulbar en pacients amb ELA mitjançant models d'aprenentatge automàtic i l'anàlisi de bio-sons.

El primer estudi va consistir en desenvolupar una metodologia per diagnosticar eficaçment l'afectació bulbar a través dels paràmetres acústics de les vocals pronunciades en castellà. El mètode es va centrar en l'extracció de característiques del subsistema fonatori: fluctuació, brillantor, relació harmònica-soroll i to, a partir de la pronunciació de les cinc vocals espanyoles. A continuació, vam utilitzar diversos algorismes de classificació supervisats, precedits per l'anàlisi de components principals de les característiques obtingudes.

En el segon estudi, vam dissenyar una nova metodologia per a la detecció automàtica de l'afectació bulbar basada en característiques del subsistema fonatori i de temps-freqüència. La metodologia es va centrar en proporcionar un conjunt de 50 caractarístiques per detectar aquesta deficiència en homes i dones a partir de la pronunciació de les cinc vocals espanyoles. A continuació, es va utilitzar l'anàlisi multivariant de la variança per seleccionar les característiques estadísticament significatives i es van ajustar els models de classificacions supervisades més habituals en el diagnòstic clínic per analitzar el seu rendiment.

El tercer estudi va consistir en proporcionar una nova metodologia per detectar automàticament aquesta disfunció en les primeres etapes de la malaltia. La metodologia es va centrar en la creació d'una empremta digital de la veu basada en un patró generat a partir dels components quasi periòdics d'una porció constant de cadascuna de les cinc vocals espanyoles i del càlcul dels cinc components principals i independents d'aquest patró. Després es van obtenir un conjunt de característiques estadísticament significatives i es van implementar els models de classificació supervisats i semi-supervisats més comuns.

A més, es va realitzar un quart i darrer estudi per dissenyar una metodologia ràpida i eficient de lliure disposició per a la detecció automàtica de la COVID-19 en fitxers d'àudio en brut. La metodologia es va basar en l'extracció automàtica de les característiques temps-freqüència de la tos i de la selecció d'aquelles més significatives per diagnosticar la COVID-19 mitjançant algoritmes d'aprenentatge automàtic supervisat.

**Resultats:** En el primer estudi, les support vector machines van tenir un rendiment millor (Precisió 95,8 %) que els models analitzats en la literatura. També mostrem com el model pot millorar el diagnòstic humà, que sovint pot cometre errors en el diagnòstic de l'afectació bulbar.

En el segon estudi, vam obtenir un conjunt de característiques estadísticament significatives per a homes i dones per captar aquesta disfunció. La precisió obtinguda (98,01 % amb random forest per a les dones i 96,10 % amb random forest per als homes), va superar els resultats dels models del nostre primer estudi i la dels models trobats, fins ara, a la literatura.

En el tercer estudi, el model random forest va obtenir la millor Precisió (93,5%). A més, comparant directament pacients amb ELA amb i sense aquesta disfunció, les support vector machines van obtenir un $91,0\%$ de precisió i un $100,0\%$ de Especificitat. El nostre model va proporcionar una anotació alternativa de subjectes bulbars i no bulbars mitjançant algoritmes d'aprenentatge automàtic semi-supervisat que van millorar encara més el rendiment de la nostra proposta.

En el quart estudi, el model random forest va obtenir un millor rendiment en la detecció de la tos de la COVID-19 que la resta de models analitzats. Es va obtenir una Precisió propera al 90%.

**Conclusions:** Els resultats obtinguts són molt encoratjadors i demostren l'eficiència i l'aplicabilitat de models d'aprenentatge automàtic i de l'anàlisi de bio-sons per a la detecció automatitzada de determinades malalties.

Pot ser una eina adequada per als equips clínics multidisciplinaris per ajudar en el diagnòstic de l'ELA, en particular per millorar el diagnòstic d'afectació bulbar. També pot ser útil per ajudar a donar una resposta ràpida enfront d'altres brots de la COVID-19 o d'altres pandèmies que puguin sorgir en el futur.

El primer estudi mostra com el model pot millorar el diagnòstic humà, que sovint pot cometre errors en el diagnòstic de l'afectació bulbar.

Afegint característiques de temps-freqüència a les característiques més clàssiques obtingudes del subsistema fonador augmenten la capacitat de predicció dels models d'aprenentatge automàtic per detectar l'afectació bulbar. Estudiar homes i dones per separat ha proporcionat una millora addicional.

Els resultats obtinguts per millorar l'anotació de pacients amb ELA en els quals encara no s'havia detectat afectació bulbar mitjançant l'ús dels enfocaments subjectius actuals són molt encoratjadors i demostren l'eficiència i l'aplicabilitat de la metodologia presentada. Pot ser una eina adequada per al cribratge de l'afectació bulbar en les primeres etapes de la malaltia.

Finalment, el quart estudi demostra la viabilitat del diagnòstic automàtic de COVID-19 a partir de la tos i la seva aplicabilitat a la detecció de nous brots.

# Resumen

**Introducción:** En este trabajo se han investigado diferentes métodos basados en modelos de aprendizaje automático y en el análisis de bio-sonidos para la identificación automática de diferentes enfermedades. Concretamente, realizamos tres estudios para investigar la identificación automática de la afectación bulbar en pacientes

con esclerosis lateral amiotrófica (ELA) a través del análisis de la voz. Además, se realizó un estudio para detectar casos positivos de COVID-19 mediante la identificación automática de tos de la COVID-19.

El subcomité bulbar del Northeast Amyotrophic Lateral Sclerosis Consortium (NEALS) publicó una declaración reciente sobre la necesidad de enfoques basados en parámetros objetivos para diagnosticar la afectación bulbar en pacientes con ELA. La afectación bulbar es un término utilizado en la ELA que se refiere al deterioro de las neuronas motoras en el área corticobulbar del tronco encefálico que conduce a una disfunción del habla y la deglución. Uno de los primeros síntomas de la afectación bulbar es el deterioro de la voz, caracterizado por una articulación muy defectuosa, habla extremadamente lenta y laboriosa, hipernasalidad marcada y aspereza severa. La afectación bulbar requiere intervenciones oportunas y cuidadosamente coordinadas. Por tanto, la detección precoz es fundamental para mejorar la calidad de vida y alargar la esperanza de vida de aquellos pacientes con ELA que presentan esta disfunción. Recientemente, los esfuerzos de investigación se han centrado en el análisis de la voz para capturar esta disfunción.

De manera análoga, la fácil detección de la COVID-19 es un desafío. Las pruebas biológicas rápidas no son suficientemente precisas. El éxito en la lucha contra nuevos brotes depende no solo de la eficiencia de las pruebas utilizadas, sino también del coste, el tiempo transcurrido y la cantidad de pruebas que se pueden realizar de forma masiva. Nuestra propuesta brinda una solución a este desafío.

**Métodos:** Se han llevado a cabo tres estudios para la detección automática de la afectación bulbar en pacientes con esclerosis lateral amiotrófica mediante análisis de bio-sonidos por aprendizaje automático.

El primer estudio consistió en desarrollar una metodología para diagnosticar la afectación bulbar de manera eficiente a través de los parámetros acústicos de las vocales pronunciadas en español.

El método se centró en la extracción de características del subsistema fonatorio (jitter, shimmer, relación de armónicos a ruido y tono) a través de la pronunciación de las cinco vocales españolas. Luego, utilizamos varios algoritmos de clasificación supervisados, precedidos por el análisis de componentes principales de las características obtenidas.

En el segundo estudio, diseñamos una nueva metodología para la detección automática de afectación bulbar basada en características del subsistema fonatorio y de tiempo-frecuencia. La metodología se centró en proporcionar un conjunto de 50 características para detectar esta deficiencia en hombres y mujeres a partir de la pronunciación de las cinco vocales españolas. Luego, se utilizó el análisis multivariante de la varianza para seleccionar las características estadísticamente significativas, y se ajustaron los modelos de clasificación supervisados más utilizados en el diagnóstico clínico para analizar su rendimiento.

El tercer estudio consistió en aportar una nueva metodología para detectar de forma automática esta disfunción en etapas tempranas de la enfermedad. La metodología se centró en la creación de una huella de voz calculada en base a un patrón

generado a partir de los componentes cuasi periódicos de una porción constante de cada una de las cinco vocales españolas y del cálculo de los cinco componentes principales e independientes de este patrón. Luego se obtuvo un conjunto de características estadísticamente significativas y se implementaron los modelos de clasificación supervisados y semi-supervisados más comunes.

Adicionalmente, se realizó un cuarto y último estudio para diseñar una metodología de acceso libre, rápida y eficiente para la detección automática de la COVID-19 en archivos de audio sin procesar. La metodología se basó en la extracción automatizada de características tiempo-frecuencia de la tos y de la selección de aquellas más importantes para diagnosticar la COVID-19 mediante un algoritmo de aprendizaje automático supervisado.

**Resultados:** En el primer estudio, los modelos basados en support vector machines funcionaron mejor (Precisión 95,8%) que los modelos analizados en la literatura. También mostramos cómo el modelo puede mejorar el diagnóstico humano, que a menudo puede cometer errores en el diagnóstico de la afectación bulbar.

En el segundo estudio, obtuvimos un conjunto de características estadísticamente significativas para hombres y mujeres para capturar esta disfunción. Hasta la fecha, la Precisión obtenida (98,01 % con modelos random forest para el caso de las mujeres y 96,10 % con random forest para el de los hombres) superó los modelos de nuestro primer estudio y los modelos encontrados en la literatura.

En el tercer estudio, random forest obtuvo la mejor Precisión (93.5%) cuando se compararon controles y pacientes con ELA con afectación bulbar. Además, comparando directamente pacientes con ELA con y sin esta disfunción, las support vector machines obtuvieron un 91.0% de Precisión y un 100.0% de Especificidad. Nuestro modelo proporcionó una anotación alternativa de sujetos bulbar y no bulbar mediante algoritmos de aprendizaje automático semi-supervisados que mejoraron aún más el rendimiento de nuestra propuesta.

En el cuarto estudio, los random forest tuvieron un mejor desempeño para detectar la tos de la COVID-19 que los otros modelos analizados. Se obtuvo una Precisión cercana al 90%.

**Conclusiones:** Los resultados obtenidos son muy alentadores y demuestran la eficacia y aplicabilidad de métodos basados en modelos de aprendizaje automático y en el análisis de bio-sonidos para la detección automatizada de determinadas enfermedades.

Estos métodos pueden ser una herramienta adecuada para ayudar a los equipos clínicos multidisciplinares en el diagnóstico de la ELA, en particular para mejorar el diagnóstico de la afectación bulbar. También podrían ser útiles para ayudar a dar una respuesta rápida frente a los brotes de COVID-19 u otras pandemias que puedan surgir en el futuro.

El primer estudio muestra cómo el modelo puede mejorar el diagnóstico humano, que a menudo puede cometer errores a la hora de diagnosticar la afectación bulbar.

Añadir características de tiempo-frecuencia a características más clásicas como las

mencionadas del subsistema fonatorio aumenta las capacidades de predicción de los modelos de aprendizaje automático para detectar la afectación bulbar. El estudio de hombres y mujeres como grupos separados ha proporcionado una mejora adicional en el rendimiento de los modelos.

Los resultados obtenidos para mejorar la anotación de los pacientes con ELA en los que aún no se detectó afectación bulbar mediante el uso de los enfoques subjetivos actuales son muy alentadores y demuestran la eficacia y aplicabilidad de la metodología presentada. Puede ser una herramienta adecuada para detectar la afectación bulbar en las primeras etapas de la enfermedad.

Finalmente, el cuarto estudio demuestra la viabilidad del diagnóstico automático de la COVID-19 a través de la tos y su aplicabilidad para detectar nuevos brotes.

# Acknowledgments

Doing this PhD has been a great experience and an incredible opportunity to go beyond what I would have ever thought. It has been hard at times, however it has taught me plenty professionally and even more personally. It would not have been possible without all those people who helped me along this journey. Firstly, I would like to thank my supervisors, Dr. Francesc Solsona and Dr. Francisco Clarià, whose patient, knowledge, constant attention and patience supported me throughout this way. I would also like to thank all the members of the Department of Computer Science and Industrial Engineering, at the University of Lleida.

I want to acknowledge Prof. Einar Meister for being an excellent host and supervisor during my doctoral stay at the Tallinn University of Technology, and for providing valuable comments and suggestions on improving my work.

I would like to thank all the staff in CIMNE, in particular, Prof. Eugenio Oñate and Jordi Jiménez whose support and advise always guide me in my professional development.

I would also want to have some grateful words for all the staff of the Neurology Department of the Bellvitge University Hospital in Barcelona, especially, Dra. Mònica Povedano who believed in my work and gave me advice during my research, Carlos Augusto Salazar Talavera who illustrated the clinical records of the ALS patients, Dr. Marta Fulla and Maria Carmen Majos Bellmunt who advised me about the process of eliciting the sounds. I would like to have some special words for the patients who were always predisposed to collaborate making everything easy.

I want also like to thank Alfredo Jover, Cristina Acosta and Maria Ramirez from the UFIN of the Arnau de Vilanova University Hospital in Lleida, who recorded coughs using the website covid.udl.cat. We also thank the University of Cambridge (in particular, Cecilia Mascolo), the Coswara and Virufy projects, who voluntary share their cough datasets. We also thank everyone who volunteered their cough.

I would also like to thank my family for the support they have provided me with throughout my life, especially my parents, Santiago and Nicodema, who always

encouraged me to do my best and gave me warmth, sweetness and care.

My final words are for the most wonderful person I have had the opportunity to meet, my partner Noelia Vanesa. She helped me registering the sounds of the ALS patients. Her sweetness and kindness were very appreciated by them. Besides that, her care, love and patience supported me during this journey and meanwhile I was doing my research, life gave us the most wonderful present, our newborn Leire.

*Alberto Tena*

*Gelida, Catalonia*

# Contents

# Chapter 1

# Introduction and scope of the research

This work investigated different methods based on machine learning and bio-sound analysis such as voice and coughs for the automatic identification of different conditions.

Concretely, we investigated the automatic identification of bulbar involvement in patients with amyotrophic lateral sclerosis (ALS) through voice analysis and COronaVIrus Disease of 2019 (COVID-19) positive cases through the automatic identification of COVID-19 coughs.

Firstly, the background and the context of the addressed problems are presented, both for the detection of the bulbar involvement in ALS patients and for the detection of cough of COVID-19.

Then, the related work in machine learning and voice analysis in ALS patients and in machine-learning tools to detect COVID-19 based on the sound of voices, and the sounds we make when we breathe or cough is reviewed.

Later on, the objectives and the main contributions of this thesis are introduced.

Subsequently, a summary of the main methodologies developed in this thesis is presented. Basically, it consists of the three types of analysis which were performed in this work to obtain the features (phonatori subsystem, time-frequency representation, and pattern analysis). Next, the feature analysis and feature extraction techniques studied are presented and the visualization tools used for exploring the features.

Then, the machine-learning algorithms used are explained, as well as the corpus and datasets employed both for ALS and COVID-19 for the experimentation.

Finally, the publications obtained in the thesis and the 3-month doctoral stay which took place at the Computer Science department of the University of Tallinn are presented.

## 1.1 Background

### 1.1.1 Detecting bulbar involvement in ALS patients

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease with an irregular and asymmetric progression. This is characterized by a progressive loss of both upper and lower motor neurons leading to muscular atrophy, paralysis and death, mainly from respiratory failure. The life expectancy of these patients from the onset of symptoms is from 3 to 5 years. ALS produces muscular weakness and difficulties of mobilization, communication, feeding and breathing, creating a great dependence of the patient on caregivers and relatives and generating significant social costs. Currently, there is no cure for ALS, although early detection can slow progress [1].

The disease is referred to as spinal ALS (80% of cases) when the first symptoms appear in the arms and legs (limb or spinal onset), and bulbar ALS (20% of cases) when it begins in cranial nerve nuclei (bulbar onset). The patients with the latter form tend to have a shorter life span because of the critical nature of bulbar muscle function responsible for speech and swallowing. The first bulbar symptoms appear at the beginning of the disease in bulbar ALS, but may appear in later stages of spinal ALS. Early identification of bulbar involvement is critical for improving diagnosis and prognosis and may be the key to slowing the disease effectively. However, its diagnosis is challenging due to the difficulties of assessing this impairment through subjective measures. There is no standardised diagnostic procedure for assessing bulbar dysfunction yet and new methodologies based on objective measures are needed

[2].

### 1.1.2 Detecting COVID-19 cough

Meanwhile we conducted the ALS project, the COVID-19 pandemic abruptly burst in our daily-life, and the scientific community focused their efforts to fight against this pandemic. As we considered that the bio-sounds analysis could be applied to detect other conditions, and with the aim to provide new hints to fight against COVID-19 pandemic, we performed a study to detect COVID-19 positive cases from the automatic identification of COVID-19 coughs.

COVID-19, caused by the Severe Acute Respiratory Syndrome (SARS-CoV2) virus, was announced as a global pandemic on February 11, 2020 by the World Health Organisation (WHO). By mid-February, 2021, one year after the beginning of the COVID-19 pandemic, over 108 million confirmed cases of COVID-19 had been reported worldwide, with almost 2,400,000 deaths [3]. During this time, it has been demonstrated that COVID-19 outbreaks are very hard to contain with current testing approaches unless region-wide confinement measures are sustained. This is partly because of the limitations of current viral and serological tests and the lack of complementary pre-screening methods [4].

According to the WHO-China Joint Mission report (COVID-19) [5], typical signs and symptoms of COVID-19 are fever (87.9%), dry cough (67.7%), fatigue (38.1%), sputum production (33.4%), shortness of breath (18.6%), sore throat (13.9%), headache (13.6%), myalgia or arthralgia (14.8%), chills (11.4%), nausea or vomiting (5.0%), nasal congestion (4.8%), diarrhoea (3.7%), hemoptysis (0.9%), and conjunctival congestion (0.8%).

Several researchers have proposed methods for identifying cough sounds from audio recordings [6, 7]. Automatic cough classification is an active research area in which several researchers have proposed methods for identifying a wide range of respiratory diseases and types of coughs (namely dry and wet coughs) through cough analysis and machine-learning algorithms [8, 9].

## 1.2 Related Work

### 1.2.1 Related Work in machine learning and voice analysis in ALS patients

Table 1.1 summarizes the most recently published work related to machine learning and voice analysis in ALS patients.

**Citation**: Connaghan et al [10], 2019.

**Purpose/Thesis**: Addressed the utility of Beiwe smartphone-based digital phenotyping to identify and track speech decline in ALS.

**Data Collection**: 12 participants with ALS used the Beiwe app weekly to record reading passages and self-report (ALSFRS-R) ratings of bulbar (speech) function.

**Methodology**: Speaking rate and pause variables were automatically extracted from recordings offline. Speech function measures at baseline were significantly different for participants with and without bulbar symptoms.

**Outcomes**: They observed that the rate of decline of all measured speech functions was greater for participants with bulbar symptoms.

**Citation**: Lee et al [11], 2019.

**Purpose/Thesis**: Investigate vowel-specific intelligibility and acoustic patterns of individuals with different severities of dysarthria secondary to ALS.

**Data Collection**: 23 individuals with dysarthria secondary to ALS and 22 typically aging individuals participated as speakers.

**Methodology**: For vowel-specific intelligibility data, 135 listeners participated in the study. Vowel-specific intelligibility, intrinsic vowel duration, 1st and 2nd formants (F1 and F2), vowel inherent spectral change (VISC), and absolute VISC were examined.

**Outcomes**: A significant interaction between severity group and the vowel-specific intelligibility pattern as well as F1, F2 VISC, and absolute F2 VISC was observed.

**Citation**: Chiaramonte et al [12], 2019.

**Purpose/Thesis**: Examine the role of different specialists in the diagnosis of ALS, to understand changes in verbal expression and phonation, respiratory dynamics and swallowing that occurred rapidly over a short period of time.

**Data Collection**: 22 patients with bulbar ALS

**Methodology**: Voice assessment, ears, nose and throat (ENT) evaluation, Multi-Dimensional Voice Program (MDVP), spectrogram, electroglottography, fiberoptic endoscopic evaluation of swallowing were performed for each patient.

**Outcomes**: In the early stage of the disease, the oral tract and velopharyngeal port were involved. Values of MDVP were altered. Spectrogram showed an additional formant, due to nasal resonance. Electroglottography showed periodic oscillation of the vocal folds only during short vocal cycle.

**Citation**: Suhas et al [13], 2019.

**Purpose/Thesis**: Speech based automatic classification of patients with ALS and healthy subjects.

**Data Collection**: 25 ALS patients and 25 healthy subjects.

**Methodology**: Sustained phoneme production (PHON), diadochokinetic task (DDK) and spontaneous speech (SPON) were used as speech tasks. Support vector machines (SVM) and deep neural networks (DNNs) were used as classifiers and suprasegmental features based on mel frequency cepstral coefficients (MFCCs) were considered.

**Outcomes**: The best classification Accuracy of 92.2% is obtained using a high quality microphone.

---

**Citation**: Garcia-Gancedo et al [14], 2019.

**Purpose/Thesis**: Investigate the feasibility of a novel digital platform for remote data collection of multiple symptoms including digital speech characteristics to explore the impact of the devices on patients' everyday life.

**Data Collection**: 25 patients with ALS in an observational clinical trial setting.

**Methodology**: Patients attended a clinical site visit every 3 months to perform activity reference tasks while wearing a sensor, to conduct digital speech tests and for conventional ALS monitoring. In addition, patients wore the sensor in their daily life for approximately 3 days every month for the duration of the study.

**Outcomes**: The platform can measure physical activity in patients with ALS in their home environment; patients used the equipment successfully, and it was generally well tolerated. Good-quality in-clinic speech data were successfully captured for analysis at home.

---

**Citation**: Gutz et al [15], 2019.

**Purpose/Thesis**: Machine learning approach to detect ALS prior to the onset of overt speech symptoms.

**Data Collection**: 123 participants who were stratified by sex and into three groups: healthy controls, ALS symptomatic, and ALS presymptomatic.

**Methodology**: They compared models trained on three group pairs (symptomatic-control, presymptomatic-control, and all ALS-control participants). Using acoustic features obtained with the OpenSMILE ComParE13 configuration, they tested several feature filtering techniques. machine learning classification was achieved using an SVM model and leave-one-out cross-validation.

**Outcomes**: The most successful model, which was trained on symptomatic-control data, yielded an Area Under the Curve (AUC)=0.99 for females and AUC=0.91 for males. Models trained on all ALS-control participants had high diagnostic accuracy for classifying symptomatic and presymptomatic ALS participants (females: AUC=0.85; males: AUC=0.91).

---

**Citation**: Vashkevich et al [16], 2019.

**Purpose/Thesis**: Verify the suitability of the sustain vowel phonation test for automatic detection of patients with ALS

**Data Collection**: 15 ALS patients with signs of bulbar dysfunction and 39 healthy speakers.

**Methodology**: They developed a procedure for separation of voice signal into fundamental periods for calculation of perturbation measurements.

**Outcomes**: Linear discriminant analysis (LDA) attained 90.7% Accuracy with 86.7% Sensitivity and 92.2% Specificity.

**Citation**: Wang et al [17], 2018.

**Purpose/Thesis**: This research aimed to automatically predict intelligible speaking rate for individuals with ALS based on speech acoustic and articulatory samples.

**Data Collection**: 12 participants with ALS and 2 control subjects produced a total of 1831 phrases.

**Methodology**: Northern Digital Inc.(NDI) Wave system was used to collect tongue and lip movement and acoustic data synchronously. A machine-learning algorithm (i.e. SVM) was used to predict intelligible speaking rate (speech intelligibility · speaking rate) from acoustic and articulatory features of the recorded samples.

**Outcomes**: The results revealed that the proposed analyses predicted the intelligible speaking rate of the participant with reasonably high accuracy by extracting the acoustic and/or articulatory features from one short speech sample.

**Citation**: Norel et al [18], 2018.

**Purpose/Thesis**: Identification of acoustic speech features in naturalistic contexts which characterize disease progression and development of machine models which can recognize the presence and severity of the disease.

**Data Collection**: Prize4Life Israel dataset. The dataset was generated using the ALS Mobile Analyzer.

**Methodology**: The subjects were evaluated subjects using a variety of frequency, spectral, and voice quality features.

**Outcomes**: Classification via leave-five-subjects-out cross-validation resulted in an Accuracy of 79% (61% chance) for males and 83% (52% chance) for females.

**Citation**: An et al [19], 2018.

**Purpose/Thesis**: Explore the feasibility of automatic detection of patients with ALS at an early stage from highly intelligible speech

**Data Collection**: 13 newly diagnosed patients with ALS and 13 age and gender-matched healthy controls.

| |
|---|
| **Methodology**: Convolutional Neural Networks (CNNs), including time-domain CNN and frequency domain CNN, were used to classify the intelligible speech produced by patients with ALS and those by healthy individuals.<br>**Outcomes**: The best result was obtained by frequency-CNN (76.9% Sensitivity and 92.3% Specificity). Results demonstrated the possibility of early detection of ALS from intelligible speech signals. |
| **Citation**: Spangler et al [20], 2017.<br>**Purpose/Thesis**: Fully automated approach of detecting dysarthria in ALS patients.<br>**Data Collection**: 49 ALS patients and 34 healthy speakers.<br>**Methodology**: The proposed method used novel features based on fractal analysis. Acoustic and associated articulatory recordings of a standard speech diagnostic task, the diadochokinetic test (DDK), were used for classification.<br>**Outcomes**: Overall results obtained 90.2% accuracy with 94.2% sensitivity and 85.1% specificity. |
| **Citation**: Shellikeri et al [21], 2016.<br>**Purpose/Thesis**: Identify the effects of ALS on tongue and jaw control, both cross-sectionally and longitudinally. The data were examined in the context of their utility as a diagnostic marker of bulbar disease.<br>**Data Collection**: Cross-sectional data: 33 ALS individuals, 13 controls longitudinal data: 10 ALS individuals.<br>**Methodology**: Tongue and jaw movements were recorded using a three-dimensional electromagnetic articulography system during the production of the sentence Buy Bobby a puppy. The movements were examined for evidence of changes in size, speed, and duration and with respect to disease severity and time in the study. |

**Outcomes**: Maximum speed of tongue movements and movement durations were significantly different only at an advanced stage of bulbar ALS compared with the healthy control group. The longitudinal analysis revealed a reduction in tongue movement size and speed with time at early stages of disease, which was not seen cross-sectionally.

**Citation**: Horwitz-Martin et al [22], 2016.

**Purpose/Thesis**: Identify acoustic features that aid in predicting intelligibility loss and speaking rate decline in individuals with ALS.

**Data Collection**: Longitudinal data from 123 subjects with ALS.

**Methodology**: Features were derived from statistics of the first (F1) and second (F2) formant frequency trajectories and their first and second derivatives.

**Outcomes**: F2 features, particularly mean F2 speed and a novel feature, mean F2 acceleration, were most strongly correlated with intelligibility and speaking rate, respectively.

**Citation**: Rong et al [23], 2016.

**Purpose/Thesis**: Determine the mechanisms of speech intelligibility impairment due to neurologic impairments.

**Data Collection**: 66 individuals diagnosed with ALS were studied longitudinally.

**Methodology**: The disease-related changes in articulatory, resonatory, phonatory, and respiratory subsystems were quantified using multiple instrumental measures.

**Outcomes**: Declines in maximum performance tasks such as the alternating motion rate preceded declines in intelligibility, thus serving as early predictors of bulbar dysfunction.

**Citation**: Tomik et al [24], 2015.

**Purpose/Thesis**: Analyze the phonatory function of the larynx in ALS patients.

**Data Collection**: 17 patients with ALS. Examinations were performed three times at 6-month intervals.

|  |
|---|
| **Methodology**: They were evaluated with subjective perceptual voice assessment (including the grade, roughness, breathiness, asthenia, strain (GRBAS) scale), video-laryngostroboscopy including voice range and maximum phonation time (MPT), and objective acoustic voice analysis evaluation of jitter, shimmer, mean fundamental frequency, and harmonics-to-noise ratio (HNR)).<br>**Outcomes**: Analysis of voice qualities among patients with ALS allows for the detection of various abnormalities associated with the natural progression of the disease. |

Table 1.1: Related work. Machine learning and voice analysis in ALS patients.

Previous voice and speech production studies reveal significant differences in specific acoustic parameters in ALS patients. Carpenter et al. [25] studied the articulatory subsystem of individuals with ALS and found different involvement of articulators, i.e. tongue function was more involved than jaw function. In recent studies, Shellikeri et al. [21] found that the maximum speed of tongue movements and movement durations were only significantly different at an advanced stage of bulbar ALS compared with the healthy control group and Connaghan et al. [10] used a smartphone app to identify and track speech decline. Lee et al [11], obtained acoustic patterns of vowels in relation to the severity of dysarthria in ALS patients.

Other works such as [12, 17, 20, 22, 24, 26, 27] demonstrated the efficiency of features obtained from the phonatory subsystem in detecting early deterioration in ALS. Studies such as [12, 24, 26] demonstrated significant differences between jitter, shimmer and HNR in ALS patients. Alternative approaches used formant trajectories to classify the ALS condition [22], correlating formants with articulatory patterns [27], fractal jitter [20], MFCCs [13] or combining acoustic and motion-related features [17]. Other related studies such as Frid et al. [28] used speech formants and their ratios to diagnose neurological disorders. Teixeira et al. [29] and Mekyska et al. [30] suggested jitter, shimmer and HNR as good parameters to be used in intelligent diagnosis systems for dysphonia pathologies.

Besides that, time-frequency representations (TFR) are broadly applied to the detection of several conditions [31, 32, 33, 34] and was recently used to detect pathological changes in the voice signals [35]. TFR enables the evolution of the periodicity and frequency components with time to be observed, allowing the analysis of non-stationary signals, such as voice signals [36]. Quasi-periodic waveform analysis has been applied in several clinical applications such as heartbeat detection, cardiopulmonary modeling and intrinsic brain activity detection [37, 38]. Garcia-Gancedo et al. [14] demonstrated the feasibility of a novel digital platform for remote data collection of digital speech characteristics among others parameters from ALS patients.

Concerning machine learning, classification models are widely used to test the performance of acoustic parameters in the analysis of pathological voices. R. Norel et al. [18] identified acoustic speech features in naturalistic contexts and machine-learning models developed for recognizing the presence and severity of ALS using a variety of frequency, spectral, and voice quality features. Wang et al. [17] explored the classification of the ALS condition using the same features with SVM and neuronal networks (NNs) classifiers. Rong et al. [23] used SVM with two feature selection techniques (decision tree and gradient boosting) to predict the intelligible speaking rate from speech acoustic and articulatory samples. Suhas et al. [13] implemented SVM and DNNs for automatic classification by using MFCCs. An et al. [19] used CNNs to classify the intelligible speech produced by patients with ALS compared with healthy individuals. Gutz et al. [15] merged SVM and feature filtering techniques (SelectKBest). In addition, Vashkevich et. al [16] used LDA to verify the suitability of the sustain vowel phonation test for automatic detection of patients with ALS.

### 1.2.2 Related work in machine learning and cough analysis in COVID-19.

Dry cough sound analysis has proven successful in diagnosing respiratory conditions like pertussis [8], asthma, and pneumonia [9].

Cough detection is an active research area in which several researchers have proposed methods for identifying cough sounds from audio recordings [8]. Martinek et al. [6] reported good results distinguishing between voluntary cough sound and speech.

However, their method was subject-dependent. They argued that cough sound would have a higher degree of irregularity compared with speech. They computed the sample entropy [39]. This is a measure of the irregularity and unpredictability of the signal. Therefore, it is higher for noisy signals compared to periodic oscillations.

Barry et al. [40] used digital signal processing to calculate characteristic spectral coefficients of sound events, which are then classified into cough and non-cough events by the use of a probabilistic neural network (PNN). Such parameters as the total number of coughs and cough frequency as a function of time were obtained from the results of the audio processing.

Swarnkar et al. [41] used other spectral features such as formant frequencies, kurtosis, and B-score together with MFCCsfeatures for cough detection. These were fed into a neural network. Matos et al. [42] used thirteen MFCCs which were classified using a Hidden Markov Model (HMM). Liu et al. [43] proposed a feature extraction method called Gammatone Cepstral Coefficient (GMCC) with SVM classification. They showed that together GMCC and MFCC surpassed MFCC used alone. Lucio et al. [44] extracted 79 MFCC and Fast Fourier Transform (FFT) coefficients and used k-Nearest Neighbor (kNN) for classification.

Several algorithms for automatic cough classification have been published to identify various cough types. Chatrzarrin et al. [7] studied the different phases of dry and wet coughs and found the second phase of dry coughs to have lower energy compared to wet coughs. They also noted that, during this phase, most of the signal power is contained between 0-750 Hz in the case of wet coughs and 1500-2250 Hz in the case of dry coughs. Using a simple thresholding method, they successfully identified 14 wet and dry coughs with 100% Accuracy. Swarnkar et al. [41] used a Logistic Regression (LR) to discriminate between dry and wet coughs from pediatric patients with different respiratory illnesses. B–score, non-Gaussianity, formant frequencies, kurtosis, zero crossing rate and MFCCs were used as features. Kosasih et al. [45] used MFCCs, non–Gaussianity index and wavelet features with a LR classifier to differentiate between pneumonia and non–pneumonia cough sounds. Parker et al. [46] used 13 MFCCs features and the energy level to detect pertussis cough from sounds available on the Internet. NNs, random forest (RF) and kNN classifiers were used.

Various studies have begun to work on the design of machine-learning tools to detect COVID-19 cough [47, 48, 49, 50, 51, 52, 53, 54, 55] as complementary pre-screening method (see Table 1.2).

**Citation**: Feng et al [55], 2021

**Purpose/Thesis**: Developping a method for the automatic diagnosis of COVID-19 by detecting cough during recorded conversations.

**Data Collection**: They used Coswara [56] and Virufy [57] datasets.

**Methodology**: Their method was composed of five modules: sound extraction, sound feature extraction, cough detection, cough classification, and COVID-19 diagnosis. The method extracted relevant features from the audio signal and then used machine-learning and deep learning models to make the prediction.

**Outcomes**: Overall Accuracies of 81.25% (AUC of 0.79) were obtained.

**Citation**: Pahar et al [51], 2021

**Purpose/Thesis**: Machine learning based COVID-19 cough classifier which can discriminate COVID-19 positive coughs from both COVID-19 negative and healthy coughs recorded on a smartphone.

**Data Collection**: They used the publicly available Coswara dataset [56]

**Methodology**: Dataset skew was addressed by applying the synthetic minority over-sampling technique (SMOTE). A leave-p-out cross-validation scheme was used to train and evaluate seven machine-learning classifiers: LR, kNN, SVM, multilayer perceptron (MLP), CNNs, long short-term memory (LSTM) and a residual-based neural network architecture (Resnet50)

**Outcomes**: Resnet50 classifier was best able to discriminate between the COVID-19 positive and the healthy coughs with an AUC of 0.98. An LSTM classifier was best able to discriminate between the COVID-19 positive and COVID-19 negative coughs, with an AUC of 0.94 after selecting the best 13 features from a sequential forward selection (SFS).

**Citation**: Vrindavanam et al [53], 2021

**Purpose/Thesis**: Contactless detection of COVID-19 patients by analyzing their respective cough audio samples.

**Data Collection**: 86 cough audios, sampled at 44 kHz out of which 54 COVID-19 positive cough audio samples and 32 healthy individuals cough audio samples.

**Methodology**: 65 features were extracted using librosa [58] and implemented three machine-learning classifiers consisted of LR, SVM and RF.

**Outcomes**: RF obtained the best Accuracy (83.9%) with a Sensitivity of 81.2% and a Precision of 76.9%.

**Citation**: Laguarta et al [47], 2020

**Purpose/Thesis**: They hypothesized that COVID-19 subjects, especially including asymptomatics, could be accurately discriminated only from a forced-cough cell phone recording using Artificial Intelligence.

**Data Collection**: They built a data collection pipeline of COVID-19 cough recordings through a website (opensigma.mit.edu) between April and May 2020 and created an audio COVID-19 cough dataset with 5,320 subjects.

**Methodology**: MFCCs of coughs were inputted into a CNN made up of one Poisson biomarker layer and 3 pre-trained ResNet50's in parallel. Transfer learning was used to learn biomarker features on larger datasets.

**Outcomes**: When validated with subjects diagnosed using an official test, the model achieves COVID-19 Sensitivity of 98.5% with a Specificity of 94.2% (AUC: 0.97). For asymptomatic subjects it achieves Sensitivity of 100% with a Specificity of 83.2%

**Citation**: Brown et al [52], 2020

**Purpose/Thesis**: Data analysis over a large-scale crowdsourced dataset of respiratory sounds collected to aid diagnosis of COVID-19.

**Data Collection**: 141 cough and breathing samples from COVID-19 patients. 298 non-COVID-19 samples. 32 non-COVID-19 samples from subjects who presented cough as a symptom.

**Methodology**: Handcrafted features were extracted covering frequency-based, structural, statistical and temporal attributes. Features were also obtained though transfer learning from VGGish. They tested the performance of LR, Gradient Boosting Trees (GBTs) and SVM classifiers.
**Outcomes**: The preliminary results of the models achieved an AUC of above 80%.

**Citation**: Imram et al [50], 2020
**Purpose/Thesis**: Develop and test an Artificial Intelligence-powered screening solution for COVID-19 infection to be deployable via a smartphone app.
**Data Collection**: 96 bronchitis cough samples; 130 pertussis cough samples; 70 COVID-19 cough samples; and 247 normal cough samples.
**Methodology**: They investigated the distinctness of alterations in the respiratory system induced by COVID-19 infection when compared to other respiratory infections. They exploited transfer learning and classical machine-learning approaches.
**Outcomes**: Overall, the models used achieved 90% of Accuracy.

Table 1.2: Related work. Machine-learning tools to detect COVID-19 cough.

These are based on the analysis of the sound of voices, and the sounds we make when we breath or cough and which change when our respiratory system is affected. These changes range from coarse, clearly audible changes, to minute changes (called micro signatures) that are inaudible to the untrained listener, but nevertheless present [48]. These works have been performed in own datasets and no idenfication of the main features has been performed. We are also interested in the automatic identification of COVID-19 cough from any raw audio recording. Overall, finding a general method and the main cough features from audio records for diagnosing COVID-19 is a challenge. The difficulty is to find good machine-learning features. Some works in the literature, as we have mentioned before, advocate some features, but in the particular case of COVID-19, it remains to be seen which properties, brands, signs (that is, features) are those that uniquely identify COVID-19. So, the big challenge is to identify the best features that discriminate the COVID-19 cough. In addition, we want to find the group of features with better performance for each type of experiment, as for example, comparing COVID-19 and pertussis coughs.

## 1.3 Objectives and contributions

### 1.3.1 Objectives and contributions for the automatic identification of bulbar involvement in ALS patients

Three different studies have been conducted.

In the first study, we suggested that the acoustic parameters obtained through automated signal analysis from a steady portion of sustained vowels may be used efficiently as predictors for early detection of bulbar involvement in ALS patients. The study focused on the extraction of features from the phonatory subsystem: jitter, shimmer, HNR and pitch, from the utterance of each Spanish vowel. The features chosen for analysis were selected to provide information regarding changes in the vocal signal believed to reflect physiologic changes of the vocal folds. For that purpose, the main objectives (and contributions) of this research were:

1. To design a methodology for diagnosing bulbar involvement efficiently through the acoustic parameters of uttered vowels in Spanish.

2. To demonstrate the better performance of automated diagnosis of bulbar involvement compared with human diagnosis.

In a second study, we used TFR to obtain additional features. Starting from our previous study [59], we used the phonatory subsystem features obtained in the previous work and added time-frequency features to improve the performance of the classification models for early detection of bulbar involvement in ALS. To that end, the main objectives (and contributions) of this research were:

1. To design a new methodology for automatic detection of bulbar involvement in males and females based on phonatory subsystem and time-frequency features;

2. To obtain a set of statistically significant features for diagnosing bulbar involvement efficiently.

3. To analyze the performance of the most common supervised classification models to improve the bulbar involvement diagnosis.

Alternatively, in the third study, we conjectured that the diagnosis of ALS patients with bulbar dysfunction would greatly benefit from the creation of a voice fingerprint able to detect bulbar dysfunction in ALS before the first symptoms can be detected by the human hearing. This could be effectively done by means of the analysis of a pattern generated from the quasi-periodic waveform produced by the vocal folds when a vowel is elicited. Furthermore, performance could increase by correcting the bias as well as enlarging the corpus upsampling it [60], and relabeling bulbar and non-bulbar ALS patients by using semi-supervised classifiers, as was pointed out in [61, 62].

Our objective (and contribution) of this study was creating a machine-learning model obtained by applying supervised and unsupervised classifiers and upsampling to improve the corpus for diagnosing bulbar dysfunction by the creation of a voice fingerprint consisting of a pattern generated from the quasi-periodic components of a steady portion of the five Spanish vowels, and the five principal and independent components of this pattern. This model should be behaving properly with small and usually, badly annotated corpus, the kind of corpus associated to rare diseases (i.e. ALS with bulbar involvement).

### 1.3.2 Objectives and contributions for the automatic identification of COVID-19 cough

The goal was to develop a pre-screening method that could lead to automated identification of COVID-19 through the analysis of cough TFR with similar performance presented in [47, 48, 49, 50, 51, 52, 53, 54, 55]. TFR permit the evolution of the periodicity and frequency components over time to be observed, allowing the analysis of non-stationary signals. Moreover, this representation, which maintains the time dependence of signal features, gives the possibility of introducing more related features than traditional analysis. This way, we go a step further by finding the set of time-frequency features that could allow COVID-19 coughs to be distinguished from other cough patterns and validate it as a more generic proposal by applying our method to various datasets from different sources.

Prior to performing the TFR analysis, the YAMNet [63] deep neuronal network was used for the automatic identification of cough sounds in raw audio files. Then, a TFR analysis of a Choi-Williams distribution (CWD) was carried out in the cough-samples automatically identified to obtain discriminatory features for an automated diagnosis of COVID-19. 39 features were extracted and the sets which showed better performance at discriminating COVID-19 cough were selected. For that purpose, the main objectives (and contributions) of this research were:

1. To design a free, quick and efficient methodology for the automatic detection of COVID-19 in raw audio files based on the time-frequency analysis of the cough.

2. To obtain the time-frequency discriminatory features leading to automated identification of COVID-19.

3. To find an optimal supervised machine-learning algorithm to diagnose COVID-19 from the cough features found.

## 1.4 Bio-sounds Analysis

Different sounds exist in Spanish language (namely vocalic, diphthongs, occlusives, fricatives, affricates and sonorants). Sounds are different from each other depending on their geometry, physical, articulatory and acoustic characteristics. Vowels are generated by the vibration of the vocal cords (also known as vocal folds) in the glottic area. During their production, articulatory and larynx regions are involved. The flow of air from the lungs crosses the supraglottic cavities and passes through the mouth, which functions as a resonance chamber, with minimal obstruction and without audible friction. Acoustically, vowel sounds are made up of a complex periodic waveform, whose profile is recurrent at regular time intervals. This periodic waveform experiences the phenomenon of resonance when it crosses the supraglottic cavities and some of its harmonics are amplified.

Table 1.3 summarizes the geometry, physical, articulatory and acoustic characteristics and acoustic analysis associated to vocalic sounds.

| **Sounds/ Geometry** | **Articulatory Characteristics** | **Acoustic Characteristics** | **Acoustic Analysis** |
|---|---|---|---|
| a[-consonant] [+sonorant] [dorsal]AR: [+low] [+retracted] [+voicing]LR [+continuous] | LOW and CENTRAL palatal vowel. The tongue is located in the lower part of the oral cavity. | Complex periodic sound wave. Recurrent profile at regular time intervals. Amplification of some of its HARMONICS when it crosses the supraglottic cavities (RESONANCE). F1 and F2 are very close, but they differ significantly from those the back vowels because they are in higher frequencies range in the spectrum (MEDIUM VOWEL). Mean values of F1 (753 Hz) and F2 (1260 Hz). | Fundamental frequency (F0): Vibration frequency of the vocal folds. Variations of F0 may indicate changes in the opening of the glottis and modifications in the rigidity of the mucosa of the vocal folds. Oral opening: the more open the vowel, the higher the frequency of F1. Tongue position: Posterior position for vowel a. The more anterior the vowel articulation the higher F2. |

| | | | |
|---|---|---|---|
| e<br>[-consonant]<br>[+sonorant]<br>[dorsal]AR:<br>[-high] and<br>[-low]<br>[-retracted]<br>[+voicing]LR<br>[+continuous] | MEDIUM and ANTERIOR palatal vowel. The tongue approaches the anterior part of the palate and, therefore, the resonance cavity is small. | Complex periodic sound wave. Amplification of some of its HARMONICS when it crosses the supraglottic cavities (RESONANCE). High F2 values and very different from F1 (HIGH-PITCHED). Mean values of F1 (465 Hz) and F2 (1780 Hz). Anterior vowels are high-pitched. They are clearly different from posteriors and low-pitched vowels. | Fundamental frequency (F0): Vibration frequency of the vocal folds. Variations of F0 may indicate changes in the opening of the glottis and modifications in the rigidity of the mucosa of the vocal folds. Oral opening: the more open the vowel, the higher the frequency of F1. Tongue position: Approaches the front of the palate. The more anterior the vowel articulation the higher F2. |
| i<br>[-consonant]<br>[+sonorant]<br>[dorsal]AR:<br>[+high]<br>[-retracted]<br>[+voicing]LR<br>[+continuous] | HIGH and ANTERIOR (PALATAL) Vowel. The tongue, higher than in vowel e pronunciation, approaches the anterior part of the palate. | Complex periodic sound wave. Amplification of some of its HARMONICS when it crosses the supraglottic cavities (RESONANCE). High F2 values and very different from F1 (HIGH-PITCHED). Mean values of F1 (298 Hz) and F2 (2188 Hz). Anterior vowels are high-pitched. They are clearly different from posteriors and low-pitched vowels. | Fundamental frequency (F0): Vibration frequency of the vocal folds. Variations of F0 may indicate changes in the opening of the glottis and modifications in the rigidity of the mucosa of the vocal folds. Oral opening: the more open the vowel, the higher the frequency of F1. Tongue position: Approaches the anterior part of the palate higher than in e. The more anterior the vowel articulation the higher F2. |

| | | | |
|---|---|---|---|
| o<br>[-consonant]<br>[+sonorant]<br>[labial]AR:<br>[+rounded]<br>[dorsal]AR:<br>[-high] and<br>[-low] [+retracted]<br>[+voicing]LR<br>[+continuous] | MEDIUM and BACK velar Vowel. The tongue approaches the palate in the posterior area of the oral cavity. A wide and long resonance cavity is configured with a severe timbre. | Complex periodic sound wave. Amplification of some of its HARMONICS when it crosses the supraglottic cavities (RESONANCE). F1 and F2 are very close and they are in low frequencies range in the spectrum (LOW - PITCHED). Mean values of F1 (455 Hz) and F2 (910 Hz). Back vowels are low-pitched | Fundamental frequency (F0): Vibration frequency of the vocal folds. Variations of F0 may indicate changes in the opening of the glottis and modifications in the rigidity of the mucosa of the vocal folds. Oral opening: the more open the vowel, the higher the frequency of F1. Tongue position: Back of the palate. The more anterior the vowel articulation the higher F2. Tone: Differences in tonality may indicate changes in the structure and tissues of the tongue and the anterior part of the palate. |
| u<br>[-consonant]<br>[+sonorant]<br>[labial]AR:<br>[+rounded]<br>[dorsal]AR:<br>[+low]<br>[+retracted]<br>[+voiced]LR<br>[+continuous] | HIGH and BACK (VELAR) Vowel. The tongue approaches the palate in the back area of the oral cavity higher than the o vowel with a severe timbre. | Complex periodic sound wave. Amplification of some of its HARMONICS when it crosses the supraglottic cavities (RESONANCE). F1 y F2 are very close. They are in low frecuency range in the spectrum (LOW-PITCHED). Mean values of F1 (283 Hz) y F2 (865 Hz) | Fundamental frequency (F0): Vibration frequency of the vocal folds. Variations of F0 may indicate changes in the opening of the glottis and modifications in the rigidity of the mucosa of the vocal folds. Oral opening: the more open the vowel, the higher the frequency of F1. Tongue position: Back of the palate and elevated. The more anterior the vowel articulation the higher F2. Tone: Differences in tonality may indicate changes in the structure and tissues of the tongue and the anterior part of the palate. |

Table 1.3: Main properties of the Spanish vowels.

Different studies such as [22, 23, 64] based on features obtained from the fundamental frequency of the vibration of the vocal folds suggested that these features may be well suited for an early detection of bulbar involvement in ALS.

Based on that, in this work, steady portions of the five Spanish vowels were selected for analysis to provide information regarding changes in the vocal signal which reflected physiologic changes of the vocal folds.

Features obtained from voice utterance are strongly dependant from the native language of the speakers. To achieve accurately analysis is necessary to compare these features from native speakers.

Other sounds like cough or breathing are not related with the native language of the subjects. Concretely, the acoustic characteristics of coughs are more dependant on how the respiratory system is affected.

Summarizing, in this work, to detect bulbar involvement in ALS patients, three different types of analysis on the Spanish vowels were performed. These were: Phonatory subsystem analysis, Time-Frequency and Pattern analysis. Indeed, COVID-19 coughs were detected by means of Time-Frequency analysis. In this case, no distinction between languages was performed.

### 1.4.1 Phonatory subsystem

**Jitter**, **shimmer**, **HNR** and **pitch** are voice features from the phonatory subsystem which can be obtained from sounds. In our case, as was explained above, from the Spanish vowels.

**Jitter** is defined as the periodic variation from cycle to cycle of the fundamental period. Patients with lack of control of the vibration of the vocal folds tend to have higher values of jitter. Some variations of the basic Jitter feature have been computed. These were: Jitter(absolute), Jitter(relative), Jitter(rap) and Jitter(ppq5). Their definitions and the formulas used to obtain them are shown below.

**Jitter(absolute)** is the cycle-to-cycle variation of fundamental period, i.e. the average absolute difference between consecutive periods (eq. 1.1).

$$Jitter(absolute) = \frac{1}{N-1}\sum_{i=1}^{N-1} |T_i - T_{i-1}|, \tag{1.1}$$

where $T_i$ is the duration of the $i$th cycle and $N$ is the total number of cycles.

**Jitter(relative)** is the average absolute difference between consecutive periods, divided by the average period. It is expressed as a percentage (eq. 1.2).

$$Jitter(relative) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100 \tag{1.2}$$

**Jitter(rap)** is defined as the relative average perturbation, the average absolute difference between a period and the average of this and its two neighbors, divided by the average period (eq. 1.3).

$$Jitter(rap) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - \frac{1}{3}\sum_{n=i-1}^{i+1} T_n|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100 \tag{1.3}$$

**Jitter(ppq5)** is the five-point period perturbation quotient, computed as the average absolute difference between a period and the average of this and its four closest neighbors, divided by the average period (eq. 1.4).

$$Jitter(ppq5) = \frac{\frac{1}{N-1}\sum_{i=2}^{N-2}|T_i - \frac{1}{5}\sum_{n=i-2}^{i+2} T_n|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100 \tag{1.4}$$

**Shimmer** is defined as the fluctuation of the waveform amplitudes of consecutive cycles of the fundamental period. A reduction of glottal resistance causes a variation in the magnitude of the glottal period correlated with breathiness and noise emission, causing an increase in shimmer. Some variations of the basic Shimmer feature have been computed. These were: Shimmer(dB), Shimmer(relative), Shimmer(apq3), Shimmer(apq5), and Shimmer(apq11). Their definitions and the formulas used to obtain them are shown below.

**Shimmer(dB)** is expressed as the variability of the peak-to-peak amplitude, defined as the difference between the maximum positive and the maximum negative amplitude of each period, in decibels, i.e. the average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20 (eq. 1.5).

$$Shimmer(dB) = \frac{1}{N-1}\sum_{i=1}^{N-1} |20 \times log(\frac{A_{i+1}}{A_i})|, \qquad (1.5)$$

where $A_i$ is the extracted peak-to-peak amplitude data and $N$ is the number of extracted fundamental periods.

**Shimmer(relative)** is defined as the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude, expressed as a percentage (eq. 1.6).

$$Shimmer(relative) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \times 100 \qquad (1.6)$$

**Shimmer(apq3)** is the three-point amplitude perturbation quotient. This is the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbors, divided by the average amplitude (eq. 1.7).

$$Shimmer(apq3) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |A_i - (\frac{1}{3}\sum_{n=i-1}^{i+1} A_n)|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \times 100 \qquad (1.7)$$

**Shimmer(apq5)** is defined as the five-point amplitude perturbation quotient, or the average absolute difference between the amplitude of a period and the average of the amplitudes of this and its four closest neighbors, divided by the average amplitude (eq. 1.8).

$$Shimmer(apq5) = \frac{\frac{1}{N-1}\sum_{i=2}^{N-2} |A_i - (\frac{1}{5}\sum_{n=i-2}^{i+2} A_n|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \times 100 \qquad (1.8)$$

**Shimmer(apq11)** is expressed as the eleven-point amplitude perturbation quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of this and its ten closest neighbors, divided by the average amplitude (eq. 1.9).

$$Shimmer(apq11) = \frac{\frac{1}{N-1}\sum_{i=5}^{N-5} |A_i - (\frac{1}{11}\sum_{n=i-5}^{i+5} A_n)|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \times 100 \qquad (1.9)$$

**HNR** is defined as the ratio between periodic and non-periodic components of a speech sound. The HNR provides an indication of the overall periodicity of the voice signal by quantifying the ratio between the periodic (harmonic part) and aperiodic (noise) components (see eq. 1.10):

$$HNR = 10 \times log_{10} \frac{r(t=\tau)}{1-r(t=\tau)}, \tag{1.10}$$

where $r(t)$ is the normalized auto-correlation function, $r(t = \tau)$ is the second local maximum of the normalized auto-correlation and $\tau$ is the period of the signal.

The **pitch** is the frequency at which vocal chords vibrate in voiced sounds as vowels. It can be computed by means of the auto-correlation method implemented in [65]. From pitch, some features as mean **pitch(mean)**, standard deviation **pitch(SD)**, minimum **pitch(min)** and maximum **pitch(max)** can be obtained. See [65] for details.

### 1.4.2 Time-Frequency representation

Time-frequency representation (TFR), broadly applied to detecting several malfunction conditions [31, 32, 33, 34], has been recently used to detect pathological changes in voice signals [35]. TFR enables the evolution of the periodicity and frequency components to be observed over time, allowing the analysis of non-stationary signals, such as voice signals [36]. The spectrogram is the most common TFR for the analysis of audio signals. This representation corresponds to Cohen's class of time-frequency energy distributions. The depiction of a spectrogram is not optimal in terms of resolution quality. In general, the Cohen-class representations provide greater resolution quality. They are all made by smoothing the Wigner distribution, which has the finest resolution but the most detrimental interference. The smoothing functions chosen strike a balance between resolution quality and the elimination of detrimental interference terms.

The Wigner distribution (WD) has been used in different fields and applied to the study of time-varying and strongly non-stationary systems. Since the energy is a quadratic representation of the signal, the quadratic structure of the TFR is intuitive and reasonably accepted when the TFR is interpreted as an energy distribution in time and frequency [66]. From all TFRs that represent energy, the WD satisfies many desired mathematical properties. For example, the WD is always real, symmetrical with respect to the time and frequency axes, satisfying the marginal properties and the instantaneous frequency. Furthermore, the group delay may be obtained. Eq. 1.11 represents the WD of the signal *x(t)*.

$$WD(t,f) = \int x(t+\tau/2)x^*(t-\tau/2)e^{-j2\pi f\tau}d\tau, \tag{1.11}$$

where $t$ and $f$ represent time and frequency respectively, and $x^*(t)$ is the conjugate of $x(t)$.

Basically, the WD of a real signal *x(t)* is calculated in a similar way to a convolution. At each particular time, the signal is overlapped by itself and inverted on the time axis, and multiplied by itself. Finally, the Fourier transform of this product is carried out. Note that neither will the WD be necessarily zero when *x(t)* nor would the WD necessarily be zero at frequencies that do not exist in the spectrum. Evidence of this phenomenon has been called interference terms and cross-terms. The interference terms are undesired since they make it difficult to obtain a clear and intuitive spectrum of the signal, as two energy regions perfectly delimited are expected to be obtained.

The possibility of using the WD as a representation of the signal spectral density at each particular time induces the generation of another distribution from the WD to minimise these interference terms while simultaneously maintaining certain properties. To achieve this, the convolution of the WD can be calculated with the Choi–Williams exponential function *h(t,f)* [67] (Eq. 1.12). By convolving the Wigner distribution with the Choi–Williams exponential, the Choi–Williams distribution (CWD) is obtained (Eq. 1.13).

$$h(t,f) = \sqrt{\frac{4\pi}{\sigma_c}}e^{-4\pi^2\frac{(tf)^2}{\sigma_c}}, \tag{1.12}$$

where $\sigma_c$ is a scaling factor.

$$CWD(t,f) = \iint h(t-t', f-f')WD(t',f')\,dt'\,df' \tag{1.13}$$

CWD preserves the properties of WD [33, 67], such as the marginal properties and instantaneous frequency. Moreover, it is able to reduce the WD interference by estimating an adequate $\sigma_c$ parameter. So, the CWD is a new function of the time–frequency distribution that allows the interference terms to be minimised.

Then, in order to obtain statistical parameters, the density function $CWD(f,t)$ can be normalized to have an area equal to 1. So, it can be associated with a joint probability density function $CWD_N(f,t)$ of the time and frequency variables. Their marginal distributions, which do not contain the interference, still represent, although in a normalised manner, the instantaneous power (Eq. 1.14) and spectral density energy (Eq. 1.15) of the original signal.

$$m_t(t) = \int_{-\infty}^{\infty} CWD_N(f,t)df = |x(t)|^2 \tag{1.14}$$

$$m_f(f) = \int_{-\infty}^{\infty} CWD_N(f,t)dt = |X(f)|^2 \tag{1.15}$$

Therefore, the group delay (Eq. 1.16) and the mean frequency of the spectrum (Eq. 1.17) can be defined as:

$$t_g = \iint tCWD_N(t,f)\,dt\,df \tag{1.16}$$

$$f_m = \iint fCWD_N(t,f)\,dt\,df \tag{1.17}$$

The joint time-frequency moments of a non-stationary signal comprise a set of time-varying parameters that characterise the signal spectrum as it evolves over time. They are related to the conditional temporal moments and the joint time-frequency moments. The joint time-frequency moment is an integral function of frequency, given time, and marginal distribution. The conditional temporal moment is an integral function of time, given frequency, and marginal distribution. The calculation of the joint time-frequency moment $t^n f^m$ (Eq. 1.18) is a double integral through time and frequency [68].

$$\langle t^n f^m \rangle = \iint (t - t_g)^n (f - f_m)^m CWD_N(t, f)\, dt\, df \tag{1.18}$$

where $n$ and $m$ are the frequency and time moment orders.

The moments of the marginal density functions, that define the relationship between $m_t(t)$ and $m_f(f)$, $\langle m_t(t)^n m_f(f)^m \rangle$, are given in Eq. 1.19.

$$\langle m_t(t)^n m_f(f)^m \rangle = \frac{1}{std(m_t(t)^n) std(m_f(f)^m)} \iint (m_t(t) - \overline{m_t(t)}))^n (m_f(f) - \overline{m_f(f)})^m dt\, df \tag{1.19}$$

$CWD$ minimises the interference. However, negative values still remain. To solve this issue, the CWD can be reformulated as the product of its marginal distributions. Therefore, the joint probability density distribution $pD$ (Eq. 1.20) is obtained. This procedure is only possible if the marginal distributions of the CWD are statistically independent.

$$pD(f, t) = m_f(f) \cdot m_t(t) \tag{1.20}$$

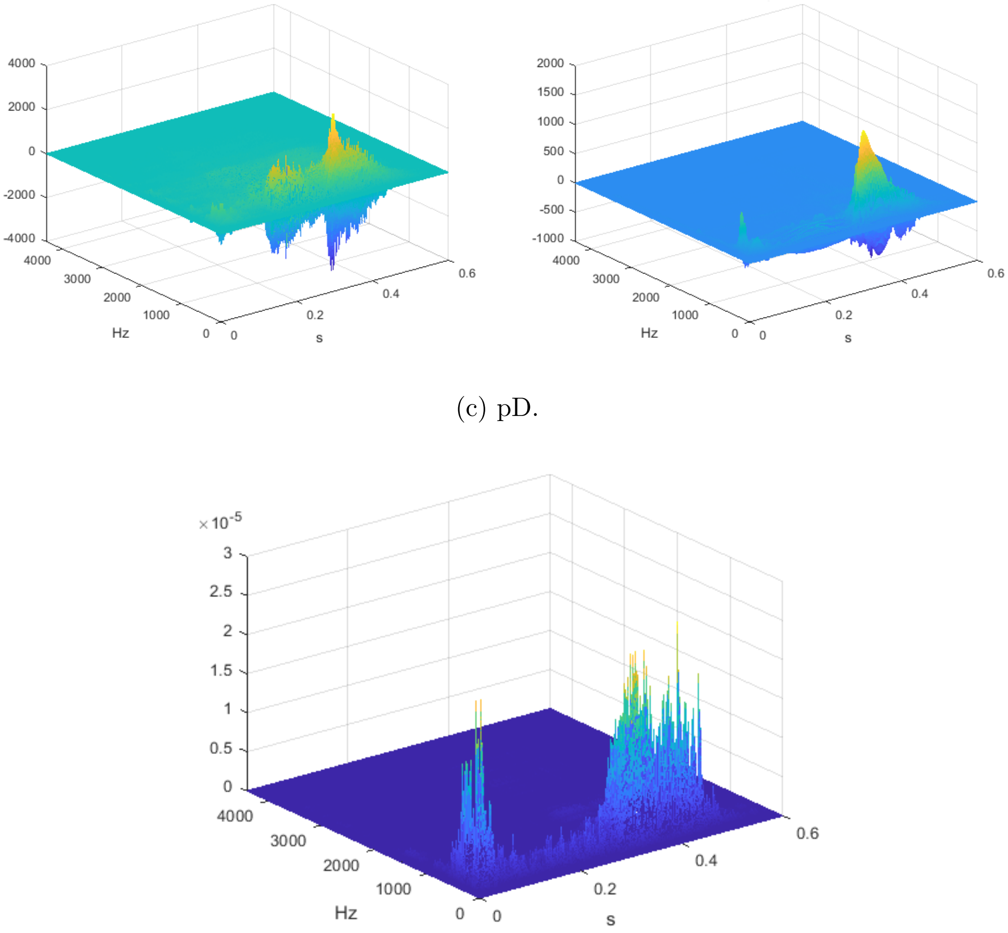
Fig. 1-1(a) corresponds to the WD of a cough sample. It shows how the interference terms of the WD make it difficult to obtain a clear and intuitive spectrum of the signal. The new function $pD(f, t)$ (Fig. 1-1(c)) is equal to WD without interference (CWD, Fig. 1-1(b))) and negative values.

Fig. 1-2(a) corresponds to a sample of the WD of vowel “a”. Fig. 1-2(b)) and Fig. 1-2(c) represent its CWD and pD distributions respectively.

### 1.4.3 Pattern analysis

Quasi-periodic waveform analysis has been applied to several clinical applications such as heartbeat detection, cardiopulmonary modeling and intrinsic brain activity detection [37, 38]. Quasi-periodic waveform are produced by the vocal folds when a vowel is elicited. We conjectured that the diagnosis of some conditions such as bulbar

Figure 1-1: Time-frequency representations of a cough sample.


(a) WD.

(b) $CWD$.

(c) pD.

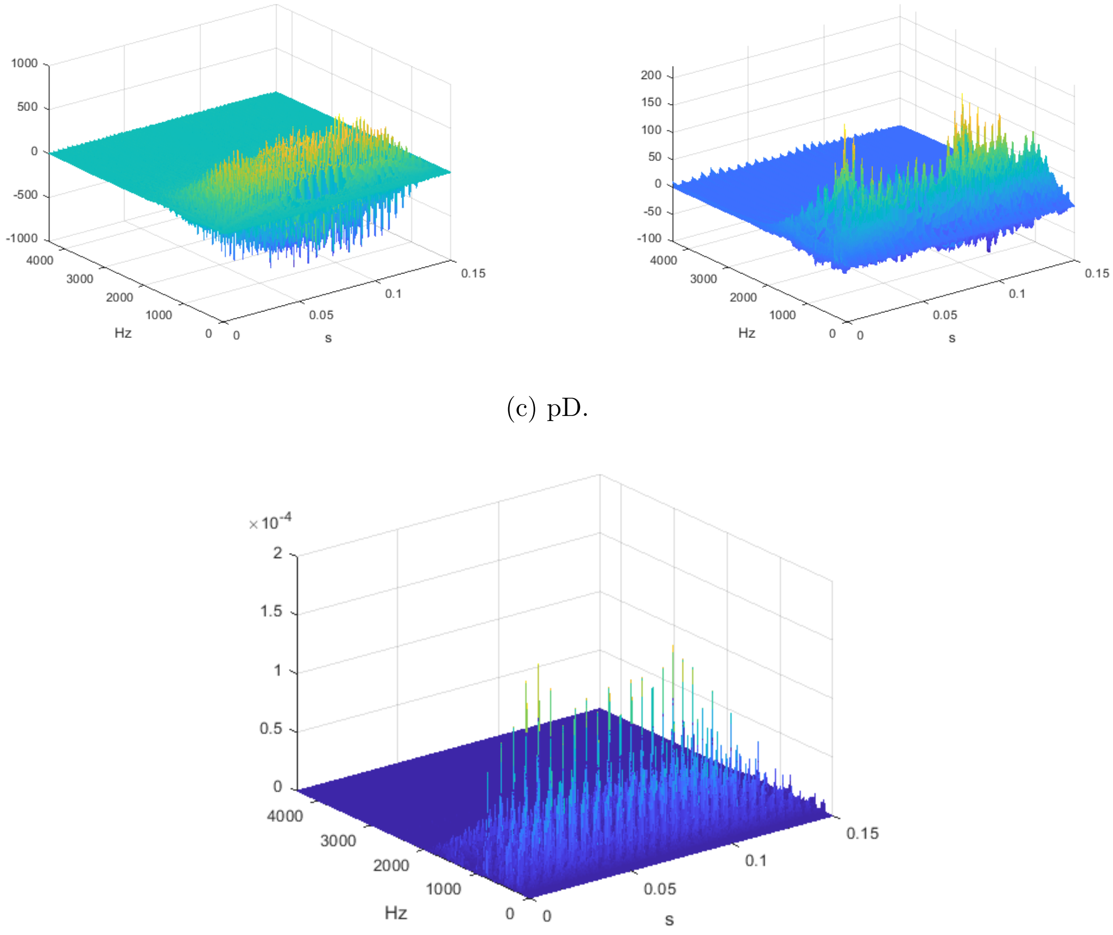
involvement in ALS patients could be greatly benefit from the creation of a voice fingerprint based on the pattern analysis of these quasi-periodic components.

A pattern generator can be developed to obtain a pattern sequence of the quasi-periodic components of the fundamental frequency of a vowel signal $x(n)$. This process consist of 3 steps:

### Step 1. Detrending Method

The baseline wandering of $x(n)$, which is a low-frequency artefact present in signal recordings, can be removed by implementing a detrending method. To obtain the trend, a six-order low-pass butterworth filter [69] can be applied twice (forward and

Figure 1-2: Time-frequency depictions of vowel a for the same patient.



(a) WD.

(b) *CWD*.

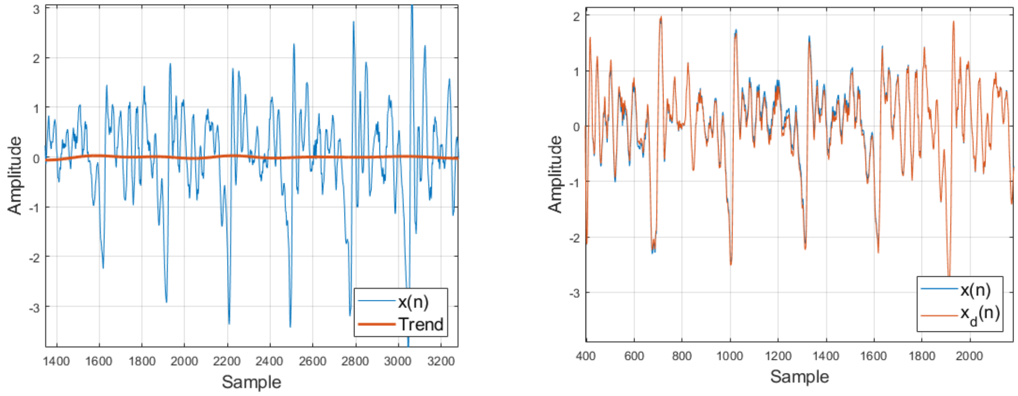
(c) pD.

backward) to $x(n)$ [70, 71]. The combined filter has zero phase distortion, a filter transfer function equal to the squared magnitude of the implemented butterworth filter transfer function, and a filter order that is double the order of the butterworth filter. Then, the detrending signal $x_d(n)$ is obtained by removing the trend from $x(n)$. Fig. 1-3a shows $x(n)$ and the trend of $x(n)$ and Fig. 1-3b shows $x(n)$ and $x_d(n)$.

### Step 2. Marking the quasi-periodic components of $x(n)$

The spectral density $|X_d(f)|^2$ of $x_d(n)$ (Fig. 1-4) can be obtained by means of the discrete Fourier transform (DFT) implementing the fast Fourier transform (FFT) algorithm. To identify the quasi-periods, the peaks of the spectral density can be

Figure 1-3: Detrending method: Removing the trend from $x(n)$.

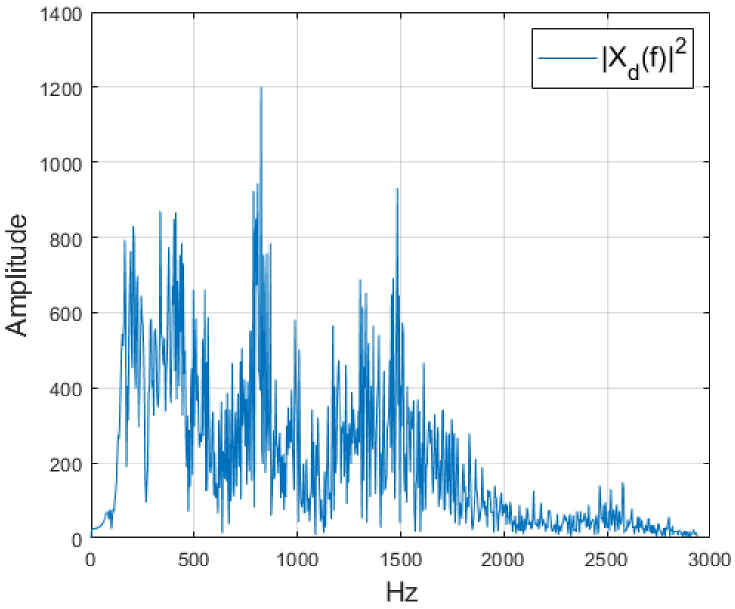


(a) Obtaining the trend of $x(n)$

(b) $x(n)$ and $x_d(n)$

used. To avoid noise, the three higher peaks can be selected. Finally, the quasi-period of $x_d(n)$ can be defined as the lower spectral component of these three peaks ($f_r$). The number of samples of each quasi-period ($n_{rep}$) can be calculated as the nearest integer of $(f_s/f_r)$, being $f_s$ the recording sampling rate.

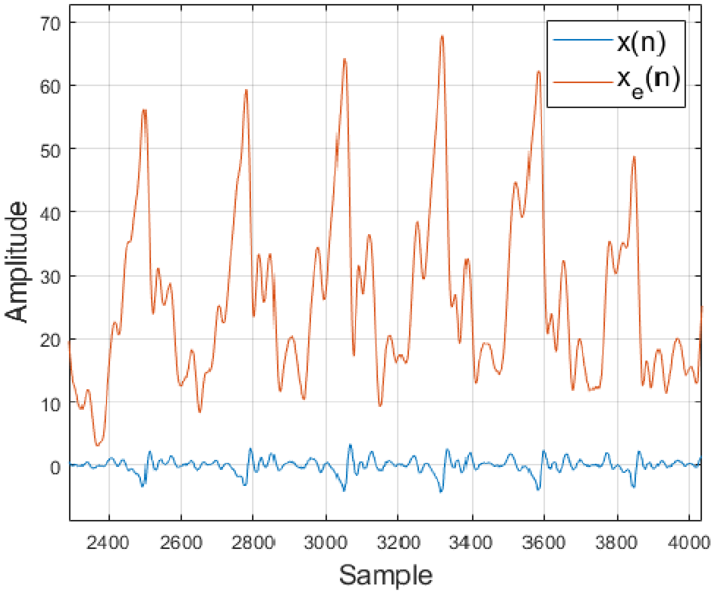
Figure 1-4: The spectral density of $x_d(n)$.



The signal envelop, $x_e(n)$, can be obtained by computing the cumulative sum of $x_d(n)$ and then calculating the envelope of the analytical signal [72]. Fig. 1-5 shows

$x_e(n)$ and $x(n)$.

Figure 1-5: The signal envelope of $x(n)$.

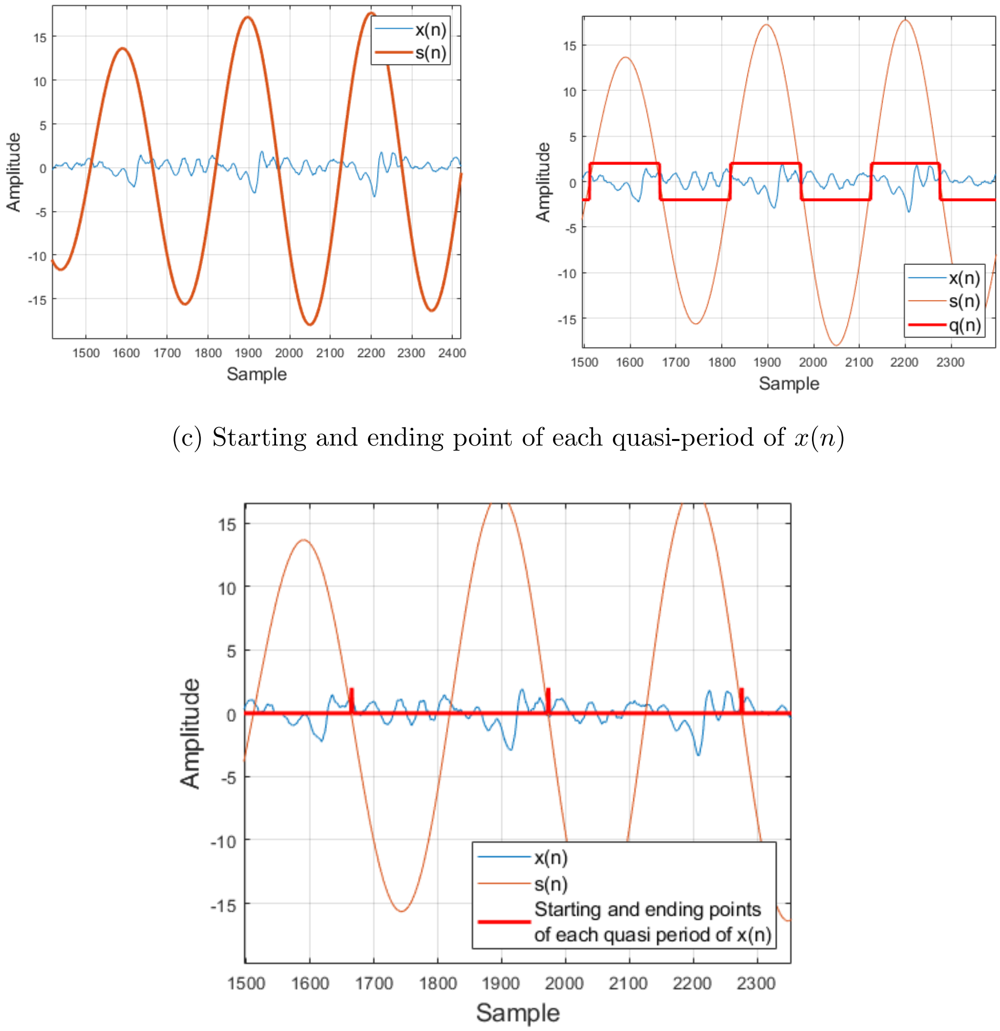


To detect the starting and ending point of each quasi-period, a quasi-sinusoidal signal, $s(n)$, synchronised with the period of $x(n)$ can be computed. It can be obtained by applying a second-order butterworth pass-band filter forward and backward to $x_e(n)$ with a cut-off frequency $f_c = f_s/n_{rep}$ Hz. From $s(n)$, a quadratic-bipolar signal, $q(n)$, is generated assigning a constant -A in those samples where $s(n) < 0$, and A in those where $s(n) > 0$. Thus, by differentiating $q(n)$, the zero crossings of the synchronized signal $s(n)$ are obtained, which represent the beginning and end of each quasi period of $x(n)$. Fig. 1-6 illustrates this process. Fig. 1-6a shows $s(n)$ synchronised with the period of $x(n)$, Fig. 1-6b represents $x(n)$, $s(n)$ and $q(n)$ and Fig. 1-6c depicts the starting and ending points detected of each quasi-period of $x(n)$.

### Step 3. Pattern function

The pattern function $p(T)$ (Fig. 1-7), is obtained as the average of the quasi-periods of $x(n)$ being $T$ the average of the number of samples of the quasi-period of $x(n)$. p(T) is compared with x_d(n) to improve the boundaries of each quasi-period. The

Figure 1-6: Detecting the starting and ending point of each quasi-period of $x(n)$.



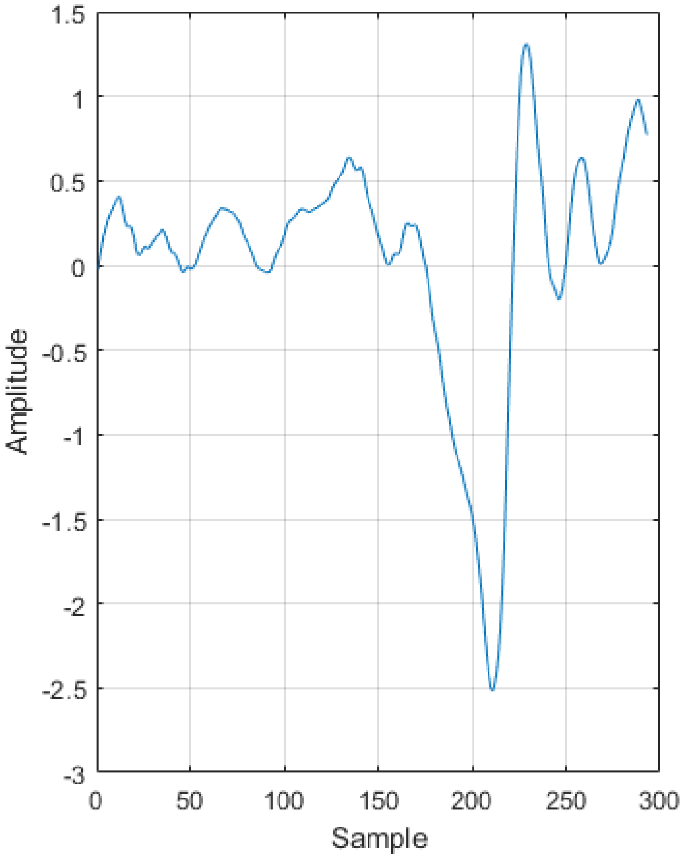
(a) $s(n)$ synchronised with the period of $x(n)$

(b) $x(n)$, $s(n)$ and $q(n)$

(c) Starting and ending point of each quasi-period of $x(n)$

pattern p(T) was inverted and the resulting signal is convolved with x_d(n) to detect the positions of p(T) in x_d(n). The positive values of the resulting signal are taken and the negative values are set at 0. Each quasi-period detected previously is centered in the position where the maximums values of the convolution are found. The refined pattern, p_ref(T), is computed as the average of the quasi-periods of x_d(n) with their new boundaries established. Finally, p_ref(T) is normalized to 550 samples and then decimated to 110 samples to obtain patterns, pN(T), with the same length.

Figure 1-7: Pattern function of vowel a.



## 1.5 Feature selection techniques

Feature selection is the automatic selection of features that are most relevant to a given predictive modeling problem.

### 1.5.1 Multivariate analysis of variance

To select a subset of relevant features for use in the classification model construction, the multivariate analysis of variance (MANOVA), which uses the covariance between the features in testing the statistical significance of the mean differences, was used. This procedure make it possible to contrast the null hypothesis in the features obtained by means of the bio-sounds analysis. The features that rejected the null hypothesis were selected for the model construction.

The selection of the subset of relevant features for constructing the classification models, by using MANOVA, was performed with IBM SPSS Statistics [73].

Table 1.4 shows the MANOVA performed to obtain the statistically significant features when comparing ALS patients with bulbar involvement and controls. The analysis was performed for males. An initial set of 35 time-frequency features was used:

- Average instantaneous spectral energy for each frequency band (E_Bn1. . . E_Bn7).
- Instantaneous frequency peak for each frequency band (f_Cres1 . . . f_Cres7).
- Average instantaneous frequency for each frequency band (f_Med1. . . f_Med7).
- Spectral information for each frequency band (IE_Bn1. . . IE_Bn7).

7 additional features were added:

- Instantaneous (″_t), and spectral (H_f), information entropies. Furthermore, the joint Shannon entropy (H_tf) was also used.
- Kurtosis (K).
- Joint time-frequency moment $\langle t^n f^m \rangle$, where $n$,$m$=1,7,15.

From these 35 features, a set of 6 statistically significant features (p-value<0.05) was obtained.

Table 1.4: Significant features for males.

| **Feature** | **p-value** |
|---|---|
| *f_Cres1* | 0.270 |
| *f_Cres2* | 0.046 |
| *f_Cres3* | 0.429 |
| *f_Cres4* | 0.357 |
| *f_Cres5* | 0.924 |
| *f_Cres6* | 0.046 |
| *f_Cres7* | 0.151 |
| *Enr_Bn1* | 0.461 |
| *Enr_Bn2* | 0.326 |
| *Enr_Bn3* | 0.234 |
| *Enr_Bn4* | 0.831 |
| *Enr_Bn5* | 0.777 |
| *Enr_Bn6* | 0.060 |
| *Enr_Bn7* | 0.274 |
| *f_Med1* | 0.703 |
| *f_Med2* | 0.001 |
| *f_Med3* | 0.559 |
| *f_Med4* | 0.304 |
| *f_Med5* | 0.952 |
| *f_Med6* | 0.008 |
| *f_Med7* | 0.103 |
| *IE_Bn1* | 0.614 |
| *IE_Bn2* | 0.278 |
| *IE_Bn3* | 0.770 |
| *IE_Bn4* | 0.563 |
| *IE_Bn5* | 0.694 |
| *IE_Bn6* | 0.228 |
| *IE_Bn7* | 0.694 |
| *H_tf* | 0.251 |
| *H_t* | 0.147 |
| *H_f* | 0.152 |
| $K$ | 0.027 |
| $\left\langle t^{1} f^{1} \right\rangle$ | 0.002 |
| $\left\langle t^{7} f^{7} \right\rangle$ | 0.900 |
| $\left\langle t^{15} f^{15} \right\rangle$ | 0.870 |

### 1.5.2 Recursive Feature Elimination

The Recursive Feature Elimination (RFE) is a recursive process that ranks features according to some measure of their importance. At each iteration, feature importance is measured and the less relevant one is removed. The recursion is needed because for some measures, the relative importance of each feature can change when evaluated over a different subset of features during the stepwise elimination process. RFE was implemented to select the set of features which obtained the best accuracy for a given classification model.

For example, for an automated COVID-19 cough detection when comparing COVID-19 cough from patients tested positive and patients who presented pertussis cough but who were COVID-19 negative, an initial set of 39 time-frequency features was used. This set of features consisted of the same set as in the previous section but adding 4 new time-frequency features: the average instantaneous frequency $f_{mi}(t)$, and joint time-frequency moment $\langle t^n f^m \rangle$, where $n,m$=1,7,15.

From these 39 features, RF performed better ($Accuracy = 94.81\%$) by previously implementing RFE which selected a set of 16 features to fit the model. These features were *IE_Bn3*, *Enr_Bn4*, *Enr_Bn3*, *IE_Bn2*, *Enr_Bn2*, *f_Med1*, *IE_Bn1*, *f_Med7*, *f_Med4*, *f_Cres1*, *Enr_Bn1*, *IE_Bn6*, *f_Cres2*, *IE_Bn7*, *f_Med2*, *f_Cres6*.

The results for RF worsened if additional features were included or if any of the selected ones were deleted.

## 1.6 Feature extraction techniques

Feature extraction is a process of dimensional reduction by which an initial set of features is reduced while preserving the information in the original dataset.

### 1.6.1 Principal component analysis

The Principal component analysis (PCA) is a ranking feature extraction technique used to decompose an original dataset into principal components (PCs) to obtain another dataset whose data is linearly independent and therefore uncorrelated. It can be performed by means of singular value decomposition (SVD) [74].

By applying SVD to the original standardized dataset, a decomposition is obtained $X = USV^\top$ where $X$ is the matrix of the standardized dataset, $U$ is a unitary matrix and $S$ is the diagonal matrix of singular values $s_i$. PCs are given by $US$ and $V$ contains the directions in this space that capture the maximal variance of the features of the matrix $X$. The number of PCs obtained were the same as the original number of features and the total variance on all the PCs were equal as the total variance among all of the features. So, all of the information contained in the original data is preserved.

Fig. 1-8 is an example of PCA cumulative percentage of the explained variance of a dataset of 15 features.

Figure 1-8: Example of PCA Cumulative percentage of the explained variance.

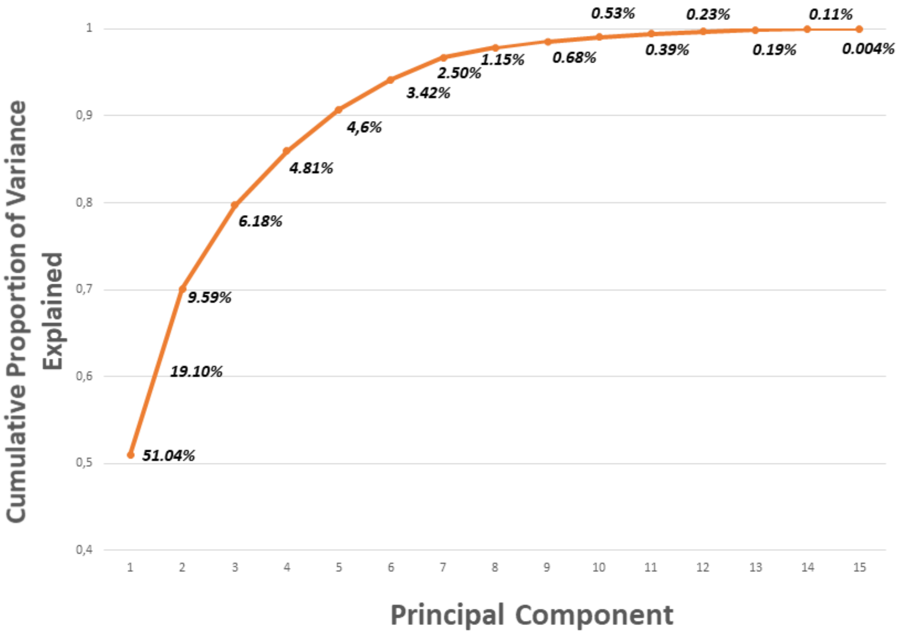
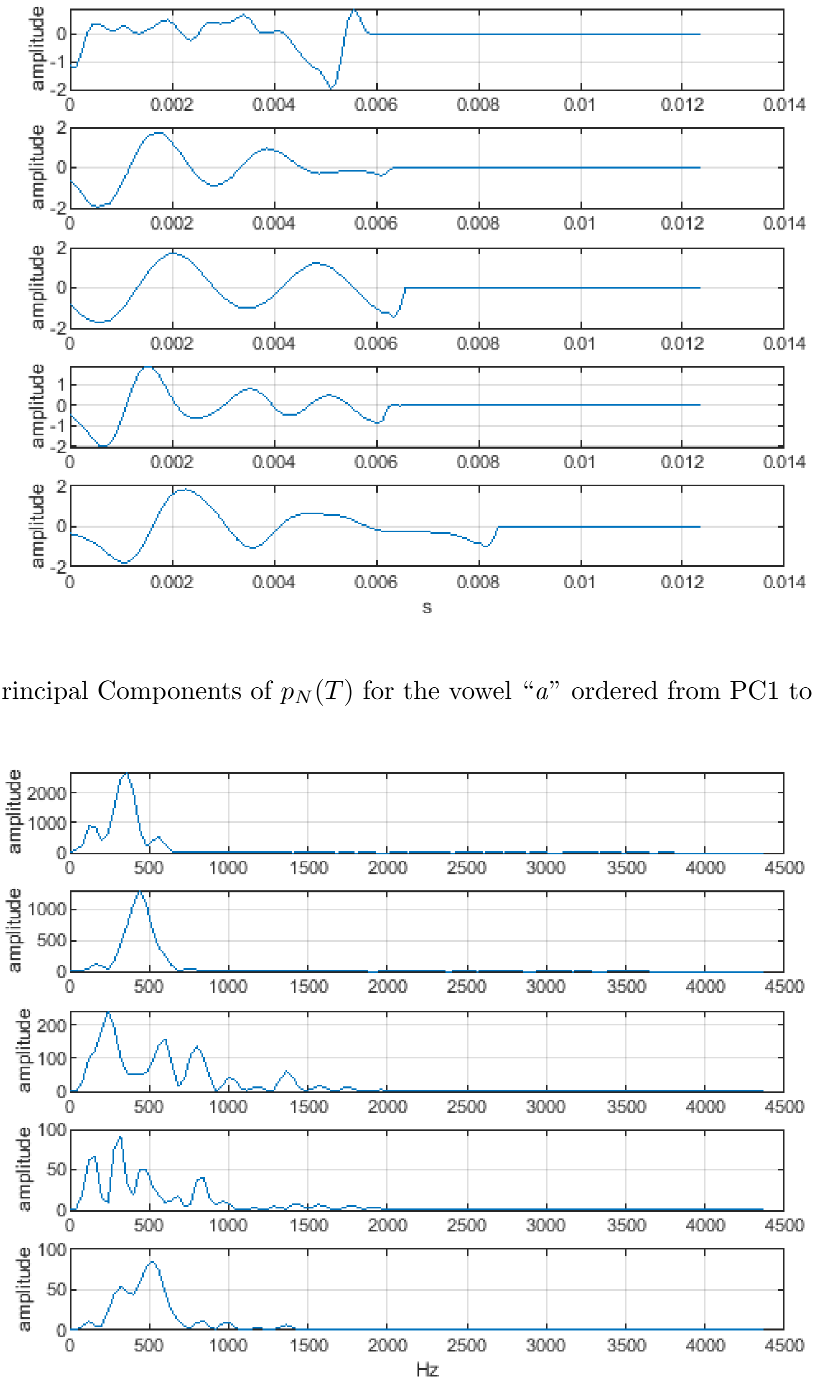


Fig. 1-9 illustrates the $p_N(T)$ the pattern function normalized) of the five Spanish vowels and five PCs of this signal.

Figure 1-9: $p_N(T)$ of the five Spanish vowels and five PCs of $p_N(T)$.



(a) $p_N(T)$ of the five Spanish vowels: a (top), e, i, o, u (bottom)

(b) Principal Components of $p_N(T)$ for the vowel "*a*" ordered from PC1 to PC5

### 1.6.2 Independent component analysis

Independent components analysis (ICA) is a technique of array processing and data analysis, aiming at recovering unobserved signals from observed mixtures, exploiting only the assumption of mutual independence between the signals. ICA is used to reduce the dimensions of an original dataset. Unlike principal components analysis (PCA), which assumes that the components are uncorrelated in both spatial and temporal domains, ICA components are maximally statistically independent in only one domain. The rationale for ICA is that a signal measured can be regarded as a linear combination of a smaller number of independent component sources.

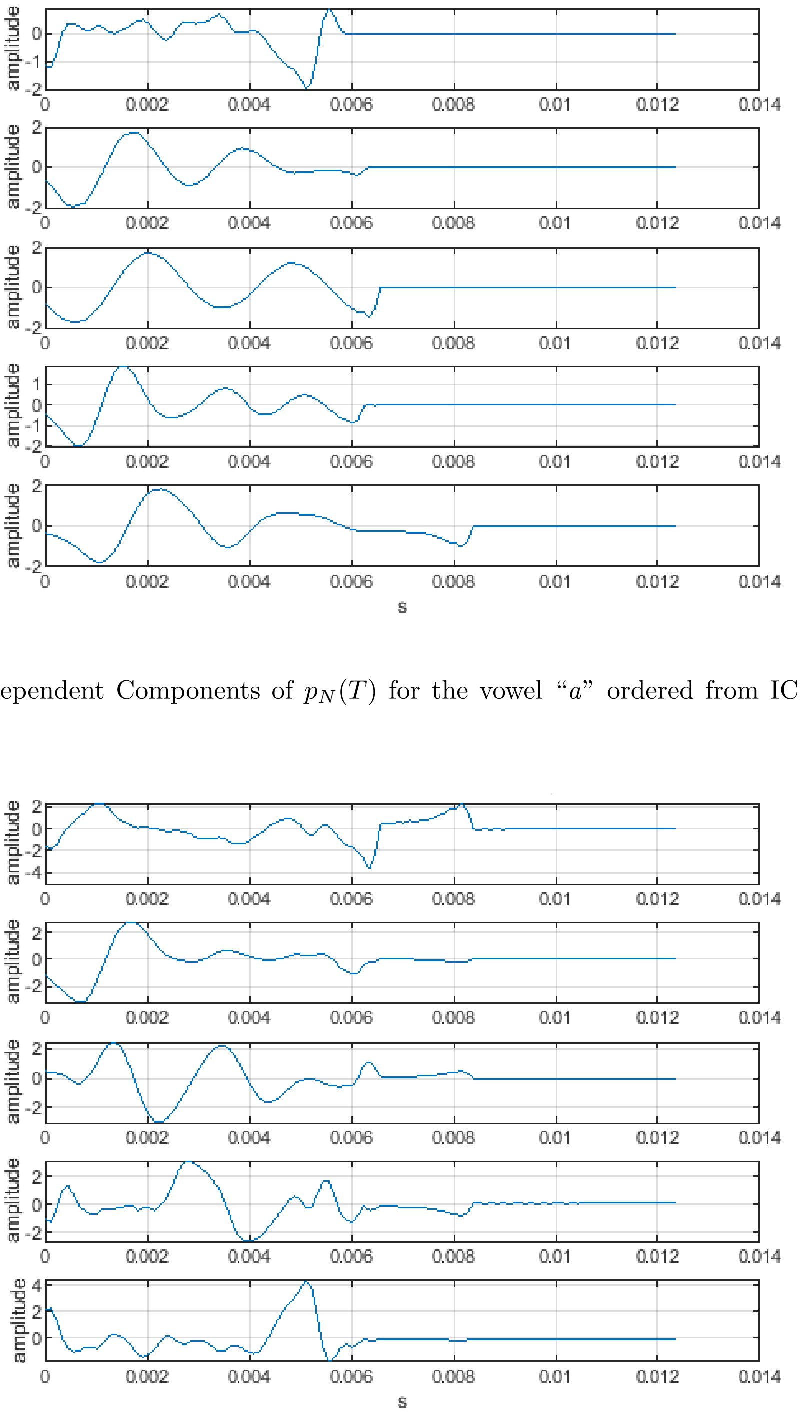
Fig. 1-10 shows the $p_N(T)$ of the five Spanish vowels and five ICs of $p_N(T)$.

### 1.6.3 Autoencoders

An Autoencoder [75] is a specific type of a neural network, one mainly designed to encode the input data into a compressed and meaningful representation, and then decode it back so that the reconstructed input is similar as possible to the original. The Autoencoder maps the input data $x$ to a hidden representation using the function $z = f(Px + b)$ parameterised by $\{P, b\}$. $f$ is the activation function. The hidden representation is then mapped linearly to the output using $\hat{x} = Wz + b'$. The parameters are optimised to minimise the mean square error of $\|\hat{x} - x\|_2^2$ over all training points.

Figure 1-11 shows the Autoencoder architecture employed. It consists of three modules: the encoder, the decoder and the bottleneck. The encoder is formed by an input layer of 39 nodes and two hidden layers of 30 and 20 nodes respectively. The bottleneck has 15 nodes and the decoder consists of two hidden layers of 20 and 30 nodes respectively and an output layer of 39 nodes. The activation function selected was the *tanh* function. As the purpose of our Autoencoder was to reduce the feature range of our original dataset, we took the compressed data contained in the bottleneck layer. So, the 39 original features were reduced to 15.

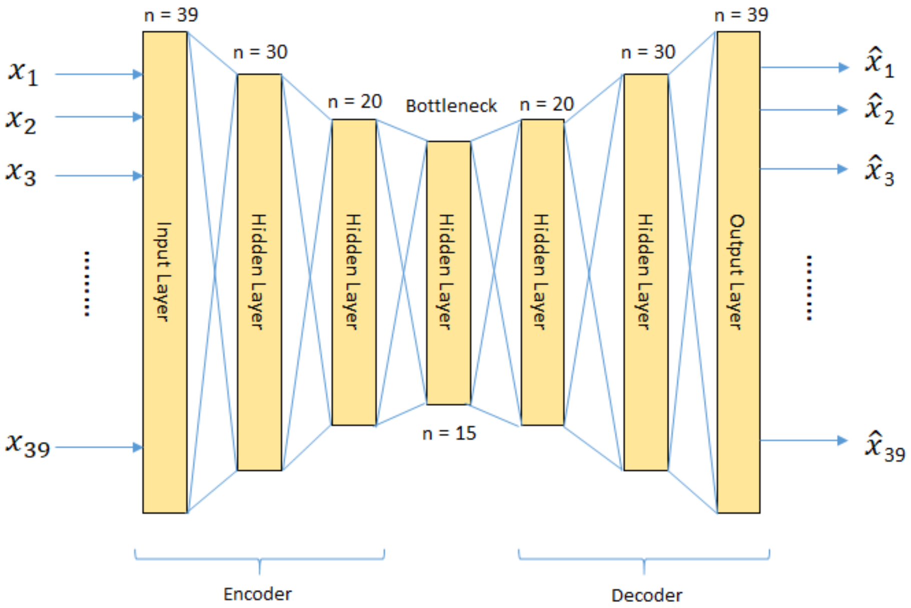
Figure 1-10: $p_N(T)$ of the five Spanish vowels and five ICs of $p_N(T)$.



(a) $p_N(T)$ of the five Spanish vowels: a (top), e, i, o, u (bottom)

(b) Independent Components of $p_N(T)$ for the vowel "*a*" ordered from IC1 to IC5

Figure 1-11: Autoencoder Architecture.



## 1.7 Principal component analysis biplot charts

From the PCA, a biplot chart can be obtained for a visual appraisal of the data. Biplot allowed to visualize the dataset structure, identify the data variability and clustering participants, displaying the variances and correlations of the analyzed features.

Fig. 1-12 shows a PCA biplot chart of a set of features of three different groups of participants (Controls (C), ALS patients without bulbar involvement (NB) and ALS patients with bulbar involvement (B)). The two axes represented the first (Dim1) and second (Dim2) principal components of the data. The biplot uses the diagonalization method to give a graphical display of its dimensional approximation [76, 77]. The interpretation of the biplot involves observing the lengths and directions of the vectors of the features, the data variability and the clusterization of the participants.

## 1.8 Machine-learning models

Various classification algorithms can be used to perform predictions for classification purposes. Table 1.5 summarizes the classification models used in this thesis including those which use supervised and semi-supervised classification training.

Table 1.5: Classification models

| **Model** | **Definition** |
|---|---|
| SVM: Support Vector Machines. | SVM is a powerful, kernel-based classification paradigm.<br>Support vector machines (SVMs) are particular linear classifiers which are based on the margin maximization principle. They perform structural risk minimization, which improves the complexity of the classifier with the aim of achieving excellent generalization performance. The SVM accomplishes the classification task by constructing, in a higher dimensional space, the hyperplane that optimally separates the data into two categories. |
| NN: Neuronal Networks | NNs are networks that utilize complex mathematical models for data processing. A neural network connects simple nodes, also known as neurons or units. And layers of such nodes forms a network of nodes. An array of algorithms are used to identify and recognize relationships in data sets. |
| LDA: Linear Discriminant Analysis | LDA estimates the mean and variance in the training set and computed the covariance matrix to capture the covariance between the groups to make predictions by estimating the probability that the test set belonged to each of the groups. |
| LR: Logistic Regression | LR uses a Gaussian generalized linear model for binomial distributions. A logit link function is used to model the probability of "success". The purpose of the logit link is to take a linear combination of the covariate values and convert those values into a probability scale. |
| NaB: Naive Bayes | NaBs models used Bayes' theorem for classification purposes |
| RF: Random Forest | RF classifier is a combination of tree predictors. Each decision tree performed the classification independently and RF computed each tree predictor classification as one "vote". The majority of the votes computed by all of the tree predictors decided the overall RF prediction. |
| S4VM: Safe Semi-supervised SupportVector Machine | S4VM is a semi-supervised classification model which returns predicted labels for unlabeled instances. It randomly generates multiple low-density separators and merges their predictions by solving a linear programming problem meant to penalize the cost of decreasing the performance of the classifier, compared to the supervised SVM. |

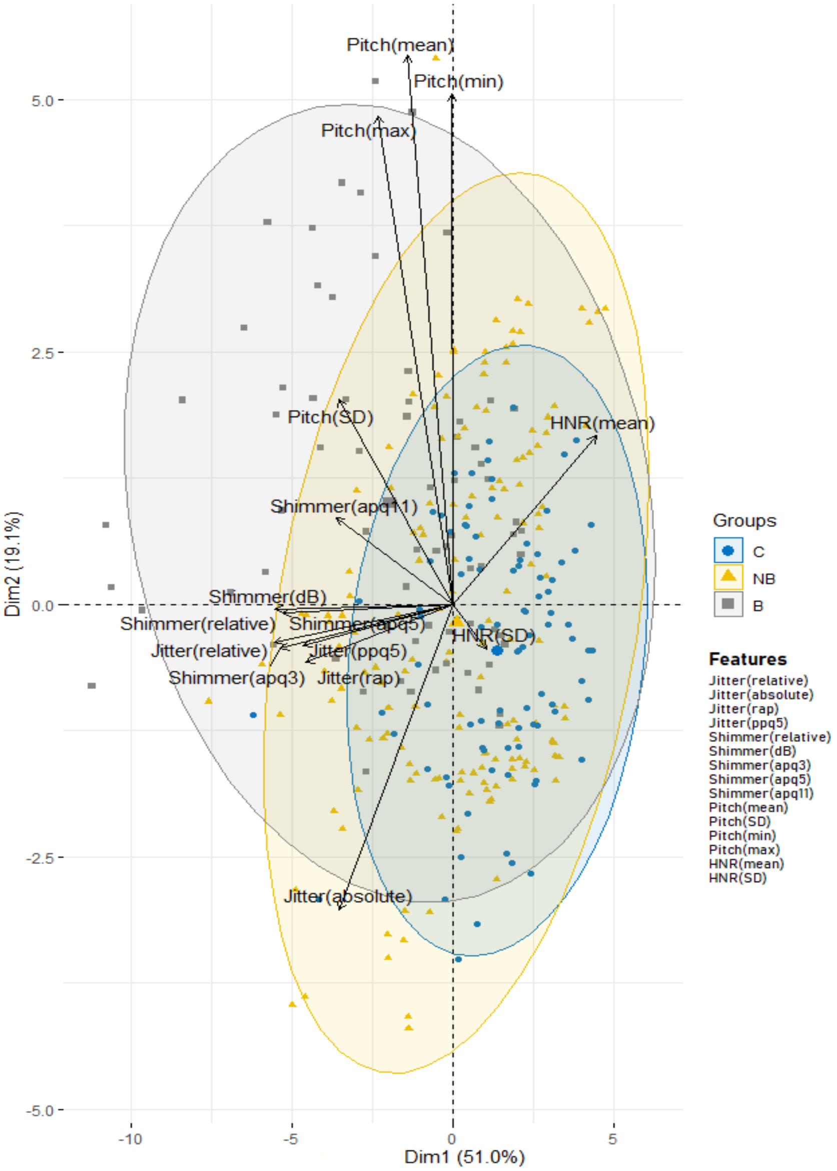


Figure 1-12: PCA biplot chart representing the variance of Dim1 and Dim2 in C, NB and B groups.

In addition to traditional SVM [13, 15, 17, 18, 23], NN [13, 17, 19] and LDA [16], LR is one of the most frequently used model for classification purposes [78, 79], RF [80] is an ensemble method in machine-learning which involves construction of multiple tree predictors that are classic predictive analytic algorithms [23], and Naïve Bayes (NaB) is still a relevant topic [81] and is based on applying Bayes' theorem.

Additionally, semi-supervised models such as the Safe Semi-supervised Support

Vector Machine (S4VM) can be used to curate the datasets. It predicts labels for unlabeled instances.

## 1.9 Performance metrics

There are four possible results in the classification task: If the sample is positive and it is classified as positive, it is counted as a *true positive* (TP) and when classified as negative, it is considered a *false negative* (FN). If the sample is negative and is classified as negative or positive, it is considered a *true negative* (TN) or *false positive* (FP) respectively.

Based on that, the Accuracy, Sensitivity (also known as recall), Specificity, Precision and F-score metrics [82] are the most relevant metrics used to evaluate the performance of the classification models. The AUC is also useful.

- **Accuracy** (Eq. 1.21). Ratio between the correctly classified samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1.21}$$

- **Sensitivity** (Eq. 1.22). Proportion of correctly classified positive samples compared to the total number of positive samples.

$$Sensitivity = \frac{TP}{TP + FN} \tag{1.22}$$

- **Specificity** (Eq. 1.23). Proportion of correctly classified negative samples compared to the total number of negative samples.

$$Specificity = \frac{TN}{TN + FP} \tag{1.23}$$

- **Precision** (Eq. 1.24). Proportion of positive samples that were correctly classified compared to the total number of positive predicted samples.

$$Precision = \frac{TP}{FP + TP} \tag{1.24}$$

- **F-score** (Eq. 1.25). Harmonic mean of the precision and sensitivity.

$$F\text{-}score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{1.25}$$

- **AUC** (Eq. 1.26). The Receiver operating characteristics (ROC) curve is a two-dimensional graph in which Sensitivity is plotted on the y-axis and $1 - Specificity$ is plotted on the x-axis. The points of the curve are obtained by sweeping the classification threshold from the most positive classification value to the most negative. The AUC score is a scalar value that measures the area under the ROC curve and is always bounded between 0..1.

$$AUC = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} 1_{p_i > p_j}, \tag{1.26}$$

  where i runs over all m samples with true label positive, and j runs over all n samples with true label negative; $p_i$ and $p_j$ denote the probability score assigned by the classifier to sample $i$ and $j$, respectively.

## 1.10 Datasets

### 1.10.1 ALS Dataset

The study was approved by the Research Ethics Committee for Biomedical Research Projects (CEIm) at the Bellvitge University Hospital in Barcelona. 45 ALS participants (26 males and 19 females) aged from 37 to 84 (M = 60.31 years, SD = 11.74 years) and 18 control subjects (9 males and 9 females) aged from 21 to 68 (M = 45.2 years, SD = 12.2 years) took part in this transversal study. All ALS participants were diagnosed by a neurologist. Bulbar involvement was diagnosed by following subjective clinical approaches [83] and the neurologist made the diagnosis of whether an ALS patient had bulbar involvement. Among all ALS participants, 5 reported bulbar onset and 40 spinal onset, but, at the time of the study, 14 of them presented bulbar symptoms. Summarizing, of the 63 participants, 14 were ALS participants diagnosed with bulbar involvement, 31 were ALS participants that did not present this dysfunction and 18 were control subjects. The severity of ALS and its bulbar presentation also varied among participants, as assessed by the ALS Functional Rating Scale-Revised (ALSFRS-R). The ALSFRS-R score (0-48) was obtained from 12 survey questions that assess the degree of functional impairment with the score of each question ranging from 4 – least impaired to 0 – most impaired. The scores of the 45 participants in this study ranged from 6 to 46, with a mean of 31.38, SD of 8.67 and 3 reported as not available.

### 1.10.2 COVID-19 Datasets

This section describes the data collection framework used in this work. It consisted of the COVID-19 dataset the University of Lleida collected for this study which was approved by the Research Ethics Committee for Biomedical Research Projects (CEIm) at the University Hospital Arnau de Vilanova of Lleida, and three additional existing publicly available COVID-19 datasets, namely University of Cambridge [84], Coswara [56] and Virufy [57] datasets. Additionally, the Pertussis dataset [8], which includes recordings of patients with pertussis cough, was also used.

#### University of Lleida (UdL) Dataset

Approved by the Research Ethics Committee for Biomedical Research Projects (CEIm) at the University Hospital Arnau de Vilanova of Lleida, we began an initiative consisting of a website for recording cough samples (https://covid.udl.cat) for COVID-19 cough discrimination. We collected variable length cough audio recordings (3 cough samples per subject on average) accompanied by a set of 3 questions related to the diagnosis of the disease and general subject information: age, sex; COVID-19 positive by antigen or PCR test; and diagnosed with a neurological or chronic respiratory disease. At the time of this study, 52 subjects registered their coughs, with 49 being tested COVID-19 positive and 3 negative.

### University of Cambridge (UC) Dataset

For this study, we used the publicly available University of Cambridge Dataset [52] consisting of 142 subjects tested COVID-19 positive, and 137 healthy without symptoms and 53 healthy subjects who presented cough symptoms.

### Coswara Dataset

The Coswara project [56] includes vowel records of 1,107 healthy, 107 COVID-19 positive subjects and 48 subjects who reported other non-COVID respiratory conditions. These conditions were not specified at the time of this study.

### Virufy Dataset

The data collected in the Virufy project can be freely downloaded from a github website [57]. The dataset is made up of 73 cough recordings from healthy subjects and 48 from COVID-19 positive subjects. All participants were given PCR tests before the coughs were obtained.

### Pertussis Dataset

This dataset [8] includes 20 recordings of patients with pertussis cough. It was used to compare the patterns of COVID-19 coughs with the patterns of pertussis cough. At the time of this study, the demographic data of the patients whose pertussis cough was recorded was not available.

## 1.11 Publications

The following publications in research journals are derived from the work in this thesis:

- Alberto Tena, Francesc Claria, Francesc Solsona, Einar Meister, Mònica Povedano. Detection of Bulbar Involvement in Patients With Amyotrophic Lateral Sclerosis by Machine Learning Voice Analysis: Diagnostic Decision Support Development Study. JMIR Medical Informatics 9(3):e21331. 2021. doi:10.2196/21331.
- Alberto Tena, Francesc Clarià, Francesc Solsona. Automated detection of COVID-19 cough. Biomedical Signal Processing and Control, Volume 71(A): 103175. 2022. doi:10.1016/j.bspc.2021.103175.
- Alberto Tena, Francesc Clarià, Francesc Solsona, Mònica Povedano. Detecting Bulbar Involvement in Patients with Amyotrophic Lateral Sclerosis Based on Phonatory and Time-Frequency Features. Sensors 22(3):1137. 2022. doi.org/10.3390/s22031137.
- Voice Fingerprint and Machine Learning Models for Early Detection of Bulbar Dysfunction in ALS. Submitted to the journal *Artificial Intelligence in Medicine.*

In addition, as a consequence of the collaboration done with other researchers during my PhD, I have co-authored an additional paper (not included in this thesis):

- Marc Pifarré, Alberto Tena, Francisco Clarià, Francesc Solsona, Jordi Vilaplana, Arnau Benavides, Lluis Mas, Francesc Abella. A Machine-Learning Model for Lung Age Forecasting by Analyzing Exhalations. Sensors 22(3):1106. 2022. doi.org/10.3390/s22031106.

## 1.12 Three-month doctoral stay

During the course of the present thesis, I spent a three-month doctoral stay at TalTech the Tallinn University of Technology. My supervisor was Prof. Einar Meister from the Department of Software Science.

The research activities done at the Laboratory of Language Technology consisted of developing a consistent ALS Corpus from the audios recorded. The work included tasks of segmentation and label properly all the sounds registered. Then, feature extraction algorithms were developed by using PRAAT [65] which is a free computer software package for speech analysis in phonetics. The analysis of the features obtained was performed to decide which kinds of analysis and which sort of models were the most appropriated to distinguish bulbar involvement in ALS patients considering the ALS corpus available.

The first article of this thesis is the result of this stay:

- Alberto Tena, Francesc Claria, Francesc Solsona, Einar Meister, Mònica Povedano. Detection of Bulbar Involvement in Patients With Amyotrophic Lateral Sclerosis by Machine Learning Voice Analysis: Diagnostic Decision Support Development Study. JMIR Med Inform 9(3):e21331. 2021. doi:10.2196/21331.

# Chapter 2

# Papers

## 2.1 Paper 1: Detection of Bulbar Involvement in Patients With Amyotrophic Lateral Sclerosis by Machine Learning Voice Analysis: Diagnostic Decision Support Development Study

- *Alberto Tena, Francesc Clarià, Francesc Solsona, Einar Meister, Mònica Povedano. Detection of Bulbar Involvement in Patients With Amyotrophic Lateral Sclerosis by Machine Learning Voice Analysis: Diagnostic Decision Support Development Study. JMIR Med Inform 9(3):e21331. 2021. doi:10.2196/21331.*

### Abstract

This paper suggested that the acoustic parameters obtained through automated signal analysis from a steady portion of sustained vowels may be used efficiently as predictors for the early detection of bulbar involvement in patients with ALS. For that purpose, the main objectives (and contributions) of this research were:

- To design a methodology for diagnosing bulbar involvement efficiently through the acoustic parameters of uttered vowels in Spanish.

- To demonstrate that the performance of the automated diagnosis of bulbar involvement is superior to human diagnosis.

To fulfill these objectives, 45 Spanish patients with ALS and 18 control subjects took part in the study. They were recruited by a neurologist, and the five Spanish vowel segments were elicited from each participant.

The study focused on the extraction of features from the phonatory subsystem—jitter, shimmer, harmonics-to-noise ratio, and pitch—from the utterance of the five Spanish vowels. Then, various supervised classification algorithms were used, preceded by principal component analysis of the features obtained.

Support vector machines performed better (accuracy 95.8%) than the models analyzed in the literature. It was also proved how the model can improve human diagnosis, which can often misdiagnose bulbar involvement.

Original Paper

# Detection of Bulbar Involvement in Patients With Amyotrophic Lateral Sclerosis by Machine Learning Voice Analysis: Diagnostic Decision Support Development Study

Alberto Tena[1], MSc; Francec Claria[2], PhD; Francesc Solsona[2], PhD; Einar Meister[3], PhD; Monica Povedano[4], PhD

[1]Information and Communication Technologies Group, International Centre for Numerical Methods in Engineering, Barcelona, Spain

[2]Department of Computer Science, Universitat de Lleida, Lleida, Spain

[3]Institute of Cybernetics, Tallinn University of Technology, Tallinn, Estonia

[4]Motoneuron Functional Unit, Hospital Universitari de Bellvitge, Barcelona, Spain

**Corresponding Author:**
Francesc Solsona, PhD
Department of Computer Science
Universitat de Lleida
Jaume II, 69
Lleida
Spain
Phone: 34 973702735
Email: francesc.solsona@udl.cat

## Abstract

**Background:** Bulbar involvement is a term used in amyotrophic lateral sclerosis (ALS) that refers to motor neuron impairment in the corticobulbar area of the brainstem, which produces a dysfunction of speech and swallowing. One of the earliest symptoms of bulbar involvement is voice deterioration characterized by grossly defective articulation; extremely slow, laborious speech; marked hypernasality; and severe harshness. Bulbar involvement requires well-timed and carefully coordinated interventions. Therefore, early detection is crucial to improving the quality of life and lengthening the life expectancy of patients with ALS who present with this dysfunction. Recent research efforts have focused on voice analysis to capture bulbar involvement.

**Objective:** The main objective of this paper was (1) to design a methodology for diagnosing bulbar involvement efficiently through the acoustic parameters of uttered vowels in Spanish, and (2) to demonstrate that the performance of the automated diagnosis of bulbar involvement is superior to human diagnosis.

**Methods:** The study focused on the extraction of features from the phonatory subsystem—jitter, shimmer, harmonics-to-noise ratio, and pitch—from the utterance of the five Spanish vowels. Then, we used various supervised classification algorithms, preceded by principal component analysis of the features obtained.

**Results:** To date, support vector machines have performed better (accuracy 95.8%) than the models analyzed in the related work. We also show how the model can improve human diagnosis, which can often misdiagnose bulbar involvement.

**Conclusions:** The results obtained are very encouraging and demonstrate the efficiency and applicability of the automated model presented in this paper. It may be an appropriate tool to help in the diagnosis of ALS by multidisciplinary clinical teams, in particular to improve the diagnosis of bulbar involvement.

***(JMIR Med Inform 2021;9(3):e21331)*** doi: 10.2196/21331

**KEYWORDS**

amyotrophic lateral sclerosis; bulbar involvement; voice; diagnosis; machine learning

## Introduction

### Background

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease with an irregular and asymmetric progression, characterized by a progressive loss of both upper and lower motor neurons that leads to muscular atrophy, paralysis, and death, mainly from respiratory failure. The life expectancy of patients with ALS is between 3 and 5 years from the onset of symptoms. ALS produces muscular weakness and difficulties of mobility, communication, feeding, and breathing, making the patient heavily dependent on caregivers and relatives and generating significant social costs. Currently, there is no cure for ALS, but early detection can slow the disease progression [1].

The disease is referred to as spinal ALS when the first symptoms appear in the arms and legs (limb or spinal onset; 80% of cases) and bulbar ALS when it begins in cranial nerve nuclei (bulbar onset; 20% of cases). Patients with the latter form tend to have a shorter life span because of the critical nature of the bulbar muscle function that is responsible for speech and swallowing. However, 80% of all patients with ALS experience dysarthria, or unclear, difficult articulation of speech [2]. On average, speech remains adequate for approximately 18 months after the first bulbar symptoms appear [3]. These symptoms usually become noticeable at the beginning of the disease in bulbar ALS or in later stages of spinal ALS. Early identification of bulbar involvement in people with ALS is critical for improving diagnosis and prognosis and may be the key to effectively slowing progression of the disease. However, there are no standardized diagnostic procedures for assessing bulbar dysfunction in ALS.

Speech impairment may begin up to 3 years prior to diagnosis of ALS [3], and as ALS progresses over time there is significant deterioration in speech [4]. Individuals with ALS with severe dysarthria present specific speech production characteristics [5-7]. However, it is possible to detect early, often imperceptible, changes in speech and voice through objective measurements, as suggested in previous works [8-11]. The authors concluded that phonatory features may be well suited to early ALS detection.

## Related Work

Previous speech production studies have revealed significant differences in specific acoustic parameters in patients with ALS. Carpenter et al [7] studied the articulatory subsystem of individuals with ALS and found different involvement of articulators—that is, the tongue function was more involved than the jaw function. In a recent study, Shellikeri et al [5] found that the maximum speed of tongue movements and their duration were only significantly different at an advanced stage of bulbar ALS compared with the healthy control group. Connaghan et al [12] used a smartphone app to identify and track speech decline. Lee et al [6] obtained acoustic patterns for vowels in relation to the severity of the dysarthria in patients with ALS.

Other works have demonstrated the efficiency of features obtained from the phonatory subsystem for detecting early deterioration in ALS [8-11,13-15]. Studies have shown significant differences between jitter, shimmer, and the harmonics-to-noise ratio (HNR) in patients with ALS [8,10,11]. More specifically, Silbergleit et al [8] obtained these features from a steady portion of sustained vowels that provided information regarding changes in the vocal signal that are believed to reflect physiologic changes of the vocal folds. Alternative approaches used formant trajectories to classify the ALS condition [13], correlating formants with articulatory patterns [14], fractal jitter [15], Mel Frequency Cepstral Coefficients (MFCCs) [16], or combined acoustic and motion-related features [9] at the expense of introducing more invasive measurements to obtain data. Besides, the findings revealed significant differences in motion-related features only at an advanced stage of bulbar ALS.

Other related studies, such as one by Frid et al [17], used speech formants and their ratios to diagnose neurological disorders. Teixeira et al [18] and Mekyska et al [19] suggested jitter, shimmer, and HNR as good parameters to be used in intelligent diagnosis systems for dysphonia pathologies.

Garcia-Gancedo et al [20] demonstrated the feasibility of a novel digital platform for remote data collection of digital speech characteristics, among other parameters, from patients with ALS.

In the literature, classification models are widely used to test the performance of acoustic parameters in the analysis of pathological voices. Norel et al [21] identified acoustic speech features in naturalistic contexts and machine learning models developed for recognizing the presence and severity of ALS using a variety of frequency, spectral, and voice quality features. Wang et al [9] explored the classification of the ALS condition using the same features with support vector machine (SVM) and neuronal network (NN) classifiers. Rong et al [22] used SVMs with two feature selection techniques (decision tree and gradient boosting) to predict the intelligible speaking rate from speech acoustic and articulatory samples.

Suhas et al [16] implemented SVMs and deep neuronal networks (DNNs) for automatic classification by using MFCCs. An et al [23] used convolutional neuronal networks (CNNs) to compare the intelligible speech produced by patients with ALS to that of healthy individuals. Gutz et al [24] merged SVM and feature filtering techniques (SelectKBest). In addition, Vashkevich et al [25] used linear discriminant analysis (LDA) to verify the suitability of the sustain vowel phonation test for automatic detection of patients with ALS.

Among feature extraction techniques, principal component analysis (PCA) [26] shows good performance in a wide range of domains [27,28]. Although PCA is an unsupervised technique, it can efficiently complement a supervised classifier in order to achieve the objective of the system. In fact, any classifier can be used in conjunction with PCA because it does not make any kind of assumption about the subsequent classification model.

## Hypothesis

Based on previous works, our paper suggests that the acoustic parameters obtained through automated signal analysis from a steady portion of sustained vowels may be used efficiently as predictors for the early detection of bulbar involvement in patients with ALS. For that purpose, the main objectives (and contributions) of this research were (1) to design a methodology for diagnosing bulbar involvement efficiently through the acoustic parameters of uttered vowels in Spanish; and (2) to demonstrate that the performance of the automated diagnosis of bulbar involvement is superior to human diagnosis.

To fulfill these objectives, 45 Spanish patients with ALS and 18 control subjects took part in the study. They were recruited by a neurologist, and the five Spanish vowel segments were elicited from each participant. The study focused on the extraction of features from the phonatory subsystem—jitter, shimmer, HNR, and pitch—from the utterance of each Spanish vowel.

Once the features were obtained, we used various classification algorithms to perform predictions based on supervised classification. In addition to traditional SVMs [9,16,21,22,24], NNs [9,16,23], and LDA [25], we used logistic regression (LR), which is one of the most frequently used models for classification purposes [29,30]; random forest (RF) [31], which is an ensemble method in machine learning that involves the construction of multiple tree predictors that are classic predictive analytic algorithms [22]; and naïve Bayes (NaB), which is still a relevant topic [32] and is based on applying Bayes' theorem.

Prior to feeding the models, PCA was applied to the features obtained due to the good performance observed of this technique in a wide range of domains.

## Methods

### Participants

The study was approved by the Research Ethics Committee for Biomedical Research Projects (CEIm) at the Bellvitge University Hospital in Barcelona, Spain. A total of 45 participants with ALS (26 males and 19 females) aged from 37 to 84 (mean 57.8, SD 11.8) years and 18 control subjects (9 males and 9 females) aged from 21 to 68 (mean 45.2, SD 12.2) years took part in this transversal study. All participants with ALS were diagnosed by a neurologist.

Bulbar involvement was diagnosed by following subjective clinical approaches [33], and the neurologist made the diagnosis of whether a patient with ALS had bulbar involvement. Of the 45 participants with ALS, 5 reported bulbar onset and 40 reported spinal onset, but at the time of the study 14 of them presented bulbar symptoms.

To summarize, of the 63 participants in the study, 14 were diagnosed with ALS with bulbar involvement (3 males and 11 females; aged from 38 to 84 years, mean 56.8 years, SD 12.3 years); 31 were diagnosed with ALS but did not display this dysfunction (23 males and 8 females, aged from 37 to 81 years, mean 58.3 years, SD 11.7 years); and 18 were control subjects (9 males and 9 females; aged from 21 to 68 years, mean 45.2 years, SD 12.2 years).

The severity of ALS and its bulbar presentation also varied among participants, as assessed by the ALS Functional Rating Scale-Revised (ALSFRS-R). The ALSFRS-R score (0-48) was obtained from 12 survey questions that assess the degree of functional impairment, with the score of each question ranging from 4 (least impaired) to 0 (most impaired). The scores of the 45 participants in this study ranged from 6 to 46 (mean 31.3, SD 8.6; 3 patients' scores were reported as not available). Within the subgroups, the scores of patients diagnosed with bulbar involvement ranged from 6 to 46 (mean 23.1, SD 9.8), and the scores of participants with ALS who did not present this dysfunction ranged from 17 to 46 (mean 30.2, SD 8.0; 3 patients' scores reported as not available).

The main clinical records of the participants with ALS are summarized in Multimedia Appendix 1.

### Vowel Recording

The Spanish phonological system includes five vowel segments—a, e, i, o, and u. These were obtained and analyzed from each patient with ALS and each control participant, all of whom were Spanish speakers.

Sustained samples of the Spanish vowels a, e, i, o, and u were elicited under medium vocal loudness conditions for 3-4 s. The recordings were made in a regular hospital room using a USB GXT 252 Emita Streaming Microphone (Trust International BV) connected to a laptop. The speech signals were recorded at a sampling rate of 44.100 Hz and 32-bit quantization using Audicity, an open-source application [34].

### Feature Extraction

Each individual phonation was cut out and anonymously labeled. The boundaries of the speech segments were determined with an oscillogram and a spectrogram using the Praat manual [35] and were audibly checked. The starting point of the boundaries was established as the onset of the periodic energy in the waveform observed in the oscillogram and checked by the apparition of the formants in the spectrogram. The end point was established as the end of the periodic oscillation when a marked decrease in amplitude in the periodic energy was observed. It was also identified by the disappearance of the waveform in the oscillogram and the formants in the spectrogram.

Acoustic analysis was done by taking into account the following features: jitter, shimmer, HNR, and pitch. Once the phonations of each participant had been segmented, the parameters were obtained from each vowel through the standard methods used in Praat [35]; they are explained in detail in this section and consist of a short-term spectral analysis and an autocorrelation method for periodicity detection.

Jitter and shimmer are acoustic characteristics of voice signals. Jitter is defined as the periodic variation from cycle to cycle of the fundamental period, and shimmer is defined as the fluctuation of the waveform amplitudes of consecutive cycles. Patients with lack of control of the vibration of the vocal folds tend to have higher values of jitter. A reduction of glottal resistance causes a variation in the magnitude of the glottal period correlated with breathiness and noise emission, causing an increase in shimmer [18].

To compute jitter parameters, some optional parameters in Praat were established. Period floor and period ceiling, defined as the minimum and maximum durations of the cycles of the waveform that were considered for the analysis, were set at 0.002 s and 0.025 s, respectively. The maximum period factor—the largest possible difference between two consecutive cycles—was set at 1.3. This means that if the period factor—the ratio of the duration of two consecutive cycles—was greater than 1.3, this pair of cycles was not considered in the computation of jitter.

The methods used to determine shimmer were almost identical to those used to determine jitter, the main difference being that jitter considers periods and shimmer takes into account the maximum peak amplitude of the signal.

Once the previous parameters had been established, jitter and shimmer were obtained by the formulas shown below [35].

Jitter(absolute) is the cycle-to-cycle variation of the fundamental period (ie, the average absolute difference between consecutive periods):

where $T_i$ is the duration of the *i*th cycle and *N* is the total number of cycles. If $T_i$ or $T_{i-1}$ is outside the floor and ceiling periods, or if or $\frac{T_i}{T_{i-1}}$ is greater than the maximum period factor, the term $|T_i - T_{i-1}|$ is not counted in the sum, and *N* is lowered by 1 (if *N* ends up being less than 2, the result of the computation becomes "undefined").

Jitter(relative) is the average absolute difference between consecutive periods divided by the average period. It is expressed as a percentage:

$$Jitter(relative) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100 \ (2)$$

Jitter(rap) is defined as the relative average perturbation—the average absolute difference between a period and the average of this and its two neighbors, divided by the average period:

Jitter(ppq5) is the five-point period perturbation quotient, computed as the average absolute difference between a period and the average of this and its four closest neighbors, divided by the average period:

$$Jitter(ppq5) = \frac{\frac{1}{N-1}\sum_{i=2}^{N-2}\left|T_i - \frac{1}{5}\sum_{n=i-2}^{i+2} T_n\right|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100 \ (4)$$

Shimmer(dB) is expressed as the variability of the peak-to-peak amplitude, defined as the difference between the maximum positive and the maximum negative amplitude of each period in decibels (ie, the average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20:

$$Shimmer(dB) = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20 \ \times \log\left(\frac{A_{i+1}}{A_i}\right)\right|, \ (5)$$

Where $A_i$ is the extracted peak-to-peak amplitude data and *N* is the number of extracted fundamental periods.

Shimmer(relative) is defined as the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude, expressed as a percentage:

Shimmer(apq3) is the three-point amplitude perturbation quotient. This is the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbors, divided by the average amplitude:

$$Shimmer(apq3) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|A_i - \frac{1}{3}\sum_{n=i-1}^{i+1} A_n\right|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \times 100 \ (7)$$

Shimmer(apq5) is defined as the five-point amplitude perturbation quotient, or the average absolute difference between the amplitude of a period and the average of the amplitudes of this and its four closest neighbors, divided by the average amplitude:

$$Shimmer(apq5) = \frac{\frac{1}{N-1}\sum_{i=2}^{N-2}\left|A_i - \frac{1}{5}\sum_{n=i-2}^{i+2} A_n\right|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \times 100 \ (8)$$

Shimmer(apq11) is expressed as the 11-point amplitude perturbation quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of this and its ten closest neighbors, divided by the average amplitude:

$$Shimmer(apq11) = \frac{\frac{1}{N-1}\sum_{i=5}^{N-5}\left|A_i - \frac{1}{11}\sum_{n=i-5}^{i+5} A_n\right|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \times 100 \ (9)$$

The HNR provides an indication of the overall periodicity of the voice signal by quantifying the ratio between the periodic (harmonics) and aperiodic (noise) components. The HNR was computed using Praat [35], based on the second maximum of normalized autocorrelation function detection, which is used in the following equation:

$$HNR = 10 \times log_{10}\frac{r(t = \tau)}{1 - r(t = \tau)}, (10)$$

where *r*(*t*) is the normalized autocorrelation function, *r*(*t* = τ) is the second local maximum of the normalized autocorrelation and τ is the period of the signal.

The time step, defined as the measurement interval, was set at 0.01 s, the pitch floor at 60 Hz, the silence threshold at 0.1 (time steps that did not contain amplitudes above this threshold, relative to the global maximum amplitude, were considered silent), and the number of periods per window at 4.5, as suggested by Boersma and Weenink [35].

For the purpose of this study, the mean and standard deviation of the HNR were used.

To obtain the pitch, the autocorrelation method implemented in Praat [35] was used. The pitch floor for males and females was set at 60 Hz and 100 Hz, respectively, and the pitch ceiling for males and females was set at 300 Hz and 500 Hz, respectively. The time step was set, according to Praat [35], at

0.0075 s and 0.0125 s for females and males, respectively. Pitch above pitch ceiling and below pitch floor were not estimated. The mean and standard deviation of the pitch, as well as the minimum and maximum pitch, were features obtained from the pitch metric.

Textbox 1 shows the procedure, inspired by Praat [35], that was used to obtain the features explained above. The full code is freely available online [36].

**Textbox 1.** Algorithm for obtaining the features (jitter, shimmer, harmonics-to-noise ratio [HNR], and pitch) for acoustic analysis.

1. Each individual phonation of each vowel was cut out and anonymously labeled to define the boundaries of the speech segments.
2. The values for the optional paramaters for analysis were set:
   - Optional parameters to obtain jitter and shimmer parameters
     - pitch floor: females 100 Hz and males 60 Hz
     - pitch ceiling: females 500 Hz and males 300 Hz
     - period floor: 0.002 s
     - period ceiling: 0.025 s
     - maximum period factor: 1.3
   - Optional parameters to obtain HNR
     - time step: 0.01 s
     - pitch floor: 60 Hz
     - silence threshold: 0.1
     - number of periods per windows: 4.5
   - Optional parameters to obtain pitch
     - pitch floor: females 100 Hz and males 60 Hz
     - pitch ceiling: females 500 Hz and males 300 Hz
     - time step: females 0.0075 s and males 0.0125 s
3. Compute jitter and shimmer features—jitter(absolute), jitter(relative), jitter(rap), jitter(ppq5), shimmer(dB), shimmer(relative), shimmer(apq3), shimmer(apq5), shimmer(apq11)—using the configuration parameters established and then obtain the mean of each of these parameters for each vowel.
4. Compute HNR using the configuration parameters established and then obtain the mean (HNR[mean]) and standard deviation (HNR[SD]) values.
5. Compute pitch using the configuration parameters established and then obtain the mean (pitch[mean]), standard deviation (pitch[SD]), minimum (pitch[min]), and maximum (pitch[max]) values.
6. Obtain a data set with the 15 features computed.

## PCA

The PCA technique [37], a ranking feature extraction approach, was implemented in R [38] using the Stats package [38]. PCA was used to decompose the original data set into principal components (PCs) to obtain another data set whose data were linearly independent and therefore uncorrelated. It was performed by means of singular value decomposition (SVD) [39].

Prior to applying PCA, given that the mean age of control subjects was approximately 12 years younger than patients with ALS, we removed the age effects by using the data from the control subjects and applying the correction to all the participants as in the study by Norel et al [21]. We fitted the features extracted for healthy people and their age linearly. Then, the "normal aging" of each single feature of each participant was obtained by multiplying the age of the participants by the slope parameter obtained from the linear fit. Finally, the computed "normal aging" was removed from the features. Afterward, a standardized data set was obtained by subtracting the mean and centering the age-adjusted features at 0.

Then, by applying SVD to the standardized data set, a decomposition was obtained: $X = USV^{\top}$ , where $X$ is the matrix of the standardized data set, $U$ is a unitary matrix and $S$ is the diagonal matrix of singular values $s_i$. PCs are given by $US$, and $V$ contains the directions in this space that capture the maximal variance of the features of the matrix $X$. The number of PCs obtained was the same as the original number of features, and the total variance of all of the PCs was equal to the total variance among all of the features. Therefore, all of the information contained in the original data was preserved.

From the PCA, a biplot chart was obtained for a visual appraisal of the data [40]. The biplot chart allowed us to visualize the data set structure, identify the data variability and clustering participants, and display the variances and correlations of the analyzed features. Then, the first eight PCs that explained almost 100% of the variance were selected to fit the classification models.

## Supervised Models

The participants in this study belonged to three different groups: the control group (n=18), patients with ALS with bulbar involvement (n=14), and patients with ALS without bulbar involvement (n=31). Each participant was properly labeled as control (C) if the subject was a control participant, ALS with bulbar (B) if the subject was a participant with ALS diagnosed with bulbar involvement, or ALS without bulbar (NB) if the subject was a participant with ALS without bulbar involvement. In addition, the ALS (A) label was added to every participant with ALS, with or without bulbar involvement.

Supervised models were built to obtain predictions by comparing the four labeled groups between them. Textbox 2 summarizes the procedure used to create proper classification models.

**Textbox 2.** Algorithm used to create the classification models.

1. Building the data set: each participant was classified as C (control), B (amyotrophic lateral sclerosis [ALS] with bulbar involvement), or NB (ALS without bulbar involvement) according to the features extracted from the utterance of the five Spanish vowels and the categorical attributes of the bulbar involvement.
2. "Undefined" values were found in few participants when computing the shimmer(apq11) for a specific vowel. They were handled by computing the mean of this parameter for the other vowels uttered by the same participant.
3. The age effects were removed from the data set.
4. The values of the features obtained from the acoustic analysis were zero centered and scaled by using the following equation: $(x_i - \bar{X}) / \sigma$, where $x_i$ is the feature vector, $\bar{X}$ is the mean, and $\sigma$ is the standard deviation. Scaling was performed to handle highly variable magnitudes of the features prior to computing primary component analysis (PCA).
5. The PCA was computed and a new data set was created with the first eight primary components (PCs).
6. A random seed was set to generate the same sequence of random numbers. They were used to divide the data set into chunks and randomly permute the data set. The random seed made the experiments reproducible and the classifier models comparable.
7. A 10-fold cross-validation technique was implemented and repeated for 10 trials. The data set was divided into ten contiguous chunks of approximately the same size. Then, 10 training-testing experiments were performed as follows: each chunk was held to test the classifier, and we performed training on the remaining chunks, applying upsampling with replacement by making the group distributions equal; the experiments were repeated for 10 trials, each trial starting with a random permutation of the data set.
8. Two different classification thresholds were established; 50% and 95% (more restrictive). The classification threshold is a value that dichotomizes the result of a quantitative test to a simple binary decision by treating the values above or equal to the threshold as positive and those below as negative.

Several supervised classification models were implemented in R [38] to measure the classification performance. The classification models were fitted with the first eight PCs that explained almost 100% of the data variability. Finally, 10-fold cross-validation was implemented in R using the caret package [41] to draw suitable conclusions. The upsampling technique with replacement was applied to the training data by making the group distributions equal to deal with the unbalanced data set, which could bias the classification models [42].

The first classifier employed was SVM, which is a powerful, kernel-based classification paradigm. SVM was implemented using the e1071 [43]. We used a C-support vector classification [44] and a linear kernel that was optimized through the tune function, assigning values of 0.0001, 0.0005, 0.001, 0.01, 0.1, and 1 to the C parameter, which controls the trade-off between a low training error and a low testing error. A C parameter value of 1 gave the best performance, and thus this was the SVM model chosen.

Next, a classical NN trained with the back propagation technique with an adaptive learning rate was implemented using the RSNNS package [45]. After running several trials to decide the NN architecture, a single hidden layer with three neurons was implemented because it showed the best performance. The activation function (transfer function) used was the hyperbolic tangent sigmoid function.

LDA was implemented using the MASS package [46]. It estimated the mean and variance in the training set and computed the covariance matrix to capture the covariance between the groups to make predictions by estimating the probability that the test set belonged to each of the groups.

LR was implemented by using the Gaussian generalized linear model applying the Stats package [38] for binomial distributions. A logit link function was used to model the probability of “success.” The purpose of the logit link was to take a linear combination of the covariate values and convert those values into a probability scale.

Standard NaB based on applying Bayes’ theorem was implemented using the e1071 package [43].

Finally, the RF classifier was implemented using the randomForest package [47] with a forest of 500 decision tree predictors. The optimal mtry—a parameter that indicated the number of PCs that were randomly distributed at each decision tree—was optimized for each classification problem by using the train function included in the caret package [41]. Each decision tree performed the classification independently and

RF computed each tree predictor classification as one “vote.” The majority of the votes computed by all of the tree predictors decided the overall RF prediction.

The code of these implementations is freely available online [48].

### Performance Metrics

There are several metrics to evaluate classification algorithms [49]. The analysis of such metrics and their significance must be interpreted correctly to evaluate these algorithms.

There are four possible results in the classification task. If the sample is positive and it is classified as positive, it is counted as a true positive (TP), and when it is classified as negative, it is considered a false negative (FN). If the sample is negative and it is classified as negative or positive, it is considered a true negative (TN) or false positive (FP), respectively. Based on that, three performance metrics, presented below, were used to evaluate the performance of the classification models.

- Accuracy: ratio between the correctly classified samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

- Sensitivity: proportion of correctly classified positive samples compared with the total number of positive samples.

$$Sensitivity = \frac{TP}{TP + FN} \quad (12)$$

- Specificity: proportion of correctly classified negative samples compared with the total number of negative samples.

$$Specificity = \frac{TN}{TN + FP} \quad (13)$$

Finally, paired Bonferroni-corrected Student *t* tests [50] were implemented to evaluate the statistical significance of the metrics results. To reject the null hypothesis, which entails considering that there is no difference in the performance of the classifiers, a significance level of α=.05 was established for all tests. The *P* values obtained by performing the tests with values below α=.05 rejected the null hypothesis.

## Results

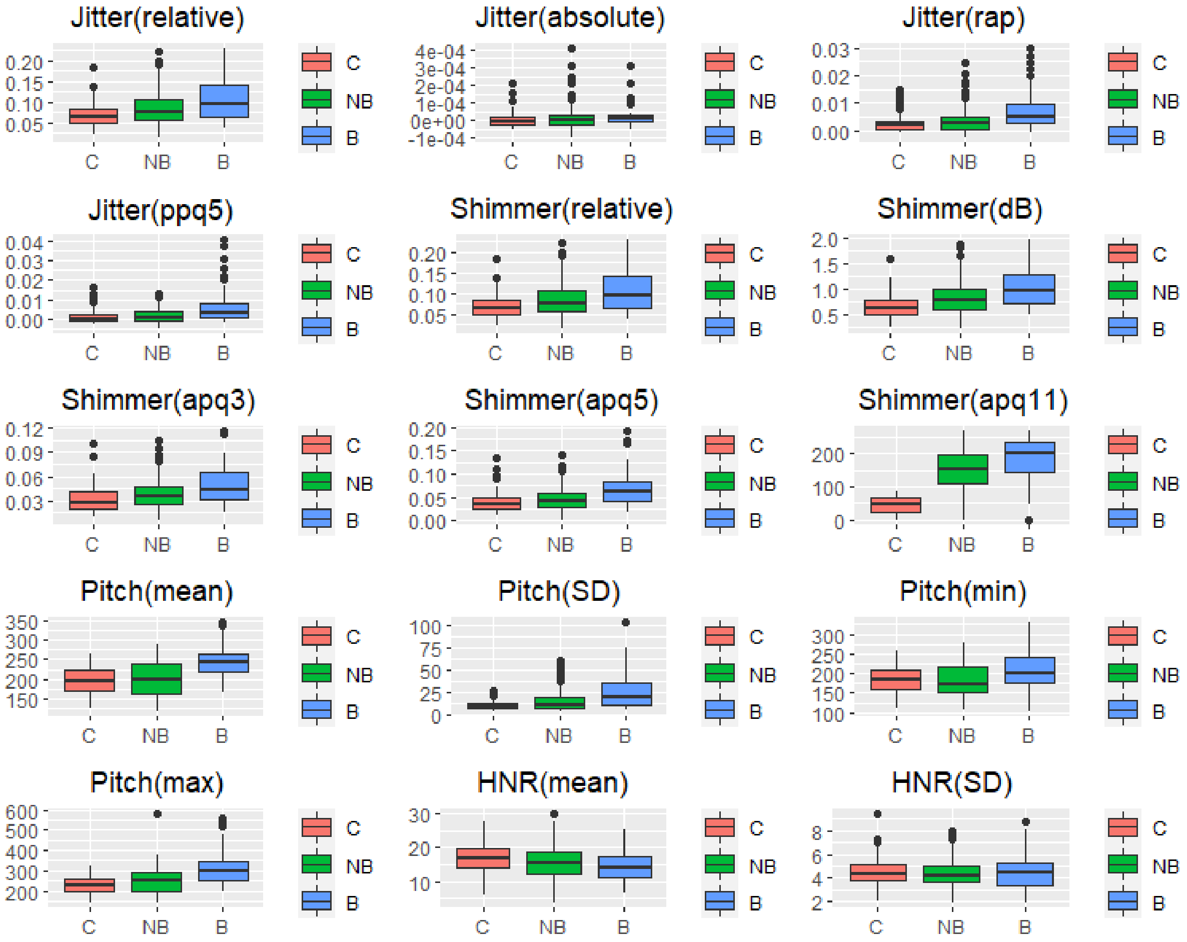
First, the distributions of the features obtained were examined. Then, the PCA was performed and the supervised models studied were evaluated.

### Data Exploration

A total of 15 features were obtained in this study. These features were jitter(absolute), jitter(relative), jitter(rap), jitter(ppq5), shimmer(relative), shimmer(dB), shimmer(apq3), shimmer(apq5), shimmer(apq11), pitch(mean), pitch(SD), pitch(min), pitch(max), HNR(mean), and HNR(SD).

Figure 1 shows the box plot of the features obtained from the control (C) group, patients with ALS with bulbar involvement (B), and patients with ALS without bulbar involvement (NB). The means in the B group were higher than those in the C and NB groups. The means in the NB group were located in the middle of the means of the C and B groups. On the contrary, the B group obtained the lowest values for the mean HNR(mean) and HNR(SD). Differences in the standard deviation between the three groups were also observed. In general, features obtained from the B group presented the highest standard deviations.

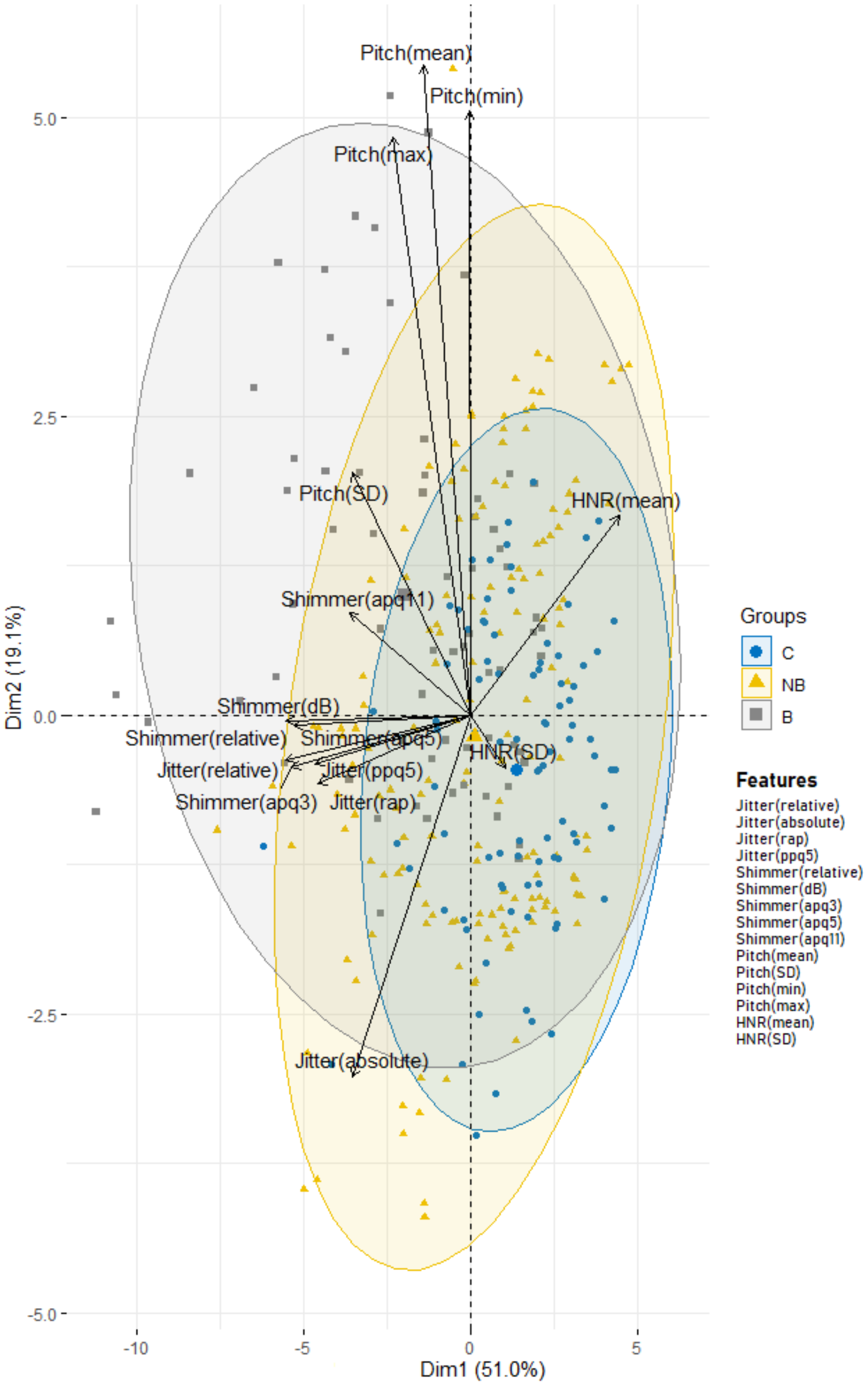
**Figure 1.** Box plots of features by group. B: patients with amyotrophic lateral sclerosis (ALS) with bulbar involvement; C: control group; HNR: harmonics-to-noise ratio; NB: patients with ALS without bulbar involvement.



## PCA

PCA was performed using the data set that contained the 15 features extracted from all of the participants. Figure 2 shows the associated PCA biplot chart. The two axes represent the first (Dim1) and second (Dim2) PCs. The biplot uses the diagonalization method to give a graphical display of its dimensional approximation [51,52]. The interpretation of the biplot involves observing the lengths and directions of the vectors of the features, the data variability, and the clusterization of the participants.

**Figure 2.** Principal component analysis biplot chart representing the variance of the first (Dim1) and second (Dim2) principal components in the control group (C), patients with amyotrophic lateral sclerosis (ALS) without bulbar involvement (NB), and patients with ALS with bulbar involvement (B). HNR: harmonics-to-noise ratio.



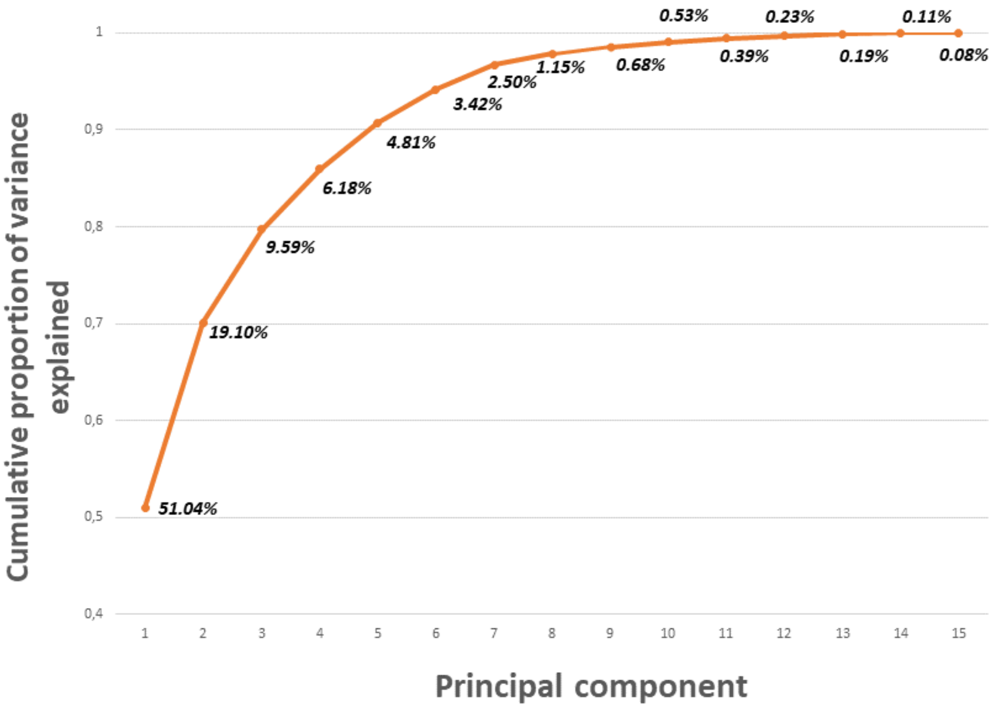
It can be observed that a considerable proportion of variance (70.1%) of the shimmer, jitter, pitch, and HNR was explained. The relative angle between any two vector features represents their pairwise correlation. The closer the vectors are to each other (<90°), the higher their correlation. When vectors are perpendicular (angles of 90° or 270°), the variables have a small or null correlation. Angles approaching 0° or 180° (collinear vectors) indicate a correlation of 1 or –1, respectively. Thus, in this case, shimmer and jitter show a strong positive correlation. Another important observation reflected in Figure 2 is the spatial proximity of the groups in relation both to each other and to the set of features. The projection of the B group onto the vector for shimmer and jitter falls to the left of the vector features. This means that subjects labelled as the B group had higher average values for those features than the average values of the other groups. Conversely, the projection of the C group onto those variables falls on the opposite side. In addition, the C and B groups are more distant from each other when projected onto shimmer and jitter. This indicates that shimmer and jitter features are the most important features for the classification of participants in the B and C groups.

The projection of subjects in the NB group requires special attention. Although the projection of these subjects has a spatial proximity with respect to the C group, their variability is higher, overflowing the gray circle corresponding to the B group.

This indicates that some features, especially shimmer and jitter, of some subjects in the NB group have similar projections to the features of the B group.

To fit the models, as explained in detail in the next section, the first eight PCs were selected in order to reduce the dimensionality but preserve almost 100% of the variability as shown in Figure 3.

**Figure 3.** Cumulative percentage of the explained variance using principal component analysis.



## Supervised Model Evaluation

The first eight PCs were selected. Then, each classification model was applied to these PCs. Consequently, better results were obtained than when applying the classification models alone. The results of the classification methods alone are not shown because of their limited contribution to the analysis.

Tables 1 and 2 show the classification performance (accuracy, sensitivity, and specificity metrics) of the supervised models tested for the four cases with the classification threshold set at 50% and 95%, respectively.

**Table 1.** Classification performance of the supervised models with the classification threshold set at 50%.

| Model and metrics | Classification performance (%) | | | |
|---|---|---|---|---|
| | C[a] vs B[b] | C vs NB[c] | B vs NB | C vs ALS[d] |
| **Random forest** | | | | |
| Accuracy | 93.6 | 91.1 | 75.5 | 90.3 |
| Sensitivity | 91.1 | 92.1 | 55.7 | 92.1 |
| Specificity | 95.5 | 89.6 | 88.4 | 85.7 |
| **Naïve Bayes** | | | | |
| Accuracy | 91.0 | 87.9 | 75.4 | 90.3 |
| Sensitivity | 89.2 | 86.7 | 62.7 | 92.1 |
| Specificity | 93.2 | 90.0 | 81.2 | 85.7 |
| **Logistic regression** | | | | |
| Accuracy | 93.8 | 91.4 | 70.1 | 91.1 |
| Sensitivity | 92.5 | 89.1 | 62.2 | 89.6 |
| Specificity | 94.8 | 95.6 | 73.5 | 93.3 |
| **Linear discriminant analysis** | | | | |
| Accuracy | 94.3 | 91.6 | 71.2 | 91.6 |
| Sensitivity | 95.6 | 87.4 | 61.8 | 88.3 |
| Specificity | 90.0 | 98.8 | 75.4 | 87.8 |
| **Neuronal network** | | | | |
| Accuracy | 94.8 | 92.5 | 70.4 | 92.2 |
| Sensitivity | 91.7 | 90.3 | 60.0 | 90.8 |
| Specificity | 97.2 | 96.4 | 75.2 | 95.6 |
| **Support vector machine** | | | | |
| Accuracy | 95.8 | 91.5 | 69.9 | 91.6 |
| Sensitivity | 91.4 | 88.4 | 59.4 | 88.9 |
| Specificity | 99.3 | 97.0 | 74.6 | 98.2 |

[a] C: control group.

[b] B: patients with amyotrophic lateral sclerosis (ALS) with bulbar involvement.

[c] NB: patients with ALS without bulbar involvement.

[d] ALS: all patients with ALS.

**Table 2.** Classification performance of the supervised models with the classification threshold set at 95%.

| Model and metrics | Classification performance (%) | | | |
|---|---|---|---|---|
| | C[a] vs B[b] | C vs NB[c] | B vs NB | C vs ALS[d] |
| **Random forest** | | | | |
| Accuracy | 58.3 | 56.1 | 68.8 | 75.1 |
| Sensitivity | 4.8 | 30.4 | 0.0 | 65.6 |
| Specificity | 100.0 | 100.0 | 100.0 | 98.8 |
| **Naïve Bayes** | | | | |
| Accuracy | 82.3 | 68.8 | 72.8 | 75.1 |
| Sensitivity | 64.7 | 54.6 | 15.8 | 65.6 |
| Specificity | 96.1 | 93.3 | 98.6 | 98.8 |
| **Logistic regression** | | | | |
| Accuracy | 92.8 | 77.7 | 74.1 | 76.0 |
| Sensitivity | 84.8 | 65.1 | 16.7 | 66.4 |
| Specificity | 99.0 | 99.6 | 100.0 | 100.0 |
| **Linear discriminant analysis** | | | | |
| Accuracy | 88.1 | 70.6 | 71.7 | 71.1 |
| Sensitivity | 72.7 | 53.5 | 0.9 | 59.5 |
| Specificity | 100.0 | 100.0 | 100.0 | 100.0 |
| **Neuronal network** | | | | |
| Accuracy | 92.6 | 84.8 | 73.1 | 86.8 |
| Sensitivity | 83.2 | 76.1 | 20.5 | 81.6 |
| Specificity | 100.0 | 100.0 | 96.8 | 99.8 |
| **Support vector machine** | | | | |
| Accuracy | 86.3 | 71.1 | 70.7 | 71.1 |
| Sensitivity | 68.8 | 54.3 | 6.1 | 59.4 |
| Specificity | 100.0 | 100.0 | 100.0 | 100.0 |

[a] C: control group.

[b] B: patients with amyotrophic lateral sclerosis (ALS) with bulbar involvement.

[c] NB: patients with ALS without bulbar involvement.

[d] ALS: all patients with ALS.

In the case of the C group versus the B group, with the classification threshold set at 50%, the results indicated that all classifiers had a good classification performance. SVM obtained the best accuracy (95.8%). The tests of significance, which are reported in Multimedia Appendix 2, revealed statistically significant differences between SVM and the other models, with the exception of LDA, which obtained an accuracy (94.3%) that closely approximated that of the SVM model. NN also showed really good results (accuracy 94.8%).

Similar behavior was obtained in the C group versus the NB group and the C group versus all patients with ALS. In these cases, NN was the best model (92.5% for C vs NB and 92.2% for C versus ALS). Meanwhile, generally poor performance was obtained in the B group versus the NB group compared with the other cases. Although RF showed the best accuracy (75.5%), the performance of specificity and especially sensitivity dropped dramatically in comparison with the previous cases.

In general, the model performance dropped with a 95% threshold. In the C group versus the B group, the accuracy of the classification models (Table 2) was worse than when the classification threshold was set at 50%. LR shows the best accuracy (92.8%). LDA, SVM, and NaB obtained accuracies of 88.1%, 86.3%, and 82.3%, respectively. RF did not seem to be a good model for this threshold, with an accuracy of 58.3%.

Lower results were obtained in the C group versus the NB group and the C group versus the group with ALS. NN showed the best performance, with accuracies of 84.8% and 86.8%, respectively.

With the 95% threshold, the performance of sensitivity dropped in all cases, especially for the B group versus the NB group, where LR obtained the best performance with an accuracy of 74.1% but a sensitivity of 16.7%.

# Discussion

## Principal Findings

This study was guided by 2 objectives: (1) to design a methodology for diagnosing bulbar involvement efficiently through the acoustic parameters of uttered vowels in Spanish, and (2) to demonstrate the superior performance of automated diagnosis of bulbar involvement compared with human diagnosis. This was based on the accurate acoustic analysis of the five Spanish vowel segments, which were elicited from all participants. A total of 15 acoustic features were extracted: jitter(absolute), jitter(relative), jitter(rap), jitter(ppq5), shimmer(relative), shimmer(dB), shimmer(apq3), shimmer(apq5), shimmer(apq11), pitch(mean), pitch(SD), pitch(min), pitch(max), HNR(mean), and HNR(SD). Then, the PCs of these features were obtained to fit the most common supervised classification models in clinical diagnosis: SVM, NN, LDA, LR, NaB, and RF. Finally, the performance of the models was compared.

The study demonstrated the feasibility of automatic detection of bulbar involvement in patients with ALS through acoustic features obtained from vowel utterance. It also confirms that speech impairment is one of the most important aspects for diagnosing bulbar involvement, as was suggested by Pattee et al [33]. Furthermore, bulbar involvement can be detected using automatic tools before it becomes perceptible to human hearing.

Voice features extracted from the B group compared with those features extracted from the C group showed the best performance of the classification model for determining bulbar involvement in patients with ALS.

Accuracy for the C group versus the B group revealed values of 95.8% for SVM with the classification threshold established at 50%. However, on increasing the threshold to 95%, the accuracy values for SVM dropped (86.3%) and LR showed the best performance (accuracy 92.8%). NN also showed a good accuracy at 92.6%. This implies that NN and LR are more robust for finding accuracy.

For that case, the results obtained reinforce the idea that it is possible to diagnose bulbar involvement in patients with ALS using supervised models and objective measures. The SVM and LR models provided the best performance for the 50% and 95% thresholds, respectively.

Great uncertainty was found in the analysis regarding bulbar involvement in the NB group. The accuracy values of the C group versus the NB group and the C group versus the group with ALS with the classification threshold at 50% were 92.5% and 92.2%, respectively, for NN. That reveals that the features extracted from the NB group differed significantly from those of the C group. Lower performance should be expected because participants labeled as the C group and NB group should have similar voice performance. This may indicate that some of the participants in the NB group probably had bulbar involvement but were not correctly diagnosed because the perturbance in their voices could not be appreciated by the human ear. Alternatively, it could be simply that a classification threshold of 50% was too optimistic. With a 95% classification threshold, lower results were obtained in the C group versus the NB group and in the C group versus patients with ALS. NN showed the best performance with accuracies of 84.8% and 86.8%, respectively, for the two cases.

The performance between the B group and C group showed better results than between the NB group and C group. Despite this, the unexpectedly high performance of the models for the C group versus the NB group still suggests that some participants in the NB group could have had bulbar involvement. Changing the classification threshold to 95% worsened the results, especially for sensitivity, although this still remained significant.

The case of the B group versus the NB group revealed that the classification models did not distinguish B group and NB group participants as well as they did with the other groups. The accuracy with the 50% threshold showed the highest performance for RF (75.5%), but the models showed difficulties in identifying positive cases. That may be due to the small difference in the variation of the data among participants in the B and NB groups. The same occurred for the 95% threshold: LR obtained the highest accuracy (74.1%) but a sensitivity of only 16.7%. These values remain far from those in the case of the C group versus the B group. These results also reinforce the idea that participants in the NB group were misdiagnosed.

The good model performance obtained in comparing the C and NB groups supports these findings and underscores the importance of using objective measures for assessing bulbar involvement. This corroborated the results obtained in the data exploration and PCA, which were presented in the Results section.

The projection of the NB group in the PCA biplot chart requires special attention. Although the projection of these subjects has a spatial proximity with regard to the C group, their variability is higher, overflowing the circle corresponding to the B group. This indicates that some features, especially shimmer and jitter, of some patients in the NB group have similar projections to those in the B group. This may reveal that these patients in the NB group could have bulbar involvement but were not yet correctly diagnosed because the perturbance in their voices could still not be appreciated by human hearing.

Figure 1 also indicates that the means of the features of the patients in the NB group were between the means of the features of the C and B groups, thus corroborating these assumptions.

## Limitations

This study has some limitations. First, using machine learning on small sample sizes makes it difficult to fully evaluate the significance of the findings. The sample size of this study was heavily influenced by the fact that ALS is a rare disease. At the time of the study, 14 of the patients with ALS presented bulbar symptoms. The relatively small size of this group was because ALS is a very heterogeneous disease and not all patients with ALS present the same symptomatology. Additionally, the control subjects were approximately 12 years younger than the patients with ALS. Vocal quality changes with age, and comparing younger control subjects' vocalic sounds with those of older participants with ALS might introduce additional variations.

Although upsampling techniques were used in this study to correct the bias and age adjustments have been applied to correct the vocal quality changes due to the age difference, it would be necessary in future studies to increase the number of participants, especially of patients with ALS with bulbar involvement and control participants of older ages, to draw definitive conclusions.

Second, the variability inherent in establishing the boundaries of the speech segments on spectrograms manually makes replicability challenging. Speakers will differ in their production, and even the same speaker in the same context will not produce two completely identical utterances. In this study, the recorded speech was processed manually in the uniform approach detailed in the Methods section. Automatic instruments have been developed, but unfortunately these methods are not yet accurate enough and require manual correction.

## Comparison with Prior Work

The PCA biplot charts indicated that shimmer and jitter were the most important features for group separation in the 2-PC model for ALS classification; however, they also revealed pitch and HNR parameters as good variables for this purpose. These results are consistent with those of Vashkevich et al [25], who demonstrated significant differences in jitter and shimmer in patients with ALS. They are also consistent with Mekyska et al [19] and Teixeira et al [18], who mentioned pitch, jitter, shimmer, and HNR values as the most popular features describing pathological voices. Finally, Silbergleit et al [8] suggested that the shimmer, jitter, and HNR parameters are sensitive indicators of early laryngeal deterioration in ALS.

Concerning the classification models, Norel et al [21] recently implemented SVM classifiers to recognize the presence of speech impairment in patients with ALS. They identified acoustic speech features in naturalistic contexts, achieving 79% accuracy (sensitivity 78%, specificity 76%) for classification of males and 83% accuracy (sensitivity 86%, specificity 78%) for classification of females. The data used did not originate from a clinical trial or contrived study nor was it collected under laboratory conditions. Wang et al [9] implemented SVM and NN using acoustic features and adding articulatory motion information (from tongue and lips). When only acoustic data were used to fit the SVM, the overall accuracy was slightly higher than the level of chance (50%). Adding articulatory motion information further increased the accuracy to 80.9%. The results using NN were more promising, with accuracies of 91.7% being obtained using only acoustic features and increasing to 96.5% with the addition of both lip and tongue data. Adding motion measures increased the classifier accuracy significantly at the expense of including more invasive measurements to obtain the data. We investigated the means of optimizing accuracy in detecting ALS bulbar involvement by only analyzing the voices of patients. An et al [23] implemented CNNs to classify the intelligible speech produced by patients with ALS and healthy individuals. The experimental results indicated a sensitivity of 76.9% and a specificity of 92.3%. Vashkevich et al [25] performed LDA with an accuracy of 90.7% and Suhas et al [16] used DNNs based on MFCCs with an accuracy of 92.2% for automatic detection of patients with ALS.

Starting with the most widely used features suggested in the literature, the classification models used in this paper to detect bulbar involvement automatically (C group versus B group) performed better than the ones used by other authors, specifically the ones obtained using NN (Wang et al [9]) and DNNs based on MCCFs (Suhas et al [16]). We obtained the best-ever performance metrics. This suggests that decomposing the original data set of features into PCs to obtain another data set whose data (ie, PCs) were linearly independent and therefore uncorrelated improves the performance of the models.

## Conclusions

This paper suggests that machine learning may be an appropriate tool to help in the diagnosis of ALS by multidisciplinary clinical teams. In particular, it could help in the diagnosis of bulbar involvement. This work demonstrates that an accurate analysis of the features extracted from an acoustic analysis of the vowels elicited from patients with ALS may be used for early detection of bulbar involvement. This could be done automatically using supervised classification models. Better performance was achieved by applying PCA previously to the obtained features. It is important to note that when classifying participants with ALS with bulbar involvement and control subjects, the SVM with a 50% classification threshold exceeded the performance obtained by other authors, specifically Wang et al [9] and Suhas et al [16].

Furthermore, bulbar involvement can be detected using automatic tools before it becomes perceptible to human hearing. The results point to the importance of obtaining objective measures to allow an early and more accurate diagnosis, given that humans may often misdiagnose this deficiency. This directly addresses a recent statement released by the Northeast ALS Consortium’s bulbar subcommittee regarding the need for objective-based approaches [53].

## Future Work

Future work is directed toward the identification of incorrectly undiagnosed bulbar-involvement in patients with ALS. A time-frequency representation will be used to detect possible deviations in the voice performance of patients in the time-frequency domain. The voice distributions of patients with ALS diagnosed with bulbar involvement and patients with ALS without that diagnosis will be compared in order to detect pattern differences between these two groups. That could provide indications to distinguish undiagnosed participants with ALS who could be misdiagnosed. Also, an improvement in the voice database by increasing the sample size is envisaged.

## Acknowledgments

This work was supported by the Ministerio de Economía y Competitividad under contract TIN2017-84553-C2-2-R. Einar Meister’s research has been supported by the European Regional Development Fund through the Centre of Excellence in Estonian Studies.

The Neurology Department of the Bellvitge University Hospital in Barcelona allowed the recording of the voices of the participants in its facilities. The clinical records were illustrated by Carlos Augusto Salazar Talavera. Dr Marta Fulla and Maria Carmen Majos Bellmunt advised about the process of eliciting the sounds.

## Conflicts of Interest

None declared.

## Multimedia Appendix 1

Summary of the clinical records of participants with amyotrophic lateral sclerosis.
[PDF File (Adobe PDF File), 37 KB-Multimedia Appendix 1]

## Multimedia Appendix 2

Paired t test with Bonferroni correction.
[PDF File (Adobe PDF File), 55 KB-Multimedia Appendix 2]

## References

1. Carmona C, Gómez P, Ferrer MA, Plamondon R, Londral A. Study of several parameters for the detection of amyotrophic lateral sclerosis from articulatory movement. loquens 2017 Dec 18;4(1):038 [FREE Full text] [doi: 10.3989/loquens.2017.038]
2. Tomik B, Guiloff R. Dysarthria in amyotrophic lateral sclerosis: A review. Amyotroph Lateral Scler 2010;11(1-2):4-15. [doi: 10.3109/17482960802379004] [Medline: 20184513]
3. Makkonen T, Ruottinen H, Puhto R, Helminen M, Palmio J. Speech deterioration in amyotrophic lateral sclerosis (ALS) after manifestation of bulbar symptoms. Int J Lang Commun Disord 2018 Mar;53(2):385-392. [doi: 10.1111/1460-6984.12357] [Medline: 29159848]
4. Tomik B, Krupinski J, Glodzik-Sobanska L, Bala-Slodowska M, Wszolek W, Kusiak M, et al. Acoustic analysis of dysarthria profile in ALS patients. J Neurol Sci 1999 Oct 31;169(1-2):35-42. [doi: 10.1016/s0022-510x(99)00213-0] [Medline: 10540005]
5. Shellikeri S, Green JR, Kulkarni M, Rong P, Martino R, Zinman L, et al. Speech Movement Measures as Markers of Bulbar Disease in Amyotrophic Lateral Sclerosis. J Speech Lang Hear Res 2016 Oct 01;59(5):887-899 [FREE Full text] [doi: 10.1044/2016_JSLHR-S-15-0238] [Medline: 27679842]
6. Lee J, Dickey E, Simmons Z. Vowel-Specific Intelligibility and Acoustic Patterns in Individuals With Dysarthria Secondary to Amyotrophic Lateral Sclerosis. J Speech Lang Hear Res 2019 Jan 30;62(1):34-59. [doi: 10.1044/2018_JSLHR-S-17-0357] [Medline: 30950759]
7. Carpenter RJ, McDonald TJ, Howard FM. The otolaryngologic presentation of amyotrophic lateral sclerosis. Otolaryngology 1978;86(3 Pt 1):ORL479-ORL484. [doi: 10.1177/019459987808600319] [Medline: 112540]
8. Silbergleit AK, Johnson AF, Jacobson BH. Acoustic analysis of voice in individuals with amyotrophic lateral sclerosis and perceptually normal vocal quality. J Voice 1997 Jun;11(2):222-231. [doi: 10.1016/s0892-1997(97)80081-1] [Medline: 9181546]
9. Wang J, Kothalkar PV, Kim M, Bandini A, Cao B, Yunusova Y, et al. Automatic prediction of intelligible speaking rate for individuals with ALS from speech acoustic and articulatory samples. Int J Speech Lang Pathol 2018 Nov;20(6):669-679 [FREE Full text] [doi: 10.1080/17549507.2018.1508499] [Medline: 30409057]
10. Chiaramonte R, Di Luciano C, Chiaramonte I, Serra A, Bonfiglio M. Multi-disciplinary clinical protocol for the diagnosis of bulbar amyotrophic lateral sclerosis. Acta Otorrinolaringol Esp 2019;70(1):25-31 [FREE Full text] [doi: 10.1016/j.otorri.2017.12.002] [Medline: 29699694]
11. Tomik J, Tomik B, Wiatr M, Składzień J, Stręk P, Szczudlik A. The Evaluation of Abnormal Voice Qualities in Patients with Amyotrophic Lateral Sclerosis. Neurodegener Dis 2015;15(4):225-232. [doi: 10.1159/000381956] [Medline: 25967115]
12. Connaghan K, Green J, Paganoni S. Use of beiwe smartphone app to identify and track speech decline in amyotrophic lateral sclerosis (ALS). Graz, Austria; 2019 Presented at: Interspeech 2019; September 15-19; Graz, Austria. [doi: 10.21437/interspeech.2019-3126]
13. Horwitz-Martin R, Quatieri T, Lammert A. Relation of automatically extracted formant trajectories with intelligibility loss and speaking rate decline in amyotrophic lateral sclerosis. 2016 Presented at: Interspeech 2016; September 16; San Francisco. [doi: 10.21437/interspeech.2016-403]
14. Rong P. Parameterization of articulatory pattern in speakers with ALS. 2014 Presented at: Interspeech 2014; September 18; Singapore.
15. Spangler T, Vinodchandran N, Samal A, Green J. Fractal features for automatic detection of dysarthria. 2017 Presented at: IEEE EMBS International Conference on Biomedical & Health Informatics (BHI); 2017; Orlando, FL. [doi: 10.1109/bhi.2017.7897299]

16. Suhas, Patel D, Rao N. Comparison of speech tasks and recording devices for voice based automatic classification of healthy subjects and patients with amyotrophic lateral sclerosis. 2019 Presented at: Interspeech 2019; September 15-19; Graz, Austria. [doi: 10.21437/interspeech.2019-1285]
17. Frid A, Kantor A, Svechin D, Manevitz L. Diagnosis of Parkinson?s disease from continuous speech using deep convolutional networks without manual selection of features. 2016 Presented at: IEEE International Conference on the Science of Electrical Engineering (ICSEE); 2016; Eilat, Israel. [doi: 10.1109/icsee.2016.7806118]
18. Teixeira JP, Fernandes PO, Alves N. Vocal Acoustic Analysis – Classification of Dysphonic Voices with Artificial Neural Networks. Procedia Comput Sci 2017;121:19-26. [doi: 10.1016/j.procs.2017.11.004]
19. Mekyska J, Janousova E, Gomez-Vilda P, Smekal Z, Rektorova I, Eliasova I, et al. Robust and complex approach of pathological speech signal analysis. Neurocomputing 2015 Nov;167:94-111. [doi: 10.1016/j.neucom.2015.02.085]
20. Garcia-Gancedo L, Kelly ML, Lavrov A, Parr J, Hart R, Marsden R, et al. Objectively Monitoring Amyotrophic Lateral Sclerosis Patient Symptoms During Clinical Trials With Sensors: Observational Study. JMIR Mhealth Uhealth 2019 Dec 20;7(12):e13433 [FREE Full text] [doi: 10.2196/13433] [Medline: 31859676]
21. Norel R, Pietrowicz M, Agurto C, Rishoni S, Cecchi G. Detection of amyotrophic lateral sclerosis (ALS) via acoustic analysis. In: Interspeech 2018. 2018 Presented at: Interspeech 2018; September 2-6; Hyderabad, India. [doi: 10.21437/interspeech.2018-2389]
22. Rong P, Yunusova Y, Wang J, Zinman L, Pattee GL, Berry JD, et al. Predicting Speech Intelligibility Decline in Amyotrophic Lateral Sclerosis Based on the Deterioration of Individual Speech Subsystems. PLoS One 2016;11(5):e0154971 [FREE Full text] [doi: 10.1371/journal.pone.0154971] [Medline: 27148967]
23. An K, Kim M, Teplansky K. Automatic early detection of amyotrophic lateral sclerosis from intelligible speech using convolutional neural networks. 2018 Presented at: Interspeech 2018; September 2-6; Hyderabad, India. [doi: 10.21437/interspeech.2018-2496]
24. Gutz S, Wang J, Yunusova Y, Green J. Early identification of speech changes due to amyotrophic lateral sclerosis using machine classification. 2019 Presented at: Interspeech 2019; September 15-19; Graz, Austria. [doi: 10.21437/interspeech.2019-2967]
25. Vashkevich M, Petrovsky A, Rushkevich Y. Bulbar ALS detection based on analysis of voice perturbation and vibrato. 2019 Presented at: IEEE 2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA); September 18-20; Poznan, Poland. [doi: 10.23919/spa.2019.8936657]
26. Jolliffe I. Principal Component Analysis. In: Lovric M, editor. International Encyclopedia of Statistical Science. Berlin Heidelberg: Springer; 2011:1094-1096.
27. Rodriguez-Lujan I, Bailador G, Sanchez-Avila C, Herrero A, Vidal-de-Miguel G. Analysis of pattern recognition and dimensionality reduction techniques for odor biometrics. Knowl-Based Syst 2013 Nov;52:279-289. [doi: 10.1016/j.knosys.2013.08.002]
28. Zhao W, Chellappa R, Krishnaswamy A. Discriminant analysis of principal components for face recognition. 1998 Presented at: Third IEEE International Conference on Automatic Face and Gesture Recognition; April 14-16; Nara, Japan. [doi: 10.1109/afgr.1998.670971]
29. Hosmer D, Lemeshow S. Applied Logistic Regression, Second Edition. Hoboken, NJ: John Wiley & Sons, Inc; 2000.
30. Dingen D, van't Veer M, Houthuizen P, Mestrom EHJ, Korsten EH, Bouwman AR, et al. RegressionExplorer: Interactive Exploration of Logistic Regression Models with Subgroup Analysis. IEEE T Vis Comput Gr 2019 Jan;25(1):246-255. [doi: 10.1109/tvcg.2018.2865043]
31. Flaxman A, Vahdatpour A, Green S, James S, Murray C, Population Health Metrics Research Consortium (PHMRC). Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. Popul Health Metr 2011 Aug 04;9:29 [FREE Full text] [doi: 10.1186/1478-7954-9-29] [Medline: 21816105]
32. Bermejo P, Gámez JA, Puerta JM. Speeding up incremental wrapper feature subset selection with Naive Bayes classifier. Knowl-Based Syst 2014 Jan;55:140-147. [doi: 10.1016/j.knosys.2013.10.016]
33. Pattee GL, Plowman EK, Focht Garand KL, Costello J, Brooks BR, Berry JD, Contributing Members of the NEALS Bulbar Subcommittee. Provisional best practices guidelines for the evaluation of bulbar dysfunction in amyotrophic lateral sclerosis. Muscle Nerve 2019 May;59(5):531-536. [doi: 10.1002/mus.26408] [Medline: 30620104]
34. Audacity Manual Contents. Audacity. 2019. URL: https://manual.audacityteam.org/ [accessed 2021-02-01]
35. Moltu C, Stefansen J, Svisdahl M, Veseth M. Negotiating the coresearcher mandate - service users' experiences of doing collaborative research on mental health. Disabil Rehabil 2012;34(19):1608-1616. [doi: 10.3109/09638288.2012.656792] [Medline: 22489612]
36. Voice features extraction. Alberto Tena. URL: https://github.com/atenad/greco [accessed 2021-02-01]
37. Phaladiganon P, Kim SB, Chen VC, Jiang W. Principal component analysis-based control charts for multivariate nonnormal distributions. Expert Syst Appl 2013 Jun;40(8):3044-3054. [doi: 10.1016/j.eswa.2012.12.020]
38. The R Project for Statistical Computing. URL: https://www.R-project.org/ [accessed 2021-02-01]
39. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Berlin: Springer; 2009.

40. Gabriel KR, Odoroff CL. Biplots in biomedical research. Stat Med 1990 May;9(5):469-485. [doi: 10.1002/sim.4780090502] [Medline: 2349401]
41. Kuhn M. Building Predictive Models in Using the caret Package. J Stat Softw 2008;28(5):1-26. [doi: 10.18637/jss.v028.i05]
42. Kuhn M, Johnson K. Applied Predictive Modeling. New York: Springer; 2013.
43. Meyer D. Misc Functions of the Department of Statistics, Probability Theory Group. R package version 1. 2019. URL: https://CRAN.R-project.org/package=e1071 [accessed 2021-02-01]
44. Boser B, Guyon I, Vapnik V. A Training Algorithm for Optimal Margin Classifiers. : Association for Computing Machinery; 1992 Presented at: COLT'92; July; Pittsburgh, Pennsylvania p. 144-152. [doi: 10.1145/130385.130401]
45. Bergmeir C, Benítez JM. Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. J Stat Softw 2012;46(7):1-26. [doi: 10.18637/jss.v046.i07]
46. Venables W, Ripley B. Modern Applied Statistics with S, Fourth Edition. USA: Springer; 2002.
47. Liaw A, Wiener M. Classification and regression by randomforest. R News 2002;2(3):18-22.
48. Supervised classification models for automated detection of bulbar involvement in als patients. Alberto Tena. URL: https://github.com/atenad/greco [accessed 2021-02-01]
49. Tharwat A. Classification assessment methods. ACI 2020 Aug 03;ahead-of-print(ahead-of-print) [FREE Full text] [doi: 10.1016/j.aci.2018.08.003]
50. Hothorn T, Leisch F, Zeileis A, Hornik K. The Design and Analysis of Benchmark Experiments. J Comput Graph Stat 2005;14(3):675-699. [doi: 10.1198/106186005x59630]
51. Gabriel KR. The biplot graphic display of matrices with application to principal component analysis. Biometrika 1971;58(3):453-467. [doi: 10.1093/biomet/58.3.453]
52. Underhill LG. Two Graphical Display Methods for Ecological Data Matrices. In: McLachlan A, Erasmus T, editors. Sandy Beaches as Ecosystems. Dordrecht: Springer; 1983:433-439.
53. Plowman EK, Tabor LC, Wymer J, Pattee G. The evaluation of bulbar dysfunction in amyotrophic lateral sclerosis: survey of clinical practice patterns in the United States. Amyotroph Lateral Scler Frontotemporal Degener 2017 Aug;18(5-6):351-357 [FREE Full text] [doi: 10.1080/21678421.2017.1313868] [Medline: 28425762]

## Abbreviations

**ALS:** amyotrophic lateral sclerosis
**ALSFRS-R:** ALS Functional Rating Scale-Revised
**CNN:** convolutional neuronal network
**DNN:** deep neuronal network
**FN:** false negative
**FP:** false positive
**HNR:** harmonics-to-noise ratio
**LDA:** linear discriminant analysis
**LR:** logistic regression
**MFCC:** Mel Frequency Cepstral Coefficient
**NaB:** naïve Bayes
**NN:** neuronal network
**PC:** principal component
**PCA:** principal component analysis
**RF:** random forest
**SVD:** singular value decomposition
**SVM:** support vector machine
**TN:** true negative
**TP:** true positive

*Edited by G Eysenbach; submitted 11.09.20; peer-reviewed by E Toki, E Beneteau; comments to author 12.10.20; revised version received 26.10.20; accepted 17.01.21; published 10.03.21*

*Please cite as:*
*Tena A, Claria F, Solsona F, Meister E, Povedano M*
*Detection of Bulbar Involvement in Patients With Amyotrophic Lateral Sclerosis by Machine Learning Voice Analysis: Diagnostic Decision Support Development Study*
*JMIR Med Inform 2021;9(3):e21331*
*URL: https://medinform.jmir.org/2021/3/e21331*
*doi: 10.2196/21331*
*PMID: 33688838*

## 2.2 Paper 2: Automated Detection of COVID-19 Cough

- *Alberto Tena, Francesc Clarià, Francesc Solsona. Automated detection of COVID-19 cough. Biomedical Signal Processing and Control, Volume 71(A):103175. 2022. doi:10.1016/j.bspc.2021.103175.*

### Abstract

In this paper, prior to performing the time-frequency representation analysis, the YAMNet [63] deep neuronal network was used for the automatic identification of cough sounds in raw audio files.

Then, a TFR analysis of a Choi-Williams distribution (CWD) was carried out in the cough-samples identified to obtain discriminatory features for an automated diagnosis of COVID-19. 39 features were extracted and the sets which showed better performance at discriminating COVID-19 cough were selected.

The main objectives (and contributions) of this research were:

1. To design a free, quick and efficient methodology for the automatic detection of COVID-19 in raw audio files based on the time- –frequency analysis of the cough.

2. To obtain the time–frequency discriminatory features leading to automated identification of COVID-19.

3. To find an optimal supervised machine-learning algorithm to diagnose COVID-19 from the cough features found.

Random Forest performed better than the other models analysed in this study. An accuracy close to 90% was obtained.

This study demonstrated the feasibility of the automatic diagnose of COVID-19 from coughs, and its applicability to detecting new outbreaks.

Biomedical Signal Processing and Control 71 (2022) 103175

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

# Automated detection of COVID-19 cough

Alberto Tena [a], Francesc Clarià [b], Francesc Solsona [b,*]

[a] CIMNE, Building C1, North Campus, UPC. Gran Capità, 08034 Barcelona, Spain
[b] Dept. of Computer Science & INSPIRES, University of Lleida. Jaume II 69, E-25001 Lleida, Spain

ARTICLE INFO

*Keywords:*
Keyword: COVID-19
Automated cough detection
Diagnosis
Signal processing
Time–frequency

ABSTRACT

Easy detection of COVID-19 is a challenge. Quick biological tests do not give enough accuracy. Success in the fight against new outbreaks depends not only on the efficiency of the tests used, but also on the cost, time elapsed and the number of tests that can be done massively. Our proposal provides a solution to this challenge. The main objective is to design a freely available, quick and efficient methodology for the automatic detection of COVID-19 in raw audio files.

Our proposal is based on automated extraction of time–frequency cough features and selection of the more significant ones to be used to diagnose COVID-19 using a supervised machine-learning algorithm.

Random Forest has performed better than the other models analysed in this study. An accuracy close to 90% was obtained.

This study demonstrates the feasibility of the automatic diagnose of COVID-19 from coughs, and its applicability to detecting new outbreaks.

## 1. Introduction

COVID19 (COronaVIrus Disease of 2019), caused by the Severe Acute Respiratory Syndrome (SARS-CoV2) virus, was announced as a global pandemic on February 11, 2020 by the World Health Organisation (WHO). By mid-February, 2021, one year after the beginning of the COVID-19 pandemic, over 108 million confirmed cases of COVID-19 had been reported worldwide, with almost 2,400,000 deaths [1].

During this time, it has been demonstrated that COVID-19 outbreaks are very hard to contain with current testing approaches unless region-wide confinement measures are sustained. This is partly because of the limitations of current viral and serological tests and the lack of complementary pre-screening methods [2].

According to the WHO-China Joint Mission report (COVID-19) [3], typical signs and symptoms of COVID-19 are fever (87.9%), dry cough (67.7%), fatigue (38.1%), sputum production (33.4%), shortness of breath (18.6%), sore throat (13.9%), headache (13.6%), myalgia or arthralgia (14.8%), chills (11.4%), nausea or vomiting (5.0%), nasal congestion (4.8%), diarrhoea (3.7%), hemoptysis (0.9%), and conjunctival congestion (0.8%).

Several researchers have proposed methods for identifying cough sounds from audio recordings [4,5]. Automatic cough classification is an active research area in which several researchers have proposed methods for identifying a wide range of respiratory diseases and types of coughs (namely dry and wet coughs) through cough analysis and machine-learning algorithms [6,7].

Various studies have begun to work on the design of machine-learning tools to detect COVID-19 [8–16] as complementary pre-screening method. These are based on the analysis of the sound of voices, and the sounds we make when we breath or cough and which change when our respiratory system is affected. These changes range from coarse, clearly audible changes, to minute changes (called *micro* signatures) that are inaudible to the untrained listener, but nevertheless present [9]. These works have been performed in own datasets and no idenfication of the main features has been performed. We are also interested in the automatic identification of COVID-19 cough from any raw audio recording. Overall, finding a general method and the main cough features from audio records for diagnosing COVID-19 is a challenge.

The difficulty is to find good machine-learning features. Some works in the literature, as we have mentioned before, advocate some features, but in the particular case of COVID-19, it remains to be seen which properties, brands, signs (that is, features) are those that uniquely identify COVID-19. So, the big challenge is to identify the best features that discriminate the COVID-19 cough. In addition, we want to find the group of features with better performance for each type of experiment,

[*] Corresponding author.
*E-mail addresses:* atena@cimne.upc.edu (A. Tena), francisco.claria@udl.cat (F. Clarià), francesc.solsona@udl.cat (F. Solsona).

https://doi.org/10.1016/j.bspc.2021.103175
Received 16 July 2021; Received in revised form 26 August 2021; Accepted 7 September 2021
Available online 13 September 2021

as for example, comparing COVID-19 and pertussis coughs.

The goal of this paper is to develop a pre-screening method that could lead to automated identification of COVID-19 through the analysis of cough time–frequency representations (TFR) with similar performance presented in [8–16]. TFRs permit the evolution of the periodicity and frequency components over time to be observed, allowing the analysis of non-stationary signals. Moreover, this representation, which maintains the time dependence of signal features, gives the possibility of introducing more related features than traditional analysis. This way, we go a step further by finding the set of time–frequency features that could allow COVID-19 coughs to be distinguished from other cough patterns and validate it as a more generic proposal by applying our method to various datasets from different sources.

In the present work, prior to performing the TFR analysis, the YAMNet [17] deep neuronal network was used for the automatic identification of cough sounds in raw audio files. Then, a TFR analysis of a Choi-Williams distribution (CWD) was carried out in the cough-samples identified to obtain discriminatory features for an automated diagnosis of COVID-19. 39 features were extracted and the sets which showed better performance at discriminating COVID-19 cough were selected. For that purpose, the main objectives (and contributions) of this research are:

- To design a free, quick and efficient methodology for the automatic detection of COVID-19 in raw audio files based on the time––frequency analysis of the cough.
- To obtain the time–frequency discriminatory features leading to automated identification of COVID-19.
- To find an optimal supervised machine-learning algorithm to diagnose COVID-19 from the cough features found.

## 2. Methods

The methods presented in this section were implemented and a synthetic dataset based on a random sample of COVID-19 and non-COVID-19 coughs is freely available online [18]. It was built in R using the synthpop package [19]. Also, the code of the machine-learning models used is also provided.

This section presents the corpus, the automatic cough identification process and the basis theory used to obtain the time–frequency features. The classification models were fitted by a set of the most important features, obtained by two different techniques, namely feature selection and feature extraction. The most popular supervised models in cough classification are then presented. Finally, the model's performance metrics are introduced.

### 2.1. Data Corpus

This section describes the data collection framework used in this work. It consisted of the COVID-19 dataset the University of Lleida collected for this study which was approved by the Research Ethics Committee for Biomedical Research Projects (CEIm) at the University Hospital Arnau de Vilanova of Lleida, and three additional existing publicly available COVID-19 datasets, namely University of Cambridge [20], Coswara [21] and Virufy [22] datasets. Additionally, the Pertussis dataset [6], which includes recordings of patients with pertussis cough, was also used.

Our analysis used four sets. The first set (C) consisted of subjects tested COVID-19 positive; the second set (N) were subjects tested COVID-19 negative; the third set (NC) were non-COVID subjects, but who had non-specified-coughs as a symptom; and the fourth set (PT) were non-COVID subjects but who presented pertussis cough.

Table 1 shows the set of participants selected and Table 2 shows the demographic data for each group.

**Table 1**
Corpus. UdL: University of Lleida; UC: University of Cambridge.

| | UdL | UC | Coswara | Virufy | Pertussis | Total |
|---|---|---|---|---|---|---|
| C | 49 | 142 | 107 | 48 | 0 | 346 |
| N | 3 | 137 | 133 | 73 | 0 | 346 |
| NC | 0 | 53 | 48 | 0 | 0 | 101 |
| PT | 0 | 0 | 0 | 0 | 20 | 20 |
| Total | 52 | 332 | 288 | 121 | 20 | 813 |

**Table 2**
Demographic dataset properties. NA: Data not-available.

| | C | N | NC | PT |
|---|---|---|---|---|
| Males (%) | 68.0 | 50.5 | 55.2 | NA |
| Females (%) | 32.0 | 49.5 | 44.8 | NA |
| Age | 48.9 ± 11.9 | 40.8 ± 9.1 | 44.6 ± 7.3 | NA |

### 2.2. Automatic Cough Identification

Fig. 1 shows an overview of the automatic cough identification process developed which was inspired by [23].

The YAMNet deep neuronal network [17] was used for the automatic identification of the cough samples registered in the raw audio files. YAMNet classifies audio segments into sound classes described by the AudioSet ontology [24] employing MobileNet [25]. The MobileNet structure is built on depthwise separable convolutions which factorises a standard convolution into a depthwise and a pointwise convolution (1 x 1 convolution kernel) [26]. Depthwise convolution applies the filter to each input channel, and 1 x 1 pointwise convolution is used to combine the outputs of the depthwise convolution. The YAMNet body architecture employing MobileNet is defined in Fig. 2.

All layers are depthwise separable convolutions except for the first layer, which is a standard convolution, and the last few layers which are pooling, fully connected layers, and a softmax layer for classification. Each convolution layer used ReLU as the activation function, and batchnorm was used for the standardised distribution of batches. The convolution layer structure is shown in Fig. 3.

To obtain the input layer passed to YAMNet, the original audio waveforms of the raw audio files were pre-processed. They were resampled to 16 kHz and buffered into L overlapping segments. Each segment was 0.98 s and the segments were overlapped by 0.8575 s. They were converted to a magnitude spectrogram with 257 frequency bins using a one-sided short-time Fourier transform (STFT) with a 25-ms periodic Hann window with a 10-ms hop and a 512-point Discrete Fourier Transform (DFT). Then, the magnitude spectrum was passed through a 64-band mel-spaced filter bank and the magnitudes of each band were summed. The audio was represented by a 96-by-64-by-1-by-L array, where 96 is the number of spectrums in the mel spectrogram and 64 is the number of mel bands. Finally, the mel spectrograms were converted to a log scale. The 96-by-64-by-1-by-L array of mel spectrograms was the input layer passed through YAMNet. The output from YAMNet (L-by-512 matrix) corresponds to confidence scores for each of the 521 sound classes over time.

The post-processing consisted of selecting the sound regions labeled as "cough" for analysis. Firstly, to detect the sound event region, the 521 confidence signals were passed through a moving mean filter with a window length of 7 and each signal through a moving median filter with a window length of 3. Although other better filters exists, combining mean and median filters offers good performance at reasonable computational costs [27]. The window length of the mean filter was computed as the $Segment\_duration/Hope\_length - 1$ where *Segment_duration* was the duration of the L segments (0.98 s) and *Hope_length* was the hope length between two consecutive segments (0.1225 s). The length of the median filter was established considering optimal computational costs.

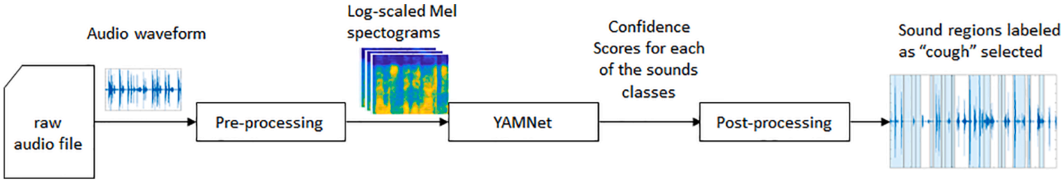

Audio waveform
raw audio file
Log-scaled Mel spectograms
Pre-processing
YAMNet
Confidence Scores for each of the sounds classes
Post-processing
Sound regions labeled as "cough" selected

**Fig. 1.** Overview of the automatic cough identification process.

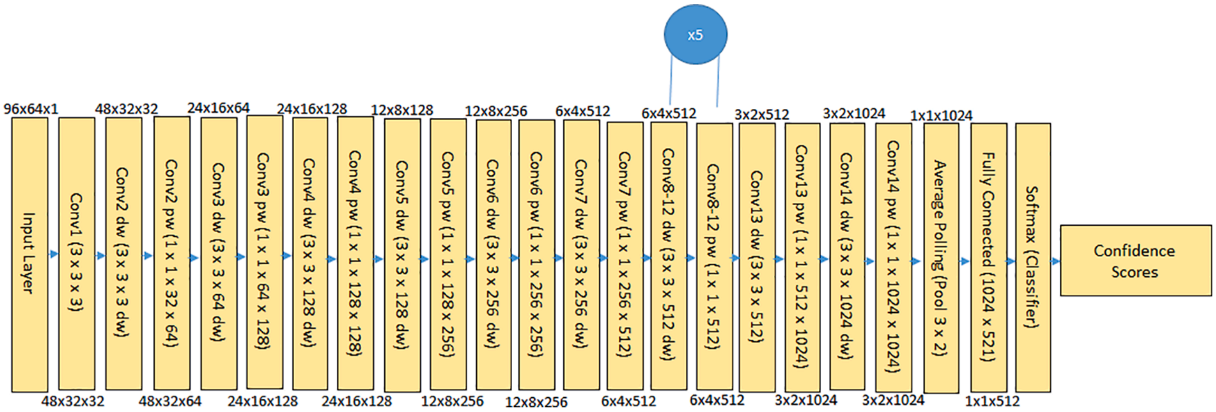

x5
96x64x1
48x32x32
24x16x64
24x16x128
12x8x128
12x8x256
6x4x512
6x4x512
3x2x512
3x2x1024
1x1x1024
Input Layer
Conv1 (3 x 3 x 3)
Conv2 dw (3 x 3 x 3 dw)
Conv2 pw (1 x 1 x 32 x 64)
Conv3 dw (3 x 3 x 64 dw)
Conv3 pw (1 x 1 x 64 x 128)
Conv4 dw (3 x 3 x 128 dw)
Conv4 pw (1 x 1 x 128 x 128)
Conv5 dw (3 x 3 x 128 dw)
Conv5 pw (1 x 1 x 128 x 256)
Conv6 dw (3 x 3 x 256 dw)
Conv6 pw (1 x 1 x 256 x 256)
Conv7 dw (3 x 3 x 256 dw)
Conv7 pw (1 x 1 x 256 x 512)
Conv8-12 dw (3 x 3 x 512 dw)
Conv8-12 pw (1 x 1 x 512)
Conv13 dw (3 x 3 x 512)
Conv13 pw (1 x 1 x 512 x 1024)
Conv14 dw (3 x 3 x 1024 dw)
Conv14 pw (1 x 1 x 1024 x 1024)
Average Polling (Pool 3 x 2)
Fully Connected (1024 x 521)
Softmax (Classifier)
Confidence Scores
48x32x32
48x32x64
24x16x128
24x16x128
12x8x256
12x8x256
6x4x512
6x4x512
3x2x1024
3x2x1024
1x1x512

**Fig. 2.** YAMNet Body Architecture. Conv: Convolution. dw: Depthwise. pw: Pointwise.

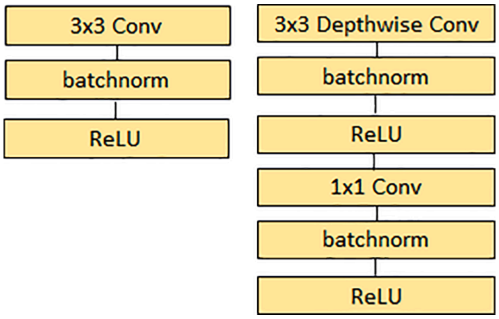


**Fig. 3.** Left: Standard convolutional layer with batchnorm and ReLU. Right: Depthwise Separable convolutions with Depthwise and Pointwise layers followed by batchnorm and ReLU.

Then, the confidence signals were converted into binary masks. After running several trials, a threshold of 0.35 was set because it showed the best performance at detecting "cough" samples. Any sound shorter than 0.5 s was discarded for analysis and regions shorter than 0.25 s were merged.

The identified sound regions that overlapped by 50% or more were consolidated into single regions. The region start time selected was the smallest start time and the region end time selected was the largest end time of all sounds in the group.

Then, the sound regions labelled as "cough" by YAMNET were selected for analysis. The boundaries of these cough samples were selected by using the `detectSpeech` algorithm available in [23], which is based on [28] using a Hann window with 0.03·*Sampling_rate* seconds hop. Finally, the first 600 ms of each cough sample identified were re-sampled at 8,820 Hz and normalised to obtain the Time–frequency representations and features.

Fig. 4 illustrates the process of the automatic identification of cough boundaries in a raw audio file. Fig. 4a shows the sound classification performed by YAMNET. Fig. 4b shows the resulting audio signal after the selection of those audio regions labelled as "cough". Fig. 4c shows the boundaries of the cough samples defined for analysis.

### 2.3. Time–frequency Representation

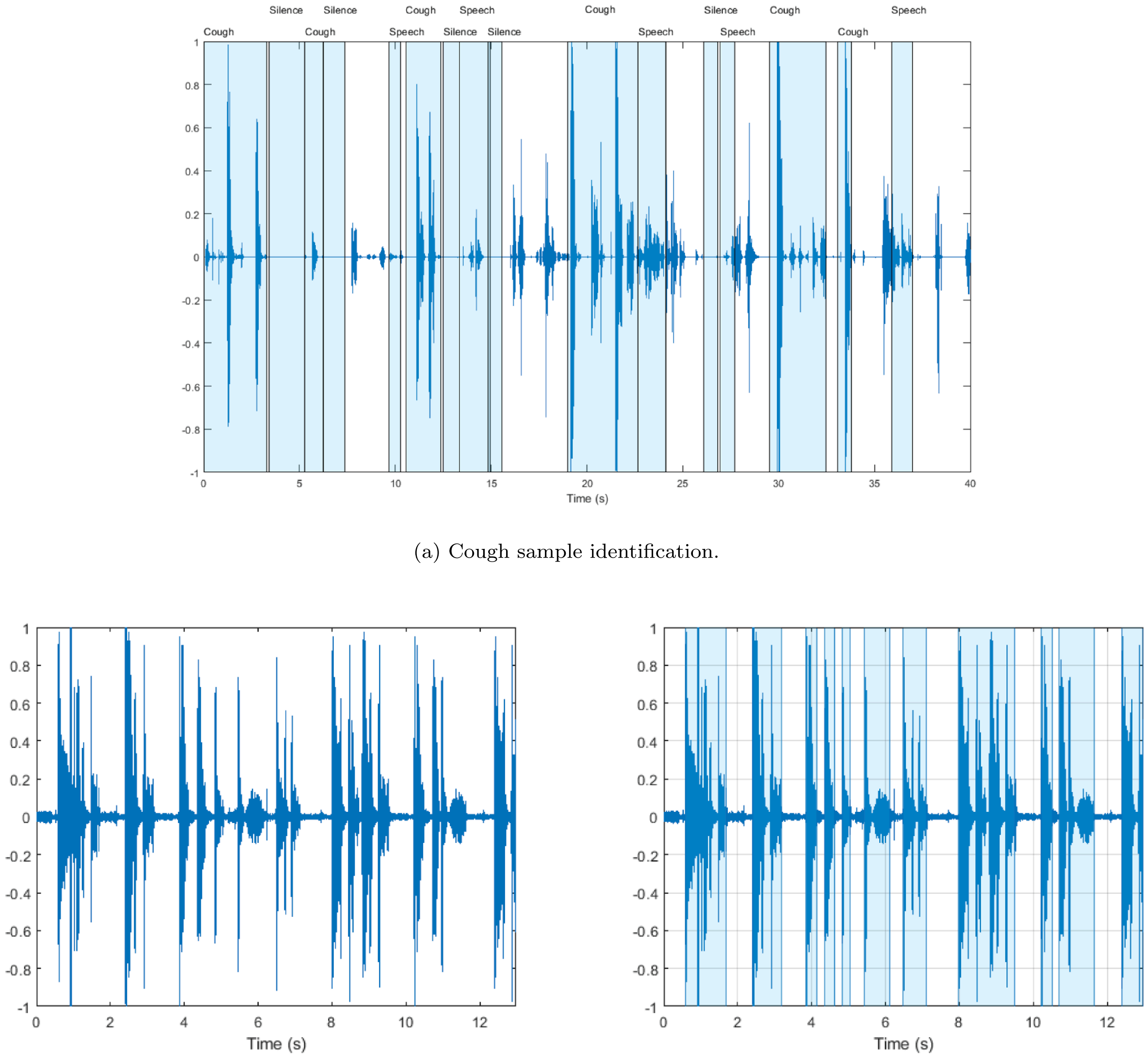
The Wigner distribution (WD) has been used in different fields and applied to the study of time-varying and strongly non-stationary systems. Since the energy is a quadratic representation of the signal, the quadratic structure of the time–frequency representation (TFR) is intuitive and reasonably accepted when the TFR is interpreted as an energy distribution in time and frequency [29]. From all TFRs that represent energy, the WD satisfies many desired mathematical properties. For example, the WD is always real, symmetrical with respect to the time and frequency axes, satisfying the marginal properties and the instantaneous frequency. Furthermore, the group delay may be obtained. Eq. 1 represents the WD of the signal *x(t)*.

$$WD\left(t,f\right) = \int x\left(t+\tau/2\right)x^{*}\left(t-\tau/2\right)e^{-j2\pi f\tau}d\tau, \quad (1)$$

where $t$ and $f$ represent time and frequency respectively, and $x^{*}(t)$ is the conjugate of $x(t)$.

Basically, the WD of a real signal *x(t)* is calculated in a similar way to a convolution. At each particular time, the signal is overlapped by itself and inverted on the time axis, and multiplied by itself. Finally, the Fourier transform of this product is carried out. Note that neither will the WD be necessarily zero when *x(t)* nor would the WD necessarily be zero at frequencies that do not exist in the spectrum. Evidence of this phenomenon has been called interference terms and cross-terms. The interference terms are undesired since they make it difficult to obtain a clear and intuitive spectrum of the signal, as two energy regions perfectly delimited are expected to be obtained.

The possibility of using the WD as a representation of the signal spectral density at each particular time induces the generation of another distribution from the WD to minimise these interference terms while simultaneously maintaining certain properties. To achieve this, we calculated the convolution of the WD of each cough sample was calculated with the Choi–Williams exponential function *h(t,f)* [30] (Eq. 2). By convolving the Wigner distribution with the Choi–Williams exponential, the Choi–Williams distribution (CWD) was obtained (Eq. 3).



(a) Cough sample identification.

(b) Selection of cough samples.

(c) Cough sample boundaries for analysis.

**Fig. 4.** Automatic identification of cough samples in a raw audio file.

$$h\left(t,f\right)=\sqrt{\frac{4\pi}{\sigma_c}}e^{-4\pi^2\frac{(tf)^2}{\sigma_c}}, \tag{2}$$

where $\sigma_c$ is a scaling factor.

$$CWD\left(t,f\right)=\iint h\left(t-t',f-f'\right)WD\left(t',f'\right)dt'\,df' \tag{3}$$

CWD preserves the properties of WD [30,31], such as the marginal properties and instantaneous frequency. Moreover, it is able to reduce the WD interference by estimating an adequate $\sigma_c$ parameter. In this study, the $\sigma_c$ parameter was established at 0.05 to eliminate the interference produced. So, the CWD is a new function of the time–frequency distribution that allows the interference terms to be minimised.

Then, in order to obtain statistical parameters, the density function $CWD(f,t)$ was normalised to have an area equal to 1. So, it can be associated with a joint probability density function $CWD_N(f,t)$ of the time and frequency variables. Their marginal distributions, which do not contain the interference, still represent, although in a normalised manner, the instantaneous power (Eq. 4) and and spectral density energy (Eq. 5) of the original signal.

$$m_t\left(t\right)=\int_{-\infty}^{\infty}CWD_N\left(f,t\right)df=|x(t)|^2 \tag{4}$$

$$m_f\left(f\right)=\int_{-\infty}^{\infty}CWD_N\left(f,t\right)dt=|X(f)|^2 \tag{5}$$

Therefore, the group delay (Eq. 6) and the mean frequency of the spectrum (Eq. 7) can be defined as:

$$t_g=\iint tCWD_N\left(t,f\right)dt\,df \tag{6}$$

$$f_m=\iint fCWD_N\left(t,f\right)dt\,df \tag{7}$$

The joint time–frequency moments of a non-stationary signal comprise a set of time-varying parameters that characterise the signal spectrum as it evolves over time. They are related to the conditional temporal moments and the joint time–frequency moments. The joint time–frequency moment is an integral function of frequency, given time, and marginal distribution. The conditional temporal moment is an integral function of time, given frequency, and marginal distribution. The calculation of the joint time–frequency moment $t^nf^m$ (Eq. 8) is a double integral through time and frequency [32].

$$\langle t^n f^m \rangle = \iint \left(t - t_g\right)^n (f - f_m)^m CWD_N \left(t,f\right) dt\,df \tag{8}$$

where $n$ and $m$ are the frequency and time moment orders.

The moments of the marginal density functions, that define the relationship between $m_t(t)$ and $m_f(f)$, $\langle m_t(t)^n m_f(f)^m \rangle$, are given in Eq. 9.

$$\langle m_t(t)^n m_f(f)^m \rangle = \frac{1}{std\left(m_t(t)^n\right) std\left(m_f(f)^m\right)} \iint \left(m_t(t) - \overline{m_t(t)}\right)^n \left(m_f\left(f\right) - \overline{m_f\left(f\right)}\right)^m dt\,df \tag{9}$$

*CWD* minimises the interference. However, negative values still remain. To solve this issue, the CWD was reformulated as the product of its marginal distributions. Therefore, the joint probability density distribution *pD* (Eq. 10) was obtained. This procedure was only possible because the marginal distributions of the CWD were statistically independent. To corroborate this, the moments of the $CWD_N$ from $n = 1$ and $m = 1$ to $n = 15$ and $m = 15$ were computed, and little covariability was observed. This meant that the marginal distributions could be considered statistically independent.

$$pD\left(f,t\right) = m_f\left(f\right) \cdot m_t\left(t\right) \tag{10}$$

Fig. 5(a) corresponds to the WD of a cough sample. It shows how the interference terms of the WD make it difficult to obtain a clear and intuitive spectrum of the signal. The new function $pD(f,t)$ (Fig. 5(c)) is equal to WD without interference (as CWD, Fig. 5(b)) and negative values.

### 2.4. Time–frequency features

This section explains how a total of 39 features were obtained from the time frequency representation of each cough sample. 28 of them corresponded to the instantaneous spectral energy, instantaneous frequency, instantaneous frequency peak and spectral information. These were obtained by dividing the spectrum (0–4,410 Hz) into 7 frequency bands: 1, 0–80 Hz; 2, 80–250 Hz; 3, 250–550 Hz; 4, 550–900 Hz; 5, 900–1,500 Hz; 6, 1,500–3,000 Hz; 7, 3,000–4,410 Hz. The mean frequency of the total spectrum, the joint, instantaneous and spectral Shannon entropies, the Kurtosis, 3 joint time–frequency moments and 3 joint moments of the marginal signals of instantaneous power and spectral density were also computed.

The instantaneous spectral energy, *E(t)* (Eq. 11), was calculated for each cough sample as the $pD(f,t)$ integral in the frequency domain. Next, the instantaneous frequency, $f_{mi}(t)$, of the spectrum was computed [31] as the average frequency of the spectrum with respect to time (Eq. 12).

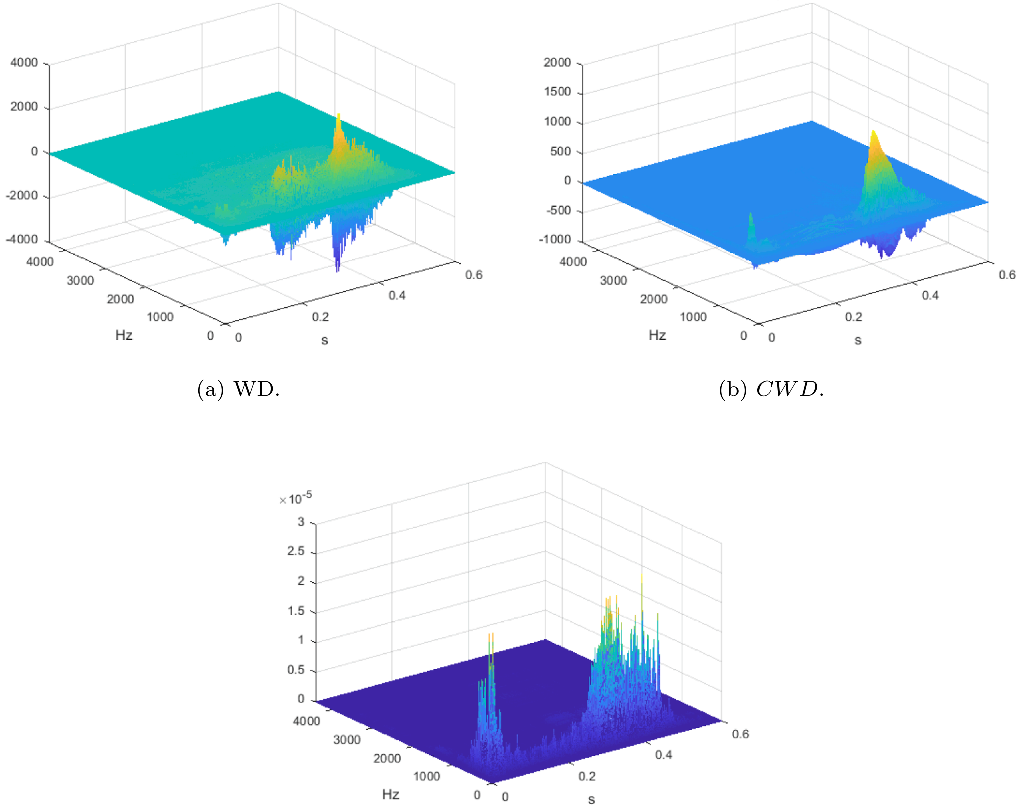
$$E\left(t\right) = \int_{f_1}^{f_2} pD\left(f,t\right) df, \tag{11}$$

where $f_1$ and $f_2$ are the lower and upper frequencies of each band.

$$f_{mi}\left(t\right) = \int_{f_1}^{f_2} \frac{1}{E(t)} f pD\left(f,t\right) df \tag{12}$$

The Instantaneous Frequency Peak, $f\_Cres(t)$ (Eq. 13), is defined as the maximum frequency value at every instant.

$$f\_Cres\left(t\right) = \frac{1}{E(t)} argmax_f \left[\prod_{f_1}^{f_2} f \cdot pD\left(f,t\right)\right] \tag{13}$$



(a) WD.

(b) *CWD*.

(c) pD.

**Fig. 5.** Time–frequency representations of the same COVID-19 subject.

Then, the joint Shannon ($H\_tf$), instantaneous ($H\_t$) and spectral information ($H\_f$) entropies were measured by means of the Shannon entropy method. Shannon entropies were used to quantify the regularity, uncertainty or randomness of these distributions. Entropy can express the information mean that an event provides when it takes place, the uncertainty about the outcome of an event and the dispersion of the probabilities with which the events take place.

Therefore, to obtain the entropy measurements from the $pD(f,t)$ and with the aim of having a range of values able to discriminate levels of spectral amplitude accurately enough, $pD(f,t)$ was quantified with $N = 2^q$ levels and $q = 20$. When the joint probability density function is quantified ($pD_N(f,t)$), the joint Shannon entropy ($H\_tf$), in this case in a range of 0 to 20 bits, can be obtained (Eq. 14).

$$H\_tf = -\int\int log_2\left(pD_N\left(t,f\right)\right)\cdot pD_N\left(t,f\right)dfdt \tag{14}$$

According to Eq. 10, the joint probability density distribution quantified ($pD_N(f,t)$) is defined in Eq. 15.

$$pD_N\left(t,f\right) = m_{fN}\cdot m_{tN}, \tag{15}$$

where $m_{tN}(t)$ is the quantified instantaneous marginal obtained from the $m_t(t)$ and $m_{fN}(f)$ is the quantified frequency marginal obtained from the $m_f(f)$. Therefore, the joint entropy can also be expressed as in Eq. 16.

$$H\_tf = H\_t + H\_f, \tag{16}$$

where $H\_t$ (Eq. 17) and $H\_f$ (Eq. 18) are the instantaneous and spectral entropy respectively.

$$H\_t = -\int log_2\left(m_{tN}\left(t\right)\right)\cdot m_{tN}\left(t\right)dt \tag{17}$$

$$H\_f = -\int log_2\left(m_{fN}\left(f\right)\right)\cdot m_{fN}\left(f\right)df \tag{18}$$

Then, the spectral information, $IE(f)$ (Eq. 19) is obtained.

$$IE(f) = -log_2(m_{fN}(f)) \tag{19}$$

Then, the Kurtosis (K) can be found (Eq. 20).

$$K = \langle m_t(t)^n m_f(f)^m \rangle \tag{20}$$

for $n = 4$ and $m = 0$.

Starting from the computed parameters $E(t), f_m, f_{mi}(t), f\_Cres(t), H\_tf, H\_t, H\_f, IE(f), \langle t^n f^m \rangle$ and $K$, a total of 39 features were obtained. The averages of $E(t), f_{mi}(t), f\_Cres(t)$ and $IE(t)$ were obtained for each of the 7 frequency bands: 1, 0–80 Hz; 2, 80–250 Hz; 3, 250–550 Hz; 4, 550–900 Hz; 5, 900–1,500 Hz; 6, 1,500–3,000 Hz; 7, 3,000–4,410 Hz. The joint time–frequency moments $\langle t^n f^m \rangle$ for $n = 1$ and $m = 1, n = 7$ and $m = 7$ and $n = 15$ and $m = 15$, and the same joint moments of the marginal signals of instantaneous power and spectral density $\langle m_t(t)^n m_f(f)^m \rangle$, were considered for analysis among all the moments computed. Then, the 39 features obtained were coded as follows:

- f_Cres1…f_Cres7: As the average of f_Crest(t) for each 7-bands.
- Enr_Bn1…Enr_Bn7: As the average of E(t) for each 7-bands.
- fm: As the value of the parameter f_m.
- f_Med1…f_Med7: As the average of $f_{mi}(t)$ for each 7-bands.
- IE_Bn1…IE_Bn7: As the average of $IE(f)$ for each 7-bands.
- H_tf: As the value of the parameter H_tf.
- H_f: As the value of the parameter H_f.
- H_t: As the value of the parameter H_t.
- kurt_Mgt: As the value of the parameter K.
- momC11: As the value of the $n = 1$ and $m = 1$ joint time–frequency moment.
- momC77: As the value of the $n = 7$ and $m = 7$ joint time–frequency moment.
- momC15: As the value of the $n = 15$ and $m = 15$ joint time–frequency moment.
- momM11: As the value of the $n = 1$ and $m = 1$ joint instantaneous power and spectral density moment.
- momM77: As the value of the $n = 7$ and $m = 7$ joint instantaneous power and spectral density moment.
- momM15: As the value of the $n = 15$ and $m = 15$ joint instantaneous power and spectral density moment.

### 2.5. Feature selection

The Recursive Feature Elimination (RFE) is a recursive process that ranks features according to some measure of their importance. At each iteration, feature importance is measured and the less relevant one is removed. The recursion is needed because for some measures the relative importance of each feature can change when evaluated over a different subset of features during the stepwise elimination process. RFE was implemented in R by using the caret package to select the set of features ($S_i$) which obtained the best accuracy for each classification model. Performance evaluation of each set of features was done by using stratified 10-fold cross-validation [33].

### 2.6. Feature extraction

Feature extraction is a process of dimensionality reduction by which an initial set of features is reduced while preserving the information in the original dataset. An Autoencoder was implemented in R using the Keras package to perform this task.

An Autoencoder is a specific type of a neural network, one mainly designed to encode the input data into a compressed and meaningful representation, and then decode it back so that the reconstructed input is similar as possible to the original. The Autoencoder maps the input data $x$ to a hidden representation using the function $z = f(Px + b)$ parameterised by $\{P, b\}$. $f$ is the activation function. The hidden representation is then mapped linearly to the output using $\widehat{x} = Wz + b'$. The parameters are optimised to minimise the mean square error of $\|\widehat{x} - x\|_2^2$ over all training points.

Fig. 6 shows the Autoencoder architecture employed. It consists of three modules: the encoder, the decoder and the bottleneck. The encoder is formed by an input layer of 39 nodes and two hidden layers of 30 and 20 nodes respectively. The bottleneck has 15 nodes and the decoder consists of two hidden layers of 20 and 30 nodes respectively and an output layer of 39 nodes. The activation function selected was the *tanh* function. As the purpose of our Autoencoder was to reduce the feature range of our original dataset, we took the compressed data contained in the bottleneck layer. So, the 39 original features were reduced to 15.

### 2.7. Classification models

Five groups of subjects (C, N, NC, PT and NNC) were defined for analysis. The C group contained COVID-19 subjects. N contained subjects tested COVID-19 negative who had no cough. NC was formed of non-COVID-19 subjects with non-specific–cough as a symptom. PT had non-COVID-19 subjects with pertussis cough. Finally, the NNC group merged all non-COVID-19 subjects (N, NC and PT). Then, four classification experiments were performed. These consisted of C vs. N, C vs. NC, C vs. PT and C vs. NNC.

The most popular supervised models in cough classification were used and were implemented in R. These were Random Forest (RF), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Logistic Regression (LR) and Naïve Bayes (NB). The classification models were fitted on the one hand to the selected features obtained by means of RFE and, on the other hand, to the features extracted by means of the Autoencoder. Finally, 10-fold cross-validation [33] was

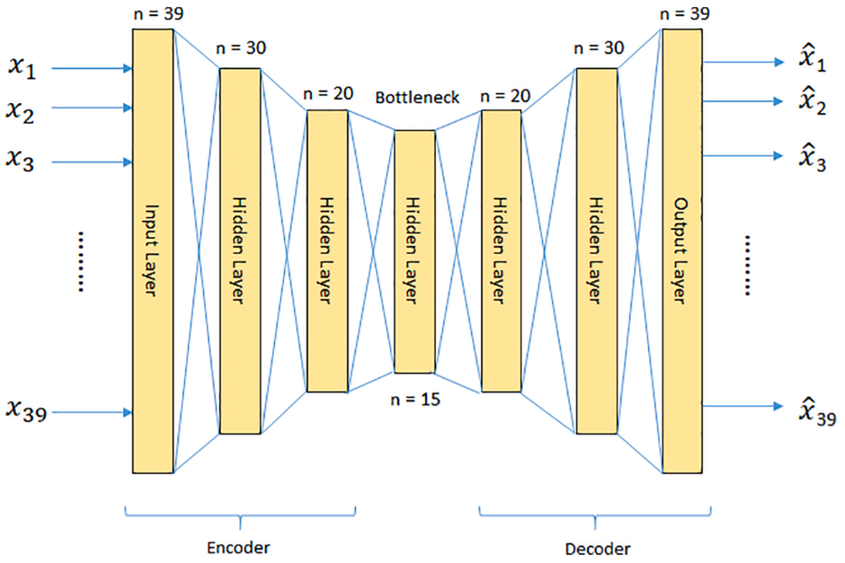


**Fig. 6.** Autoencoder Architecture.

implemented in R using the caret package to draw suitable conclusions. An upsampling technique with replacement was applied to the training data by making the group distributions equal to deal with the unbalanced dataset that could bias the classification models.

The first classifier employed was the RF. It was implemented using the R randomForest package with a forest of 500 decision tree predictors. The optimal number of features that were randomly distributed to each decision tree, was optimized for each classification problem by using the train function included in the R caret package. Each decision tree performed the classification independently and RF computed each tree predictor classification as one “vote”. The majority of the votes computed by all the tree predictors decided the overall RF prediction.

Next, SVM, a powerful kernel-based classification paradigm, was implemented using the R e1071 package. A C-Support Vector Classification [34] was used with a linear kernel that was optimised through the tune function, assigning values 0.0001, 0.0005, 0.001, 0.01, 0.1, 1, 1.25, 1.5, 1.75, 2 and 5 to the C parameter, which controls the trade-off between a low training error and a low testing error. The value of *C* which gave the best performance was chosen.

Then, LDA was implemented using the R MASS package. This estimated the mean and variance from the training set and computed the covariance matrix to capture the co-variance between the groups to make predictions by estimating the probability that the test set belongs to every group.

LR was implemented by using the Gaussian generalised linear model, applying the R Stats package for binomial distributions. A logit link function was used to model the probability of “success”. The purpose of the logit link was to take a linear combination of the covariate values and convert these into a probability scale.

Finally, standard NB based on applying Bayes’ theorem was implemented using the e1071 package [35].

### *2.8. Performance metrics*

There are four possible results in the classification task: If the sample is positive and it is classified as positive, it is counted as a *true positive* (TP) and when classified as negative, it is considered a *false negative* (FN). If the sample is negative and is classified as negative or positive, it is considered a *true negative* (TN) or *false positive* (FP) respectively. Based on that, the Accuracy, Sensitivity (also known as recall), Specificity, Precision and F-score metrics ([36]) were used to evaluate the performance of the classification models using a classification threshold of 50%. The Area Under the Curve (AUC) was also calculated.

- **Accuracy** (Eq. 21). Ratio between the correctly classified samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{21}$$

- **Sensitivity** (Eq. 22). Proportion of correctly classified positive samples compared to the total number of positive samples.

$$Sensitivity = \frac{TP}{TP + FN} \tag{22}$$

- **Specificity** (Eq. 23). Proportion of correctly classified negative samples compared to the total number of negative samples.

$$Specificity = \frac{TN}{TN + FP} \tag{23}$$

- **Precision** (Eq. 24). Proportion of positive samples that were correctly classified compared to the total number of positive predicted samples.

$$Precision = \frac{TP}{FP + TP} \tag{24}$$

- **F-score** (Eq. 25). Harmonic mean of the precision and sensitivity.

$$F - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{25}$$

- **AUC** (Eq. 26). The Receiver operating characteristics (ROC) curve is a two-dimensional graph in which Sensitivity is plotted on the y-axis and 1 −*Specificity* is plotted on the x-axis. The points of the curve are obtained by sweeping the classification threshold from the most positive classification value to the most negative. The AUC score is a scalar value that measures the area under the ROC curve and is always bounded between 0..1.

$$AUC = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}1_{p_i > p_j}, \tag{26}$$

where i runs over all m samples with true label positive, and j runs over all n samples with true label negative; $p_i$ and $p_j$ denote the probability score assigned by the classifier to sample *i* and *j*, respectively.

## 3. Results

Firstly, a visual appraisal of time–frequency representations of coughs from C, N, NC and PT subjects is presented. Then, the distributions of the features obtained for each of the five groups defined for analysis were explored. Finally, the four experiments defined were implemented and the classification models were evaluated.

### 3.1. pD Representation

Fig. 7 shows the comparison of the $pD(f,t)$ of coughs from C, N, NC and PT subjects. Fig. 7(a) corresponds to a C subject who tested positive in a PCR. Figs. 7(b), 7(c) and 7(d) correspond to N, NC and PT subjects respectively.

The visual appraisal of Fig. 7(a) shows how the energy of the $pD(f,t)$ is concentrated in the frequency range from 0 to 1 kHz. In Fig. 7(b), low-energy frequency components can be observed at higher frequencies. In Fig. 7(c), low-energy frequency components can be also observed at higher frequencies but only ranged from 0 to 2 kHz. In Fig. 7(d) energy components of the $pD(f,t)$ can be observed in the frequency range from 0 to 3 kHz although the higher amplitudes are present in frequencies ranging from 0 to 1 kHz. It can be observed that there are no interference terms in any figure.

### 3.2. Data exploration

A total of 39 time–frequency features were obtained in this study: f_Cres1:f_Cres7, Enr_Bn1:Enr_Bn7, f_Med1:f_Med7, IE_Bn1:IE_Bn7, H_tf, H_t, H_f, fm, kurt_Mgt, MomC_11, MomC_77, MomC_1515, MomM_11, MomM_77 and MomC_1515.

There were remarkable differences in the mean and standard deviation between the features, and more specifically, in the following features (Fig. 8): f_Cres1, f_Cres3, Enr_Bn1, Enr_Bn2, Enr_Bn6, f_Med1, f_Med3, f_Med7, IE_Bn2, IE_Bn3, IE_Bn5 and IE_Bn7.

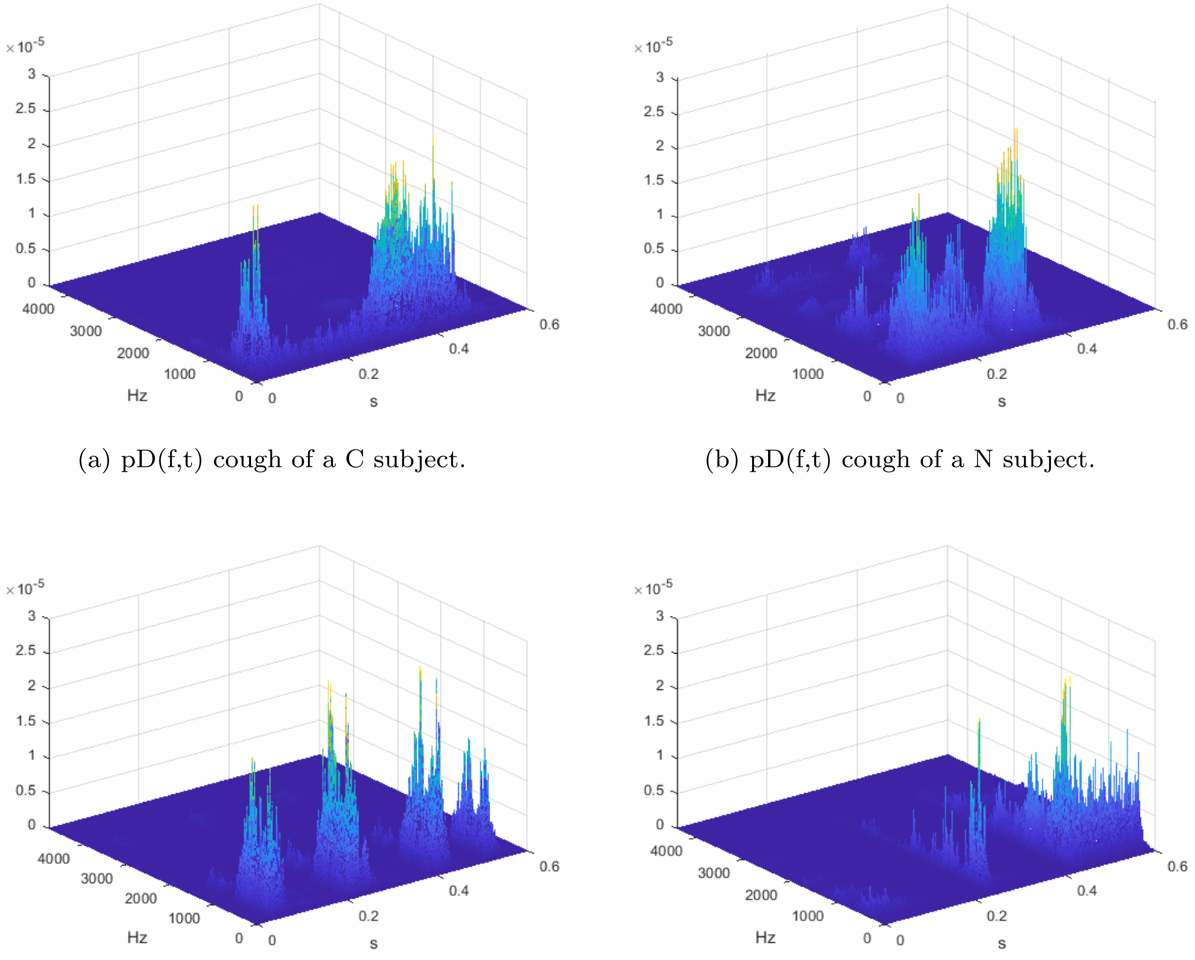
### 3.3. Feature Selection, Feature Extraction and Classification Models

The set of features which obtained the best accuracy by first applying RFE and then Autoencoder to each classification model were selected for analysis. Then, each classification model was applied to these selected features.

#### 3.3.1. RFE

Table 3 shows the classification performance of the classification models fitted with the features selected by RFE tested for the 4 experiments defined.

In the first experiment, C vs. N, the results indicate that RF obtained



(a) pD(f,t) cough of a C subject.

(b) pD(f,t) cough of a N subject.

(c) pD(f,t) cough of a NC subject

(d) pD(f,t) cough of a PT subject

**Fig. 7.** $pD(f,t)$ cough representation of C, N, NC and PT subjects.

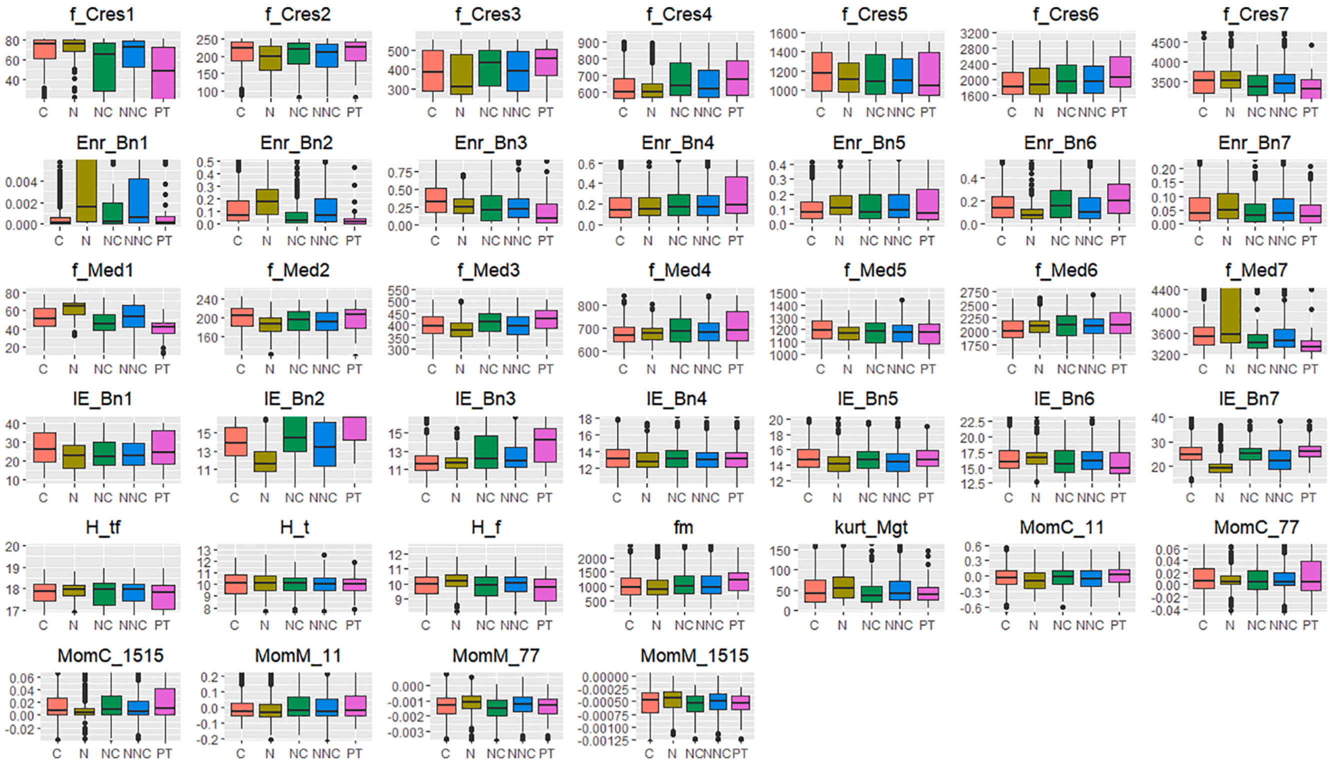


**Fig. 8.** Box plot of the time–frequency features obtained from C, N, NC, NNC and PT groups. Remarkable differences in the mean and standard deviation can be shown.

**Table 3**
Classification performance of the models fitted with the features selected previously with RFE.

| | | C vs. N | C vs. NC | C vs. NNC | C vs. PT |
|---|---|---|---|---|---|
| RF | Accuracy | **89.79** | **88.79** | **85.53** | **94.81** |
| | Sensitivity | **93.81** | 95.49 | **85.96** | **98.91** |
| | Specificity | 81.54 | **76.09** | **85.09** | 72.00 |
| | Precision | 90.97 | **88.42** | **85.14** | 95.20 |
| | F-score | **92.10** | **91.79** | **85.58** | **97.00** |
| | AUC | **96.04** | **92.53** | **89.65** | 95.67 |
| SVM | Accuracy | 83.23 | 78.33 | 74.55 | 89.49 |
| | Sensitivity | 82.57 | 80.22 | 76.79 | 90.65 |
| | Specificity | **84.90** | 74.72 | 72.35 | 83.00 |
| | Precision | **91.84** | 85.76 | 73.80 | 96.75 |
| | F-score | 86.79 | 81.59 | 74.97 | 93.58 |
| | AUC | 92.15 | 88.35 | 75.73 | **97.29** |
| LR | Accuracy | 80.78 | 75.85 | 73.38 | 89.49 |
| | Sensitivity | 79.50 | 77.16 | 74.45 | 90.29 |
| | Specificity | 83.41 | 73.37 | 72.33 | **85.00** |
| | Precision | 90.84 | 84.78 | 73.02 | **97.13** |
| | F-score | 84.67 | 80.68 | 73.49 | 93.56 |
| | AUC | 92.73 | 87.98 | 75.86 | 88.82 |
| NB | Accuracy | 80.78 | 77.86 | 71.96 | 86.41 |
| | Sensitivity | 81.65 | **95.86** | 60.79 | 87.92 |
| | Specificity | 78.98 | 43.73 | 82.99 | 78.00 |
| | Precision | 88.95 | 76.39 | 77.87 | 95.75 |
| | F-score | 84.94 | 85.01 | 68.09 | 91.57 |
| | AUC | 87.50 | 82.07 | 73.94 | 92.06 |
| LDA | Accuracy | 79.56 | 76.32 | 72.32 | 84.16 |
| | Sensitivity | 78.08 | 78.24 | 73.91 | 84.00 |
| | Specificity | 82.68 | 72.68 | 70.57 | **85.00** |
| | Precision | 90.47 | 84.60 | 71.70 | 96.86 |
| | F-score | 83.53 | 81.20 | 72.52 | 89.95 |
| | AUC | 92.69 | 88.38 | 76.04 | 96.71 |

the best overall performance with an *Accuracy* = 89.79%, *Sensitivity* = 93.81, *F-score*= 92.10 and an *AUC* = 96.04. SVM obtained the best *Specificity*, 84.90% and the best *Precision*, 91.84%.

IE_Bn7, Enr_Bn1, IE_Bn2, f_Med1, IE_Bn1 were the top five features obtained with RFE which fitted the model that obtained the best overall performance (RF). The model was also fitted with f_Med7, IE_Bn5, Enr_Bn6, Enr_Bn7, fm, f_Med2, Enr_Bn3, f_Cres7, Enr_Bn5, H_f, kurt_Mgt, Enr_Bn2, IE_Bn6, IE_Bn3, f_Med6, MomC_1515, f_Cres1, f_Med5, f_Med4, f_Cres2, H_t, f_Cres6, Enr_Bn4, IE_Bn4, MomC_11, f_Med3, MomC_77, MomM_11, f_Cres3, H_tf, f_Cres5, MomM_77, MomM_1515 and f_Cres4, which were the set of features selected by RFE.

In the second experiment, C vs. NC, the results indicate that RF obtained the best *Accuracy* = 88.79%, *Specificity* = 76.09%, *Precision* = 88.42%, *F-score*= 91.79% and *AUC* = 92.53. NB obtained the best *Sensitivity* = 95.86%.

IE_Bn3, f_Med7, IE_Bn1, Enr_Bn2 and Enr_Bn3 were the top five features which fitted the model that obtained the best overall performance (RF). The remaining features selected by REF were Enr_Bn1, f_Med1, Enr_Bn4, IE_Bn2, f_Cres1, IE_Bn7, f_Med2, f_Med4, Enr_Bn5, MomM_11, IE_Bn6, f_Cres2, H_t, f_Med6, f_Med5, IE_Bn5, f_Med3, Enr_Bn7, kurt_Mgt, IE_Bn4, H_f, Enr_Bn6, fm, MomM_1515, H_tf, f_Cres6, f_Cres7, MomM77, f_Cres3, f_Cres4, MomC_1515, f_Cres5, MomC_77 and MomC_11.

In the third experiment, C vs. NNC, the results indicate that RF obtained the best *Accuracy* = 85.53%, *Sensitivity* = 85.96, *Specificity* = 85.09, *Precision* = 85.14, *F-score*= 85.58 and *AUC* = 89.65.

IE_Bn3, f_Med7, IE_Bn1, Enr_Bn2 and Enr_Bn3 were the top five features which fitted the model that obtained the best overall performance (RF). The remaining features selected by REF were Enr_Bn1, f_Med1, Enr_Bn4, IE_Bn2, f_Cres1, IE_Bn7, f_Med2, f_Med4, Enr_Bn5, MomM_11, IE_Bn6, f_Cres2, H_t, f_Med6, f_Med5, IE_Bn5, f_Med3, Enr_Bn7, kurt_Mgt, IE_Bn4, H_f, Enr_Bn6, fm, MomM_1515, H_tf, f_Cres6, f_Cres7, MomM77, f_Cres3, f_Cres4, MomC_1515, f_Cres5, MomC_77 and MomC_11.

In the fourth experiment, C vs. PT, the results indicate that RF obtained the best *Accuracy* = 94.81%, *Sensitivity*= 98.91 and *F-score* = 97.00. LR and LDA obtained the best *Specificity* = 85.00, LR obtained the best *Precision*= 97.13 and SVM obtained the best *AUC* = 97.29.

IE_Bn3, Enr_Bn4, Enr_Bn3, IE_Bn2, Enr_Bn2 were the top five features which fitted the RF model which obtained the best overall performance. The remaining features selected by RFE were f_Med1, IE_Bn1, f_Med7, f_Med4, f_Cres1, Enr_Bn1, IE_Bn6, f_Cres2, IE_Bn7, f_Med2 and f_Cres6.

### 3.3.2. Autoencoder

Then, the classification models were fitted with 15 features extracted by means of the Autoencoder. Table 4 shows the classification performance of the classification models tested for the 4 experiments defined.

In the first experiment, C vs. N, the results indicate that RF obtained the best *Accuracy* = 83.67%, *Sensitivity* = 89.58, *F-score* = 88.04 and *AUC* = 93.56. LDA obtained the best *Specificity*, 84.90% and the best *Precision*, 91.13%.

In the second experiment, C vs. NC, the results indicate that RF obtained the best *Accuracy* = 87.73%, *Sensitivity* = 96.94%, *Specificity* = 70.25%, *Precision* = 86.22%, *F-score* = 91.21% and *AUC* = 90.73.

In the third experiment, C vs. NNC, the results show that RF obtained the best *Accuracy* = 79.74%, *Sensitivity* = 79.70, *Specificity* = 79.79, *Precision* = 79.58, *F-score* = 79.52 and *AUC* = 83.57.

In the fourth experiment, C vs. PT, the results indicate that RF obtained the best *Accuracy* = 91.92%, *Sensitivity* = 97.29 and *F-score* = 95.32. LDA obtained the best *Specificity* = 83.00, SVM obtained the best *Precision* = 96.12 and LR obtained the best *AUC* = 95.72.

## 4. Discussion

This research directly addresses a recent statement released by the WHO [1] which believes in the use of rapid tests essential to control people infected with COVID-19. We demonstrated the feasibility of automatic detection of COVID-19 positives from the time–frequency analysis of coughs.

The visual appraisal of the time–frequency representations confirmed differences in the frequency distribution of the voluntary coughs of the C, N, NC and PT subjects.

The features selected by RFE to fit the models obtained better results on the overall performance of the models than those features extracted by means of the Autoencoder. Furthermore, the rank of the features selected by RFE which fitted the model that obtained the best performance depended highly on the experiment done. This means that when comparing coughs, a good selection of the features must be chosen.

The classification models performed better when comparing C vs. PT than when comparing C vs. N, C vs. NC or C vs. NNC, although a good performance was observed for all the experiments. In C vs. PT, the metrics that performed better were *Accuracy* = 94.81%, *Sensitivity* = 98.91% for RF, *Precision* = 97.13% for LR, *F – score* = 97% for RF and *AUC* = 97.29 for SVM. This experiment better detected positive COVID-19 coughs but did not work so well for classifying pertussis coughs (*Specificity* = 85% for LR and LDA). Instead, in the other experiments, the detection of positive and negative cases was more balanced. This was specially so in the C vs. NNC experiment, which obtained the best *Specificity* = 85.09. This experiment reflects a more real case scenario where COVID-19 coughs co-exist with coughs of different patterns. In the four classification experiments done, RF showed the best overall performance.

**Table 4**
Classification performance of the models fitted with the 15 features extracted by means of the Autoencoder.

| | | C vs. N | C vs. NC | C vs. NNC | C vs. PT |
|---|---|---|---|---|---|
| RF | Accuracy | **83.67** | **87.73** | **79.74** | **91.92** |
| | Sensitivity | **89.58** | **96.94** | **79.70** | **97.29** |
| | Specificity | 71.58 | **70.25** | **79.79** | 62.00 |
| | Precision | 86.62 | **86.22** | **79.58** | 93.48 |
| | F-score | **88.04** | **91.21** | **79.52** | **95.32** |
| | AUC | **93.56** | **90.73** | **83.57** | 95.01 |
| SVM | Accuracy | 79.57 | 71.85 | 68.30 | 81.72 |
| | Sensitivity | 77.54 | 72.85 | 67.97 | 81.85 |
| | Specificity | 83.79 | 70.03 | 68.61 | 81.00 |
| | Precision | 91.01 | 82.32 | 68.23 | **96.12** |
| | F-score | 83.36 | 77.18 | 68.00 | 88.22 |
| | AUC | 91.08 | 83.43 | 69.08 | 95.23 |
| LR | Accuracy | 79.21 | 69.97 | 66.60 | 78.98 |
| | Sensitivity | 76.99 | 70.16 | 65.84 | 79.17 |
| | Specificity | 83.79 | 69.65 | 67.36 | 78.00 |
| | Precision | 90.79 | 81.45 | 66.51 | 95.41 |
| | F-score | 83.04 | 75.29 | 66.11 | 86.16 |
| | AUC | 91.16 | 83.32 | 68.61 | **95.72** |
| NB | Accuracy | 76.53 | 73.97 | 70.98 | 84.00 |
| | Sensitivity | 73.04 | 80.21 | 72.66 | 86.15 |
| | Specificity | 83.80 | 62.18 | 69.32 | 72.00 |
| | Precision | 90.16 | 80.20 | 70.25 | 94.55 |
| | F-score | 80.47 | 80.03 | 71.35 | 90.08 |
| | AUC | 89.85 | 81.84 | 73.53 | 95.14 |
| LDA | Accuracy | 77.76 | 71.61 | 67.23 | 82.21 |
| | Sensitivity | 74.30 | 72.85 | 67.25 | 79.71 |
| | Specificity | **84.90** | 69.32 | 67.19 | **83.00** |
| | Precision | **91.13** | 81.92 | 67.00 | 95.73 |
| | F-score | 81.59 | 77.01 | 67.04 | 88.55 |
| | AUC | 91.00 | 82.97 | 68.48 | 95.33 |

### 4.1. Limitations

Although in general, high performance was obtained in RF, its Specificity was not the optimal. Overall, Specificity outcomes were lower. That means that correctly classifying negative samples is an issue. This must be due to classification mistakes in the dataset. Additional efforts must be made to curate the corpus. Furthermore, further analyses comparing COVID-19 cough patterns with cough patterns from other conditions, such as asthma or bronchitis, are needed.

### 4.2. Comparison With Prior Work

Other existing works, such as Laguarta et al. [8], extracted MFCCs from cough recordings and input them into a pre-trained CNN. Their model achieved an AUC of 97% with a *Sensitivity* = 98.5% and a *Specificity* of 94.2%. Pahar et al. [12] presented a machine-learning based COVID-19 cough classifier able to discriminate COVID-19 positive coughs from both COVID-19 negative and healthy coughs recorded on a smartphone. They obtained an AUC of 98% using the Resnet50 classifier to discriminate between COVID-19 positive and healthy coughs, while an LSTM classifier was best able to discriminate between COVID-19 positive and COVID-19 negative coughs with an AUC of 94%. Brown et al. [13] used coughs and breathing to understand how discernible COVID-19 sounds are from those in asthma or healthy controls. Their results showed that a simple binary machine-learning classifier are able to classify healthy and COVID-19 sounds correctly. Their models achieved an AUC of above 80% across all tasks.

The RF model used in this paper performed similarly to the ones used by other authors (Accuracy and AUC close to, or above 90% depending on the experiment) although automated cough detection introduced some performance penalty. Additionally, our methodology allows coughs in samples of raw audio recordings to be detected automatically by using the YAMNet deep neuronal network [17]. We also found the set of time–frequency features that could lead to distinguishing COVID-19 coughs from other cough patterns. In addition, the high performance obtained in various sampling sources (UdL, UC, Virufy and Coswara) validates our method as a more generic proposal.

Newer machine-learning works have shown lower results. For example, an accuracy of 85.2% with RF and 70.6% with CNN, were obtained in [14,15] respectively. Recently [16], an accuracy of 90% was obtained with a recurrent neural network (RNN) by using the Coswara dataset. However, the accuracy dropped to 80% with Coswara and Virufy simultaneously. This fact demonstrates that obtaining good outcomes when different datasets are used is a challenge. Our proposal behaved much better even when three additional datasets (UdL, UC and Pertussis) were used.

## 5. Conclusions

This study demonstrates the feasibility of the automatic detection of COVID-19 from coughs. Excellent results were achieved by fitting an RF model with the set of the time–frequency features selected by RFE for distinguishing COVID-19 coughs. This new methodology presented could lead to automatic identification of COVID-19 by using existing simple and portable devices. It could be the core of a pre-screening mobile app for use as an early response to further COVID-19 outbreaks or other pandemics that may arise in the future.

We will gather more quality data, especially different cough patterns from other conditions, and curate the actual corpus to further train, fine-tune, and improving performance of the models.

## CRediT authorship contribution statement

**Alberto Tena:** Conceptualization, Methodology, Formal analysis, Resources, Software, Data curation, Visualization, Writing - original draft, Validation. **Francesc Clarià:** Conceptualization, Formal analysis, Data curation, Resources, Software, Investigation, Visualization, Validation. **Francesc Solsona:** Writing - review & editing, Supervision, Resources, Project administration, Investigantion, Validation, Funding acquisition.

## 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.

## Acknowledgements

This work was supported by the Intelligent Energy Europe programme and the Ministerio de Economía y Competitividad under contract TIN2017-84553-C2-2-R and the Ministerio de Ciencia e Innovación under contract PID2020-113614RB-C22. We thank Alfredo Jover and Maria Ramirez from the Functional Unit of Nosocomial Infections, and Cristina Acosta from the Internal Medicine, of the Arnau de Vilanova University Hospital in Lleida, who recorded coughs using the website covid.udl.cat. We also thank the University of Cambridge (in particular, Cecilia Mascolo), the Coswara and Virufy projects, who voluntary share their cough datasets. Finally, we also thank everyone who volunteered their cough.

## References

[1] WHO. WHO coronavirus disease (COVID-19) dashboard. https://covid19.who.int. Date accessed: August 10, 2021.

[2] Du. Zhanwei, Abhishek Pandey, Yuan Bai, et al., Comparative cost-effectiveness of SARS-CoV-2 testing strategies in the USA: a modelling study, Lancet Public Health 6 (3) (2021) e184–e191.

[3] Wannian (PRC) Aylward, Bruce (WHO); Liang. Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19), 2020.

[4] J. Martinek, M. Tatar, M. Javorka, Distinction between voluntary cough sound and Speech in volunteers by spectral and complexity analysis, J. Physiol. Pharmacol. 59 (SUPPL. 6) (2008) 433–440.

[5] Hanieh Chatrzarrin, Amaya Arcelus, Rafik Goubran, Frank Knoefel, Feature extraction for the differentiation of dry and wet cough sounds, in: MeMeA 2011–2011 IEEE International Symposium on Medical Measurements and Applications, IEEE, Proceedings, 2011, pp. 162–166.

[6] Renard Xaviero Adhi Pramono, Syed Anas Imtiaz, and Esther Rodriguez-Villegas. A cough-based algorithm for automatic diagnosis of pertussis. PLoS ONE, 11(9):1–20, 2016.

[7] Yusuf Amrulloh, Udantha Abeyratne, Vinayak Swarnkar, and Rina Triasih. Cough Sound Analysis for Pneumonia and Asthma Classification in Pediatric Population. Proceedings - International Conference on Intelligent Systems, Modelling and Simulation, ISMS, 2015-Octob:127–131, 2015.

[8] J. Laguarta, F. Hueto, B. Subirana, COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings, IEEE Open J. Eng. Med. Biol. 1 (2020) 275–281.

[9] Carnegie Mellon University. COVID Voice Detector. https://cvd.lti. cmu.edu/. Date accessed: August 25, 2021.

[10] Vocalis Health. COVID-19 Study. https://vocalishealth.com/. Date accessed: August 25, 2021.

[11] Ali Imran, Iryna Posokhova, Haneya N Qureshi, Usama Masood, Sajid Riaz, Kamran Ali, Charles N John, and Muhammad Nabeel. AI4COVID-19: AI Enabled Preliminary Diagnosis for COVID-19 from Cough Samples via an App. IEEE Access, pages 1–12, 2020.

[12] Madhurananda Pahar, Marisa Klopper, Robin Warren, and Thomas Niesler. COVID-19 Cough Classification using Machine Learning and Global Smartphone Recordings, 2020.

[13] Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, et al. Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 7 2020.

[14] Jayavrinda Vrindavanam, Raghunandan Srinath, Hari Haran Shankar, Gaurav Nagesh, Machine Learning based COVID-19 Cough Classification Models - A Comparative Analysis, in: 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 420–426.

[15] Redacción Médica. Coronavirus: síntomas `falsos' que nada tienen que ver con el Covid-19, 2020.

[16] Ke Feng, Fengyu He, Jessica Steinmann, and Ilteris Demirkiran. Deep-learning Based Approach to Identify Covid-19. In SoutheastCon 2021, pages 1–4, 2021.

[17] YAMNet. https://github.com/tensorflow/models/tree/master/ research/audioset/yamnet. Date accessed: August 25 2021.

[18] Alberto Tena. COVID-19 Models and Data repository. https://github.com/atenad/COVID. Date accessed: August 25, 2021.

[19] Beata Nowok, Gillian M Raab, and Chris Dibben. synthpop: Bespoke Creation of Synthetic Data in R. Journal of Statistical Software, 74(11):1–26, 2016.

[20] University of Cambridge. COVID-19 Sounds App. https://www.covid-19-sounds.org/en/. Date accessed: August 25, 2021.

[21] Indian Institute of Science (IISc) Bangalore. Project Coswara. https://coswara.iisc.ac.in/. Date accessed: August 25, 2021.

[22] Amil Khanzada, Chandan Chaurasia, Nikki Perez, and Lisa Chionis. Virufy. https://virufy.org/. Date accessed: August 25, 2021, 2020.

[23] Matlab. Audio Toolbox. https://github.com/atenad/COVID. Date accessed: August 25, 2021.

[24] J.F. Gemmeke, D.P.W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R.C. Moore, M. Plakal, M. Ritter, Audio Set: An ontology and human-labeled dataset for audio events, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 776–780.

[25] Andrew Howard, Zhu Menglong, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. CoRR, abs/1704.0, 2017.

[26] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, Joel S Emer, Efficient Processing of Deep Neural Networks: A Tutorial and Survey, Proc. IEEE 105 (12) (2017) 2295–2329.

[27] X.X. Sun, W. Qu, Comparison between Mean Filter and Median Filter Algorithm in Image Denoising Field, Appl. Mech. Mater. 644–650 (2014) 4112–4116.

[28] Theodoros Giannakopoulos, A Method for Silence Removal and Segmentation of Speech Signals, Implemented in MATLAB, University of Athens, Athens, 2009.

[29] F. Hlawatsch, G.F. Boudreaux-Bartels, Linear and quadratic time-frequency signal representations, IEEE Signal Process. Mag. 9 (2) (1992) 21–67.

[30] Leon Cohen, Time Frequency Analysis: Theory and Applications, Prentice-Hall, 1995.

[31] Francesc Claria, Montserrat Vallverdú, Rafał Baranowski, Lidia Chojnowska, Pere Caminal, Heart rate variability analysis based on time-frequency representation and entropies in hypertrophic cardiomyopathy patients, Physiol. Measure. 29 (3) (2008) 401–416.

[32] Patrick Loughlin. What are the time-frequency moments of a signal? Proceedings of SPIE - The International Society for Optical Engineering, 4474, 2001.

[33] Payam Refaeilzadeh, Lei Tang, Huan Liu, Cross-Validation, in: Encyclopedia of Database Systems, Springer, US, Boston, MA, 2009, pp. 532–538.

[34] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A Training Algorithm for Optimal Margin Classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT '92, page 144–152, New York, NY, USA, 1992. Association for Computing Machinery.

[35] David Meyer and others. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien, 2019.

[36] Alaa Tharwat. Classification assessment methods. Applied Computing and Informatics, 2018.

## 2.3 Paper 3: Detecting Bulbar Involvement in Patients with Amyotrophic Lateral Sclerosis Based on Phonatory and Time-Frequency Features

- *Alberto Tena, Francesc Clarià, Francesc Solsona, Mònica Povedano. Detecting Bulbar Involvement in Patients with Amyotrophic Lateral Sclerosis Based on Phonatory and Time-Frequency Features. Sensors 22(3):1137. 2022. doi.org/10.3390/s22031137.*

### Abstract

This paper suggested using phonatory-subsystem [59] and time-frequency [85] features jointly. This also is our hypothesis and main contribution. These features, extracted from a portion of the five Spanish vowels, could enhance the performance of the classification models for the early detection of bulbar involvement.

The main goals (and contributions) of this article were:

1. To design a new methodology for the automatic detection of bulbar involvement in males and females based on phonatory-subsystem and time-frequency features.

2. To obtain a set of statistically significant features for diagnosing bulbar involvement efficiently.

3. To analyze the performance of the most common supervised classification models to improve the diagnosis of bulbar involvement.

the accuracy obtained (98.01% for females and 96.10% for males employing a random forest) outperformed the models in the literature.

Adding time-frequency features to more classical phonatory-subsystem features increased the prediction capabilities of the machine-learning models for detecting bulbar involvement.

Studying men and women separately gave greater success. The proposed method could be deployed in any kind of recording device (i.e., smartphone).

sensors

Article

# Detecting Bulbar Involvement in Patients with Amyotrophic Lateral Sclerosis Based on Phonatory and Time-Frequency Features

**Alberto Tena [1], Francesc Clarià [2], Francesc Solsona [2,*] and Mònica Povedano [3]**

[1] CIMNE, Building C1, North Campus, UPC, Gran Capità, 08034 Barcelona, Spain; atena@cimne.upc.edu
[2] Department of Computer Science & INSPIRES, University of Lleida, Jaume II 69, 25001 Lleida, Spain; francisco.claria@udl.cat
[3] Neurology Department, Hospital Universitari de Bellvitge, L'Hospitalet de Llobregat, 08907 Barcelona, Spain; mpovedano@bellvitgehospital.cat
* Correspondence: francesc.solsona@udl.cat; Tel.: +34-973-702-735

**Abstract:** The term "bulbar involvement" is employed in ALS to refer to deterioration of motor neurons within the corticobulbar area of the brainstem, which results in speech and swallowing dysfunctions. One of the primary symptoms is a deterioration of the voice. Early detection is crucial for improving the quality of life and lifespan of ALS patients suffering from bulbar involvement. The main objective, and the principal contribution, of this research, was to design a new methodology, based on the phonatory-subsystem and time-frequency characteristics for detecting bulbar involvement automatically. This study focused on providing a set of 50 phonatory-subsystem and time-frequency features to detect this deficiency in males and females through the utterance of the five Spanish vowels. Multivariant Analysis of Variance was then used to select the statistically significant features, and the most common supervised classifications models were analyzed. A set of statistically significant features was obtained for males and females to capture this dysfunction. To date, the accuracy obtained (98.01% for females and 96.10% for males employing a random forest) outperformed the models in the literature. Adding time-frequency features to more classical phonatory-subsystem features increases the prediction capabilities of the machine-learning models for detecting bulbar involvement. Studying men and women separately gives greater success. The proposed method can be deployed in any kind of recording device (i.e., smartphone).

**Keywords:** ALS; bulbar involvement; voice; diagnosis; phonatory subsystem; time frequency; machine learning

check for updates

**Citation:** Tena, A.; Clarià, F.; Solsona, F.; Povedana, M. Detecting Bulbar Involvement in Patients with Amyotrophic Lateral Sclerosis Based on Phonatory and Time-Frequency Features. *Sensors* **2022**, *22*, 1137. https://doi.org/10.3390/s22031137

Academic Editors: Marcin Woźniak and Sergio A. Velastin

Received: 27 October 2021
Accepted: 28 January 2022
Published: 2 February 2022

## 1. Introduction

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease with an irregular and asymmetric progression, characterized by a progressive loss of both upper and lower motor neurons and that leads to muscular atrophy, paralysis and death, mainly from respiratory failure. The life expectancy of patients with ALS is between 3 and 5 years from the onset of symptoms.

ALS causes muscle weakness and movement, speech, eating and respiratory impediments, leaving the patient reliant on caretakers and relatives and causing considerable social costs. Currently, there is no cure for ALS, although early detection can lead to the use of more appropriate therapies that may slow progress [1].

When the disease starts in the arms and legs, it is called spinal ALS (limb or spinal onset; 80% of cases), and when it starts in the cranial nerve nuclei, it is called bulbar ALS (bulbar onset; 20%). The bulbar muscle is responsible for speech and swallowing, so patients with the later variant have a shorter life expectancy. However, dysarthria, or slurred or difficult speech articulation, affects 80% of all ALS patients [2]. In bulbar ALS, these symptoms usually appear at the onset of the disease, while in spinal ALS, they appear later. Early detection of bulbar involvement in those with ALS is crucial for better diagnosis and prognosis, and could be the key to effectively slowing the development of the disease.

The authors in [3–5] demonstrated that the deterioration of the bulbar muscle affected some phonatory-subsystem features. Among these were jitter, shimmer, harmonic-to-noise ratio (HNR), pitch, formant trajectories, correlations of formants with articulation patterns, fractal jitter, and Mel Frequency Cepstral Coefficients. Consequently, as suggested in previous works [6–13], imperceptible changes in speech and voice can be detected through objective measures.

Time-frequency representation (TFR), broadly applied to detecting several conditions [14–18], has been recently used to detect pathological changes in voice signals [19]. TFR enables the evolution of the periodicity and frequency components to be observed over time, allowing the analysis of non-stationary signals, such as voice signals [20]. The spectrogram is the most common TFR for the analysis of audio signals. This representation corresponds to Cohen's class of time-frequency energy distributions in general. The depiction of a spectrogram is not optimal in terms of resolution quality. There are Cohen class representations that have greater resolution quality. They are all made by smoothing the Wigner distribution, which has the finest resolution but the most detrimental interference. The smoothing functions chosen strike a balance between resolution quality and the elimination of detrimental interference terms.

The authors in [21] used TFR representations from the Cohen class for the onset signal of the vocal fold to diagnose various phonation problems induced by pathological alterations. To assess the voice signal, Cohen class TFRs were combined with a cone kernel distribution to provide optimum smoothness across time. The authors demonstrated that even minor pathogenic alterations in the vocal folds can be seen in TFR, allowing for sensitive affection detection and diagnosis.

In ALS, voice and speech impairment can occur up to 3 years before a diagnosis [22], and when the bulbar muscle function is damaged, voice and speech deteriorates significantly as the disease advances [23]. Features obtained from Cohen class TFRs could aid in the identification of bulbar involvement even earlier than human hearing can.

Centering attention on the subject at hand, R. Norel et al. [24] developed machine-learning models that recognize the presence and severity of ALS based on a variety of frequency, spectral, and voice quality characteristics. An et al. [25] employed Convolutional Neural Networks to classify the intelligible speech of ALS patients compared to healthy people. Finally, Gutz et al. [26] combined SVM and feature filtering techniques.

Based on previous works, and starting from our recent studies [6,18], our paper suggests using phonatory-subsystem [6] and time-frequency [18] features jointly. This also is our hypothesis and main contribution. These features, extracted from a portion of the five Spanish vowels, could enhance the performance of the classification models for the early detection of bulbar involvement, for which the main goals (and contributions) of this research are:

1. To design a new methodology for the automatic detection of bulbar involvement in males and females based on phonatory-subsystem and time-frequency features.
2. To obtain a set of statistically significant features for diagnosing bulbar involvement efficiently.
3. To analyze the performance of the most common supervised classification models to improve the diagnosis of bulbar involvement.

## 2. Methods

### 2.1. Participants

Of the 65 participants selected for this study, 14 of those with ALS had been diagnosed with bulbar involvement (11 females and 3 males; mean = 56.8 years, standard deviation = 12.3 years), 33 had ALS but had not been diagnosed with bulbar involvement (8 females and 25 males; mean = 57.6 years, standard deviation = 12.0 years) and 18 were healthy individuals (9 females and 9 males; mean = 45.2 years, standard deviation = 12.2 years). The main clinical records of the ALS participants are summarized in Table 1. It can be seen that the sample is well age-balanced.

**Table 1.** ALS participants clinical records. Notation: Age (in years). ALSFR-R (Rating Scale-Revised): scores (0–48) the severity of ALS; Bulbar: Bulbar involvement; NA: Data not available.

| Age | Sex | ALSFR-R | Bulbar | Bulbar Onset Symptoms |
|---|---|---|---|---|
| 37 | F | 37 | NO | No Symptoms |
| 38 | M | 6 | YES | NA |
| 39 | M | 43 | NO | No Symptoms |
| 41 | M | 34 | NO | No Symptoms |
| 41 | M | 34 | NO | No Symptoms |
| 43 | F | 21 | YES | Dysphagia |
| 44 | F | 19 | NO | No Symptoms |
| 48 | F | 36 | NO | No Symptoms |
| 48 | F | 29 | YES | Dysphagia |
| 48 | M | 31 | NO | No Symptoms |
| 48 | M | 45 | NO | No Symptoms |
| 49 | M | NA | NO | No Symptoms |
| 50 | M | 39 | NO | No Symptoms |
| 52 | M | 43 | NO | No Symptoms |
| 52 | F | 27 | YES | Dysphagia |
| 52 | M | 33 | NO | No Symptoms |
| 53 | F | 29 | YES | Dysphagia/Dysarthria |
| 55 | M | 26 | NO | No Symptoms |
| 55 | M | 24 | NO | No Symptoms |
| 56 | M | 35 | NO | No Symptoms |
| 56 | M | 27 | NO | No Symptoms |
| 58 | F | 46 | YES | Dysarthria |
| 58 | M | 28 | YES | NA |
| 59 | F | 33 | YES | NA |
| 60 | M | 46 | YES | NA |
| 63 | M | 22 | NO | No Symptoms |
| 63 | M | 42 | NO | No Symptoms |
| 63 | M | NA | NO | No Symptoms |
| 65 | M | 24 | NO | No Symptoms |
| 66 | F | 41 | NO | No Symptoms |
| 67 | M | NA | NO | No Symptoms |
| 67 | F | 33 | YES | Dyspnoea |
| 68 | M | NA | NO | No Symptoms |
| 68 | F | 21 | NO | No Symptoms |
| 69 | M | 37 | NO | No Symptoms |
| 70 | F | 28 | YES | Dysphagia |
| 70 | F | 17 | NO | No Symptoms |
| 70 | M | 46 | NO | No Symptoms |
| 70 | M | 27 | NO | No Symptoms |
| 70 | F | 23 | YES | Dysphagia/Dysarthria |
| 71 | M | 39 | NO | No Symptoms |
| 71 | F | 32 | YES | Dysphagia |
| 72 | M | 30 | NO | No Symptoms |
| 72 | F | 38 | NO | No Symptoms |
| 76 | F | 30 | NO | No Symptoms |
| 81 | M | 36 | NO | No Symptoms |
| 81 | M | 28 | NO | No Symptoms |
| 84 | F | 30 | YES | NA |

The ALS patients' voices were checked by a multidisciplinary clinical team and finally selected by a neurologist for this study.

The control subjects were recruited through personal advertisements in the hospital facilities by the researchers involved in this study. After contacting the volunteers, they received an information sheet explaining the procedure and goal of the study as well as the exclusion criteria. They were interviewed through a questionnaire and those who did not report any voice issue or relevant previous condition were selected for the study.

The control subjects were recruited through personal advertisements conducted in the hospital facilities by the researchers involved in this study. Most of them were companions of ALS patients. After contacting them, control subjects received an information sheet explaining the procedures and goals of the study as well as the exclusion criteria. Control subjects were informed that the study focused on voice analysis to distinguish bulbar involvement in ALS patients. They were interviewed through a questionnaire. Those who did not report any voice issue or relevant previous condition were selected for the study. When they were eligible and still willing to participate, they were invited to come to the hospital room where the voice samples were registered.

### *2.2. Vowel Recording*

There are five vowel segments in the Spanish phonological system (a, e, i, o, u). These were obtained and analyzed from each ALS patient, all of whom were Spanish speakers.

Under medium vocal loudness conditions, each participant uttered a sustained sample of each Spanish vowel for 3–4 s. The recordings were made in a standard hospital room using a laptop and a USB EMITA Streaming GXT 252 microphone calibrated for dBSPL. It has a sensitivity of $-35$ dBSPL and a maximum sound pressure level of 135 dBSPL. The participants sat on a chair with the microphone positioned approximately 30 centimeters from their mouths. The voice signals were recorded using *Audacity*, an open-source application [27], at a sampling rate of 44.100 Hz and 32-bit quantization.

A visual inspection of the spectrograms of the voice signals was conducted similarly to the procedure in [28] to analyze the signal type of the participants' voices. Their results suggested four voice types, of which only type 1 and type 2 were considered suitable for perturbation analysis.

In this study, all the control subjects presented type 1 voice signals, which were periodic without strong modulations or subharmonics. They showed multiple clearly and nearly straight defined harmonics.

Among the 14 ALS patients with bulbar involvement, 10 patients presented type 1 voice signals, which were nearly periodic and showed some clearly defined harmonics. However, a small amount of noise was observed in some voices (four of them). Four of the ALS patients with bulbar involvement presented type 2 voice signals. These had some strong modulations and subharmonics, yet still presented stable and periodic segments in their voices.

Among the 33 ALS patients without bulbar involvement, 29 presented type 1 voice signals, which were nearly periodic and showed multiple or at least some clearly defined harmonics. Instead, four of them presented type 2 voice signals with some strong modulations and subharmonics but still with stable and periodic segments.

It was observed that most of the information of the signal recordings was contained in the range from 0 to 4000 Hz. Therefore, it was decided to decimate all the recording signals sampled at 44.100 Hz using a decimated factor of 5. Signals re-sampled at 8820 Hz were obtained.

Then, each re-sampled signal was standardized by means of the z-score technique. The z-score measures the distance of a signal sample from the mean of the re-sampled signal in terms of the standard deviation. The resulting standardized signal had mean 0 and standard deviation 1, and retained the shape properties of the re-sampled signal. For the re-sampled signal with mean $\overline{X}$ and standard deviation $S$, the z-score of a signal sample $x$ was computed as:

$$z = \frac{(x - \overline{X})}{S} \tag{1}$$

Finally, a segment of 150 ms of each re-sampled and standardized signal ($x(t)$) was chosen for analysis by tacking the midpoint at the center of the phonation.

### *2.3. Phonatory-Subsystem Features*

A total of 15 features from the phonatory subsystem defined in [6,13] were used. They were computed by means of the standard methods used in Praat [29] and the setting details used were the same as in [6]. These features were:

- Fundamental period cycle-to-cycle variation (**Jitter(absolute)**, Equation (2)).

$$Jitter(absolute) = \frac{1}{N-1} \sum_{i=1}^{N-1} |T_i - T_{i-1}|, \tag{2}$$

  where $N$ is the number of cycles and $T_i$ the duration of the $i$th cycle.
- Relative period (**Jitter(relative)**, Equation (3)).

$$Jitter(relative) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |T_i - T_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100 \tag{3}$$

- Relative perturbation (**Jitter(rap)**, Equation (4)).

$$Jitter(rap) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |T_i - \frac{1}{3}\sum_{n=i-1}^{i+1} T_n|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100 \tag{4}$$

- Five-point period perturbation quotient (**Jitter(ppq5)**, Equation (5)).

$$Jitter(ppq5) = \frac{\frac{1}{N-1}\sum_{i=2}^{N-2} |T_i - \frac{1}{5}\sum_{n=i-2}^{i+2} T_n|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100 \tag{5}$$

- Variability of the peak-to-peak amplitude (**Shimmer(dB)**, Equation (6)).

$$Shimmer(dB) = \frac{1}{N-1} \sum_{i=1}^{N-1} |20 \times log\left|\left(\frac{A_{i+1}}{A_i}\right)\right|, \tag{6}$$

  where $A_i$ is the extracted peak-to-peak amplitude data and $N$ is the number of extracted fundamental periods.
- Relative amplitudes of consecutive periods (**Shimmer(relative)**, Equation (7)).

$$Shimmer(relative) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \times 100 \tag{7}$$

- Three-, five- and eleven-point amplitude perturbation (**Shimmer(apqP)**, Equation (8)).

$$Shimmer(apqP) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |A_i - (\frac{1}{P}\sum_{n=i-1}^{i+1} A_n)|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \times 100, \tag{8}$$

  where P = {3, 5 and 11}.
- Mean and standard deviation (**HNR(mean)** and **HNR(SD)**) of the harmonics-to-noise-ratio (HNR, Equation (9)).

$$HNR = 10 \times log_{10} \frac{r(T_0)}{1 - r(T_0)}, \tag{9}$$

where $r(T_0)$ is the second local maximum of the normalized auto-correlation function and $T_0$ is the period of the signal.

- Mean, standard deviation, minimum and maximum value of the pitch (**pitch(mean)**, **pitch(SD)**, **pitch(min)** and **pitch(max)**). See [29] for more details about obtaining the pitch.

### 2.4. Time-Frequency Features

The methods employed to obtain the time-frequency features were inspired by the previous work, presented in [16,17], and implemented with MATLAB [30].

First, the Wigner distribution (WD) of the real signal $x(t)$ of each voice segment was obtained and convoluted with the Choi-Williams exponential function. The resulting Choi-Williams distribution was normalized ($CWD_N(f,t)$). For more details, see [18].

Then, the joint probability density distribution $pD(f,t)$ (Equation (10)) was obtained.

$$pD(f,t) = m_t(t) \cdot m_f(f), \tag{10}$$

where $m_t(t)$, instantaneous power, and $m_f(f)$, spectral energy density, are the marginal density functions of $CWD_N(f,t)$.

According to Equation (10), $pD$ can be only computed as the product of the marginal density functions $m_t(t)$ and $m_f(f)$ (of $CWD_N$) if they are statistically independent. To corroborate this assumption, we computed the joint time-frequency moments of the $CWD_N$ ($\langle t^n f^m \rangle$ from $n = 1$ and $m = 1$ to $n = 15$ and $m = 15$ where $n$ and $m$ are the frequency and time moment orders) of the vowels of all the participants. All of these were 0 or very close to 0. This confirmed the statistical independence of $m_t(t)$ and $m_f(f)$.

$pD(f,t)$ is completely free of interference and negative values. Thus, it is very useful for extracting time-frequency features for classification.

Figure 1 shows the comparison of the $pD(f,t)$ of the vowel "*a*" from three different patients. Non-undesirable effects were observed in the $pD(f,t)$. Figure 1a corresponds to a patient without bulbar involvement. The $pD(f,t)$ shows a voice rich in harmonics. Figure 1b shows the $pD(f,t)$ of the vowel "a" of a patient diagnosed with slight bulbar involvement. Significant differences can be observed. Voice harmonics appear attenuated. Figure 1c shows the $pD(f,t)$ of an even more extreme case, diagnosed with severe bulbar involvement. It can be seen that its voice harmonics appear even more attenuated.The visual appraisal of these figures clearly shows the significant differences in the $pD(f,t)$ between ALS patients with and without bulbar involvement.

From the $pD(f,t)$, a set of 30 features per vowel was obtained. Twenty-one features were computed by dividing the spectrum (0–4410 Hz) into 7 frequency bands. These were 1, 0–80 Hz; 2, 80–250 Hz; 3, 250–550 Hz; 4, 550–900 Hz; 5, 900–1500 Hz; 6, 1500–3000 Hz; 7, 3000–4410 Hz. These bands were selected to capture the differences observed in the time-frequency representations of the two groups of ALS patients by means of the visual appraisal of $pD(f,t)$ in the range of these frequency bands. These features were:

- Average instantaneous spectral energy (E(t), Equation (11)) for each frequency band (**E_Bn1**. . . **E_Bn7**).

$$E(t) = \int_{f_1}^{f_2} pD(f,t)df, \tag{11}$$

where $f_1$ and $f_2$ are the lower and upper frequencies of each band.

- Instantaneous frequency peak ($f\_Cres(t)$, Equation (12)) for each frequency band (**f_Cres1** . . . **f_Cres7**).

$$f\_Cres(t) = \frac{1}{E(t)} argmax_f \left[ \prod_{f_1}^{f_2} f \cdot pD(f,t) \right] \tag{12}$$

- Average instantaneous frequency ($f_{mi}(t)$, Equation (13)) of the spectrum for each frequency band (**f_Med1**. . . **f_Med7**).

$$f_{mi}(t) = \int_{f_1}^{f_2} \frac{1}{E(t)} f \cdot pD(f,t) df \tag{13}$$

10 additional features were added:

- Instantaneous (**H_t**, Equation (14)) and spectral (**H_f**, Equation (15)) information entropies. Furthermore, the joint Shannon entropy (**H_tf**, Equation (16)) was also used.

$$H\_t = -\int log_2(m_{tN}(t)) \cdot m_{tN}(t) dt, \tag{14}$$

where $m_{tN}(t)$ is the quantified instantaneous marginal obtained from the $m_t(t)$ and $m_{fN}(f)$ is the quantified frequency marginal obtained from the $m_f(f)$.

$$H\_f = -\int log_2(m_{fN}(f)) \cdot m_{fN}(f) df \tag{15}$$

$$H\_tf = H\_t + H\_f \tag{16}$$

- Spectral information (IE(f), Equation (17)), for each frequency band (**IE_Bn1**. . . **IE_Bn7**).

$$IE(f) = -log_2(m_{fN}(f)) \tag{17}$$

- Kurtosis (**K**, Equation (18)).

$$K = \left\langle m_t(t)^n m_f(f)^m \right\rangle, \tag{18}$$

where $n = 4$ and $m = 0$.

- Joint time-frequency moment ($\langle t^n f^m \rangle$, [18,31]) where $n$ and $m$ (n, m = 1, 7, 15) are the frequency and time moment orders, i.e., the following time-frequency moments were used: $\langle \mathbf{t^1 f^1} \rangle$, $\langle \mathbf{t^7 f^7} \rangle$ and $\langle \mathbf{t^{15} f^{15}} \rangle$.

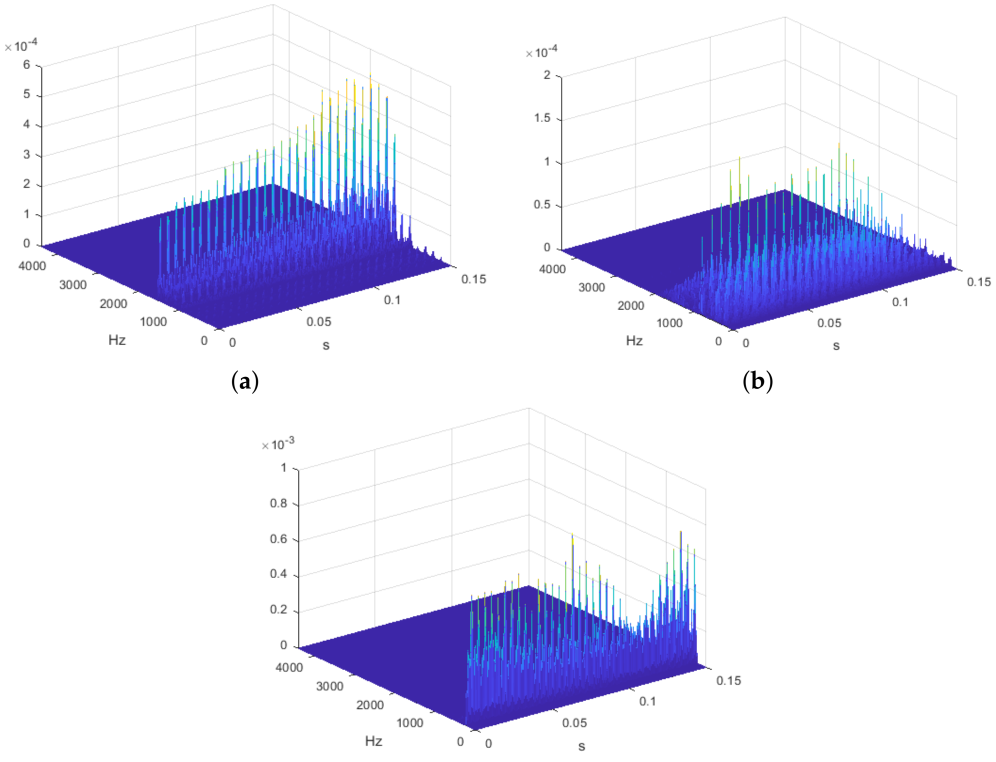
### *2.5. Feature Selection*

From a total of 65 participants, 18 were labelled C (healthy group), 14 were labelled B (ALS patients with bulbar involvement) and 33 were labelled NB (ALS patients without bulbar involvement). Furthermore, every ALS participant was labelled A.

Accordingly, four classification problems were analyzed, males and females being studied separately, these being C vs. B, C vs. NB, B vs. NB and C vs. A.

The Multivariant Analysis of Variance (MANOVA), which uses the covariance between the features in testing the statistical significance of the mean differences, was performed in IBM SPSS Statistics [32] to select a subset of relevant features for use in constructing the classification model for these four cases. This procedure made it possible to contrast the null hypothesis in the features obtained.

To perform this statistical analysis, it was assumed that the features had a multivariable normal distribution, and no assumptions were made regarding the homogeneity of the variance or the correlation between the features. A significance value of $p$-value < 0.05 was considered sufficient to assume the existence of feature differences between the four groups analyzed.



(**a**) (**b**)

(**c**)

**Figure 1.** $pD(f,t)$ of vowel "*a*" for 3 different patients with bulbar involvement. The marked difference in the graphic representation of the time-frequency between the subjects can be clearly appreciated. (**a**) Patient pD without bulbar involvement. (**b**) Patient pD with slight bulbar involvement. (**c**) Patient pD with severe bulbar involvement.

### *2.6. Classification Models*

Several supervised classification models were implemented in R [33] to measure the classification performance. These models were fitted with the features selected. These were standardized by subtracting the mean and centered at 0. Ten-fold cross-validation was implemented in R using the caret package to draw suitable conclusions. This consisted of dividing the dataset into 10 contiguous chunks, each containing approximately the same number of samples, and then performing 10 training-testing experiments as follows: for each chunk $i \in \{1, 2, \ldots, 10\}$, the current chunk was retained for testing the model and training was performed on the remaining 9 chunks, recording the results. The average performance of the 10 training-testing experiments was finally provided.

The upsampling technique with replacement was applied to the training data by making the group distributions equal to deal with the unbalanced dataset, which could bias the classification models [34].

The supervised models with classification thresholds of 50% were built in R [33]. In binary classification problems, the classification threshold is a value that converts the model prediction to positive or negative depending on whether the prediction is above or below the threshold.

The classification algorithms used were the most popular ones in ALS: Support Vector Machine (SVM), Neural Networks (NN), Linear Discriminant Analysis (LDA), Logistic Regression (LR) and Random Forest (RF). For more details, see [6].

### *2.7. Model Validation Metrics*

There are various metrics for evaluating classification models [35]. The foremost among these, accuracy, sensitivity and specificity, were used to evaluate the performance of the classification models.

## 3. Results

First, the significant features from the four cases (C vs. B, C vs. NB, B vs. NB and C vs. A) were selected. Then, the performance of the classification models was evaluated.

### *3.1. Selecting the Significant Features*

From the 50 features obtained, the MANOVA analysis was applied to select those that were statistically significant. Four comparisons were analyzed separately for males and females: C vs. B, C vs. NB, B vs. NB and C vs. A. Features not showing statistical significance ($p$-value $\geq 0.05$) were discarded.

Table 2 shows the significant features obtained for males. In the C vs. B case, this was a set of 12 statistically (half phonatory) significant features ($p$-value < 0.05); in C vs. NB, there were 13 (10 of them phonatory); in B vs NB, 9 (all time-frequency); and in C vs. A, 12 (10 of which were phonatory).

For females (Table 3), in the C vs. B case, a set of 20 statistically significant features ($p$-value < 0.05) was obtained (13 out of 20 were phonatory). In the C vs. NB case, a set of 10 statistically significant features was obtained (6 of them, phonatory). In the case B vs. NB, a set of 14 statistically significant features was obtained (12 of which were phonatory). In the C vs. A case, 20 statistically significant features were obtained (12 being phonatory).

### *3.2. Classification Models*

The classification models were fitted with the significant features selected in Section 3.1. Tables 4 and 5 show the classification performance for males and females, respectively. The results are presented for the *accuracy*, *sensitivity* and *specificity* of the models used for the four cases.

For males in C vs. B case, all the classifiers generally performed well. RF obtained the best *accuracy*, 96.1%. For LDA and NN, *accuracy* was 95.0% and for SVM and LR, 93.3% and 91.9% respectively. LR gave the best *sensitivity* (95.0%), and RF and LDA the best $specificity = 97.5\%$.

Similar performance was achieved in C vs. NB and C vs. A cases. In these, SVM was the best model (an *accuracy* of 93.1% was reached for C vs. NB and 92.6% for C vs. A).

Otherwise, the outcomes worsened in B vs. NB compared with the other cases. Despite RF obtained the best *accuracy* (91.8%), the *sensitivity* it achieved was the worst.

For females, in the C vs. B case, the results also indicate that the performance of all classifiers was excellent. RF gave the best *accuracy*, 98.1%, *sensitivity*, 96.6%, and *specificity*, 100%.

Similar behavior was obtained in the C vs. NB and C vs. A cases. In these, RF was also the best model (obtaining *accuracy* of 94.1% and 95.8% for C vs. NB and C vs. A respectively). In both cases, LDA achieved the best *specificity*.

Meanwhile, the results were worse in B vs. NB compared with the other cases. Although RF obtained the best *accuracy* at 84.8%, the outcomes obtained with it for *specificity* and especially *sensitivity* were very low.

In general, the best model was RF. Special attention should be paid to female outcomes. Poor results were obtained for both genders in the B vs. NB case.

## 4. Discussion

### *4.1. Principal Findings*

The results obtained demonstrate that it is possible to diagnose bulbar involvement using supervised gender-specific models fitted to the significant phonatory and time-frequency features.

**Table 2.** Significant Features for males.

| Comparison | Feature | *p*-Value |
|---|---|---|
| C vs. B | *shimmer(dB)* | 0.039 |
| | *shimmer(apq11)* | <0.001 |
| | *pitch(mean)* | 0.001 |
| | *pitch(SD)* | 0.023 |
| | *pitch(min)* | 0.016 |
| | *pitch(max)* | <0.001 |
| | *f_Cres2* | 0.046 |
| | *f_Cres6* | 0.046 |
| | *f_Med2* | <0.001 |
| | *f_Med6* | 0.008 |
| | *K* | 0.027 |
| | $\langle t^1 f^1 \rangle$ | 0.002 |
| C vs. NB | *jitter(relative)* | 0.008 |
| | *shimmer(dB)* | 0.001 |
| | *shimmer(relative)* | 0.008 |
| | *shimmer(apq3)* | 0.035 |
| | *shimmer(apq11)* | <0.001 |
| | *pitch(mean)* | 0.001 |
| | *pitch(SD)* | 0.002 |
| | *pitch(min)* | 0.023 |
| | *pitch(max)* | 0.001 |
| | *HNR(mean)* | 0.037 |
| | *IE_Bn1* | 0.045 |
| | *H_tf* | 0.015 |
| | *H_f* | 0.045 |
| B vs. NB | *f_Cres1* | 0.044 |
| | *f_Cres2* | 0.028 |
| | *f_Med2* | <0.001 |
| | *f_Med6* | 0.011 |
| | *f_Med7* | 0.024 |
| | *H_tf* | 0.009 |
| | *H_f* | 0.009 |
| | *K* | 0.045 |
| | $\langle t^1 f^1 \rangle$ | <0.001 |
| C vs. A | *jitter(relative)* | 0.009 |
| | *shimmer(dB)* | 0.001 |
| | *shimmer(relative)* | 0.009 |
| | *shimmer(apq3)* | 0.044 |
| | *shimmer(apq11)* | <0.001 |
| | *pitch(mean)* | 0.001 |
| | *pitch(SD)* | 0.002 |
| | *pitch(min)* | 0.015 |
| | *pitch(max)* | <0.001 |
| | *HNR(mean)* | 0.046 |
| | *H_tf* | 0.048 |
| | $\langle t^1 f^1 \rangle$ | 0.034 |

**Table 3.** Significant Features for females.

| Comparison | Feature | *p*-Value |
|---|---|---|
| C vs. B | *jitter(relative)* | 0.001 |
| | *jitter(absolute)* | <0.001 |
| | *jitter(rap)* | <0.001 |
| | *jitter(ppq5)* | <0.001 |
| | *shimmer(relative)* | <0.001 |
| | *shimmer(dB)* | <0.001 |
| | *shimmer(apq3)* | <0.001 |
| | *shimmer(apq5)* | <0.001 |
| | *shimmer(apq11)* | <0.001 |
| | *pitch(mean)* | <0.001 |
| | *pitch(SD)* | <0.001 |
| | *pitch(max)* | <0.001 |
| | *HNR(mean)* | <0.001 |
| | *f_Cres2* | 0.004 |
| | *f_Cres6* | 0.029 |
| | *f_Cres7* | 0.020 |
| | *E_Bn2* | 0.003 |
| | *f_Med2* | <0.001 |
| | *f_Med6* | 0.013 |
| | $\langle t^1 f^1 \rangle$ | 0.028 |
| C vs. NB | *jitter(absolute)* | <0.001 |
| | *shimmer(apq11)* | <0.001 |
| | *pitch(mean)* | <0.001 |
| | *pitch(SD)* | 0.003 |
| | *pitch(min)* | 0.008 |
| | *pitch(max)* | <0.001 |
| | *f_Cres7* | 0.011 |
| | *E_Bn2* | 0.015 |
| | *f_Med1* | 0.014 |
| | $\langle t^7 f^7 \rangle$ | 0.022 |
| B vs. NB | *jitter(relative)* | <0.001 |
| | *jitter(absolute)* | <0.001 |
| | *jitter(rap)* | <0.001 |
| | *jitter(ppq5)* | <0.001 |
| | *shimmer(relative)* | <0.001 |
| | *shimmer(dB)* | <0.001 |
| | *shimmer(apq3)* | <0.001 |
| | *shimmer(apq5)* | <0.001 |
| | *shimmer(apq11)* | <0.001 |
| | *pitch(SD)* | <0.001 |
| | *pitch(max)* | 0.029 |
| | *HNR(mean)* | <0.001 |
| | *H_tf* | 0.026 |
| | *H_f* | 0.048 |
| C vs. A | *jitter(relative)* | <0.001 |
| | *jitter(rap)* | 0.001 |
| | *jitter(ppq5)* | 0.004 |
| | *shimmer(relative)* | <0.001 |
| | *shimmer(dB)* | <0.001 |
| | *shimmer(apq3)* | <0.001 |
| | *shimmer(apq5)* | 0.001 |
| | *shimmer(apq11)* | <0.001 |
| | *pitch(mean)* | <0.001 |
| | *pitch(SD)* | <0.001 |
| | *pitch(max)* | <0.001 |
| | *HNR(mean)* | 0.003 |
| | *f_Cres2* | 0.006 |
| | *f_Cres7* | 0.005 |
| | *E_Bn2* | 0.003 |
| | *f_Med1* | 0.049 |
| | *f_Med2* | 0.001 |
| | *f_Med7* | 0.049 |
| | *H_t* | 0.039 |
| | $\langle t^1 f^1 \rangle$ | 0.018 |

In the case of B vs. C, the *accuracy* achieved was up to 98.1% (RF) and 96.1% (RF) for females and males, respectively.

Lower performance was obtained in C vs. NB but this was still higher than expected. The voice performance in C or NB should be similar. This indicates that some participants in the NB group were probably incorrectly diagnosed. This is coherent with [6]. Similarly, the excellent performance achieved in C vs. A suggests that some of the members of A (14 out of 47) have bulbar involvement. Alternatively, although the most stable segments of the voice samples were selected for analysis, many co-articulatory effects could have influenced the results. Moreover, phonatory-subsystem features are subject to inherently large variability, even for Cs.

On the whole, huge uncertainty was observed in the evaluation concerning bulbar involvement among the participants in the NB group. The case of B vs. NB disclosed that the models did not differentiate between the B and NB subject groups as well as they did with the other groups. RF achieved the best overall performance (accuracy = 91.8%) in males. However, the model presented problems for spotting positive cases (sensitivity = 55.0%). In females, RF achieved an *accuracy* of 84.8%. These values are still far from the ones obtained in the C vs. B case. These outcomes additionally reinforce the idea that NB subjects were misdiagnosed.

The outcomes of each comparison between groups depend on the significant features chosen (between phonatory and time-frequency). In other words, the optimal results in each experiment are obtained with an ad-hoc set of features. This means the differentiation between the participants in different groups depends on different features. However, classifiers obtained very similar results for each experiment, showing a lesser influence.

The results obtained proved that combining phonatory-subsystem and time-frequency features improves the ability of the machine-learning models to detect bulbar involvement. In addition, detecting bulbar involvement also depends on the ad-hoc set of significant features found for such a case.

**Table 4.** Performance of male models. RF: Random Forest; LR: Logistic Regression; LDA: Linear Discriminant Analysis; NN: Neuronal Networks; SVM: Support Vector Machines.

| | | C vs. B | C vs. NB | B vs. NB | C vs. A |
|---|---|---|---|---|---|
| RF | Accuracy | **96.1** | 91.9 | **91.8** | 92.0 |
| | Sensitivity | 86.1 | **92.1** | 55.0 | **93.8** |
| | Specificity | **97.5** | 91.0 | **97.5** | 87.0 |
| LR | Accuracy | 91.9 | 89.2 | 88.5 | 91.3 |
| | Sensitivity | **95.0** | 90.3 | 75.0 | 90.7 |
| | Specificity | 92.0 | 86.9 | 89.5 | 94.0 |
| LDA | Accuracy | 95.0 | 91.1 | 81.3 | 92.0 |
| | Sensitivity | 85.0 | 88.6 | **90.0** | 90.7 |
| | Specificity | **97.5** | 98.0 | 80.5 | 96.0 |
| NN | Accuracy | 95.0 | 90.0 | 86.1 | 92.0 |
| | Sensitivity | 90.0 | 91.3 | 75.0 | 91.5 |
| | Specificity | 95.0 | 86.5 | 88.4 | 93.0 |
| SVM | Accuracy | 93.3 | **93.1** | 86.1 | **92.6** |
| | Sensitivity | 85.0 | 91.2 | 85.0 | 90.7 |
| | Specificity | 95.0 | **98.0** | 86.7 | **98.0** |

### *4.2. Comparison with Prior Work*

This study is consistent with [6–8,36] which demonstrated that such phonatory-subsystem features as jitter, shimmer, pitch and HNR were sensitive indicators for describing pathological voices in ALS. It is also consistent with [6] where great uncertainty was found in the diagnosis of NBs participants.

**Table 5.** Performance of female models. RF: Random Forest; LR: Logistic Regression; LDA: Linear Discriminant Analysis; NN: Neuronal Networks; SVM: Support Vector Machines.

| | | C vs. B | C vs. NB | B vs. NB | C vs. A |
|---|---|---|---|---|---|
| RF | Accuracy | **98.1** | **94.1** | **84.8** | **95.8** |
| | Sensitivity | **96.6** | **92.5** | **92.3** | **95.8** |
| | Specificity | **100** | 95.5 | **75.0** | 96.0 |
| LR | Accuracy | 91.4 | 93.0 | 74.7 | 91.3 |
| | Sensitivity | 91.3 | 90.0 | 75.0 | 93.4 |
| | Specificity | 91.5 | 95.5 | **75.0** | 87.0 |
| LDA | Accuracy | 93.1 | 90.4 | 72.1 | 90.7 |
| | Sensitivity | 87.6 | 82.5 | 70.0 | 87.3 |
| | Specificity | 86.6 | **97.5** | **75.0** | **98.0** |
| NN | Accuracy | 93.2 | 86.9 | 71.1 | 90.6 |
| | Sensitivity | 93.3 | 85.0 | 72.3 | 93.6 |
| | Specificity | 94.0 | 89.0 | 70.0 | 84.5 |
| SVM | Accuracy | 95.1 | 91.6 | 74.2 | 93.6 |
| | Sensitivity | 93.3 | 90.0 | 73.6 | 94.7 |
| | Specificity | 97.5 | 93.0 | **75.0** | 91.5 |

Besides the 15 phonatory-subsystem features obtained in [6], this study also provides 35 time-frequency features. The combination of phonatory-subsystem and time-frequency features, after performing MANOVA for feature selection, enhanced the outcomes of [6], which achieved the best results to date for detecting bulbar involvement in ALS using only acoustic features, ahead of [8,13,24].

*Accuracies* of up to 98.1% (RF) and 96.1% (RF) for females and males respectively were achieved when comparing the bulbar and control participants (case B vs. C). This *accuracy* exceeded the one obtained in [24] with SVM (79.0%) by 17.1% for males and 15.1% for females. The other studies found did not distinguish the classification problems by gender. In [6], SVM obtained an accuracy of 95.8%. In [13], NN based on Mel Frequency Cepstral Coefficients (coefficients for speech representation based on human auditory perception) obtained 90.7%. In [8], NN based on phonatory-subsystem features obtained 91.7% and adding motion sensors for both lip and tongue data increased the accuracy to 96.5% at the expense of including more invasive measurements. For females, our results outperformed those from the aforementioned studies by 2.3%, 7.4% and 6.4% respectively. For males, ours were 0.3% above those obtained in [6] and 5.4% and 4.4% above those obtained in [8,13].

When comparing ALS patients diagnosed with bulbar involvement with those patients in whom bulbar involvement has yet to be detected (B vs. NB), the outcomes outperformed the ones obtained in [6]. The respective accuracy for males and females increased by 16.3% and 9.3% with the same classifier (RF) (91.8% and 84.8% as against 75.5%). This is an important outcome which indicates that the use of time-frequency features increases the identification of bulbar involvement among patients with ALS.

The outcomes obtained in the C vs. NB and C vs. A cases were very similar to those in [6], reinforcing the idea that some NBs could have bulbar involvement.

The most important gains were obtained when comparing B and NB. The selection of the significant features for this comparison improved the outcomes. Thus, involvements (i.e., bulbar) could be detected through a separate, and more closely adjusted, set of features. Consequently, by increasing the identification of particular features, treatment could be better customized for each ALS patient.

In addition, only studies showing C vs. B have been presented in the literature (except in [6]). No attempts to distinguish other subjects have been made to date. We highlight this differentiating issue, and the importance of future research into it.

### 4.3. Limitations

The use of classification models with small datasets hinders the full assessment of the importance of the findings. The size of the dataset was, in part, determined by the low prevalence of ALS, which is considered a rare disease. The small number of samples in the B group was influenced by the heterogeneity of the ALS disease in which patients' symptomatology is very diverse.

Furthermore, hand editing the segments of the voice recordings is inherently subjective and may introduce subtle and unintended selection biases. Although automatic instruments have been created, these methods are currently insufficiently accurate and require manual correction.

## 5. Conclusions and Future Work

This research directly addresses a recent statement released by the NEALS bulbar subcommittee regarding the need for methodologies based on objective measurements [37]. The outcomes achieved reinforce the idea that machine learning can be a suitable tool for helping with the diagnosis of ALS with bulbar involvement using common recording or mobile (i.e., smartphone) devices.

We demonstrate the usefulness of assessing bulbar involvement properly using phonatory-subsystem and time-frequency features from a study of the Spanish vowels that outperformed previous works, specifically [6,8,13,24]. It was also demonstrated that group identification depends on the significant features found for such an experiment.

The main contribution is the differentiation of diagnosis by gender. This outperformed all the results in the literature.

The next steps of this work will consist of improving the corpus for diagnosing bulbar dysfunction. It is planned to increase the sample size and enhance the annotation of the ALS patients without bulbar involvement. Novel methods based on the creation of vowel patterns and semi-supervised classification models will be developed to provide hints for distinguishing those ALS patients without bulbar involvement who may have been misdiagnosed.

Vowel patterns could be generated from the quasi-periodic components of a short stable segment of the five Spanish vowels. Principal and independent component analysis of these patterns is also envisioned.

Moreover, additional research is required to develop this concept properly. Longitudinal research studies are conceived in which patients' diagnoses are obtained at multiple follow-ups. Several repetitions of the sustained phonations will be required to minimize sampling variability even for the control subjects.

**Author Contributions:** Conceptualization, A.T., F.C. and F.S.; methodology, A.T. and F.C.; software, A.T. and F.C.; validation, A.T., F.C., F.S. and M.P.; formal analysis, A.T. and F.C.; investigation, A.T., F.C. and F.S.; resources, A.T., F.C., F.S. and M.P.; data curation, A.T. and F.C.; writing—original draft preparation, A.T. and F.S.; writing—review and editing, A.T. and F.S.; visualization, A.T., F.C. and F.S.; supervision, F.S.; project administration, F.S.; funding acquisition, F.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Spanish *Ministerio de Economía, Industria y Competitividad* and the *Ministerio de Ciencia e Innovación* under contracts TIN2017-84553-C2-2-R and PID2020-113614RB-C22, respectively. AT is a member of CIMNE, a Severo Ochoa Centre of Excellence (2019-2023) under grant CEX2018-000797-S, funded by MCIN/AEI/10.13039/501100011033.

**Institutional Review Board Statement:** The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Research Ethics Committee for Biomedical Research Projects (CEIm) at the Bellvitge University Hospital in Barcelona. (ref.: 065/19). Date: 29 May 2019.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study. Written informed consent has been obtained from the patient(s) to publish this paper.

**Acknowledgments:** Neurology Department of the Bellvitge University Hospital in Barcelona allowed the recording of the voices of the participants in its facilities. Carlos Augusto Salazar Talavera reported the clinical records of the participants. Marta Fulla and Maria Carmen Majos Bellmunt gave advise about the process of eliciting the sounds.

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

**Sample Availability:** Source and Synthetic Data Repository. All the models implemented in R are available in a GitHub repository in *ELA source files* (https://github.com/atenad/ALS), accessed on 27 January 2022. A synthetic dataset of the features obtained is also provided. It was built in R from the original values using the synthpop package [38].

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| TFR | Time-frequency representation |
| Jitter (absolute) | inter-cycle variation of the fundamental period |
| Jitter(relative) | relative period |
| Jitter(rap) | relative perturbation |
| Jitter(ppq5) | five-point period perturbation quotient |
| Shimmer(dB) | Variability of the peak-to-peak amplitude |
| Shimmer(relative) | relative amplitudes of consecutive periods |
| Shimmer(apqP) | three, five and eleven-point amplitude perturbation |
| HNR | harmonics-to-noise ratio |
| HNR(mean) | mean HNR |
| HNR(SD) | standard deviation of HNR |
| WD | Wigner distribution |
| CWD | Choi-Williams exponential function |
| pD | joint probability density distribution |
| E(t) | average instantaneous spectral energy |
| $f_{mi}(t)$ | average instantaneous frequency |
| $f\_Cres(t)$ | instantaneous frequency peak |
| H_t | instantaneous information entropy |
| H_f | spectral information entropy |
| H_tf | joint Shannon entropy |
| IE(f) | spectral information |
| K | kurtosis |
| $\langle t^n f^m \rangle$ | joint time-frequency moment |
| SVM | Support Vector Machine |
| NN | Neural Networks |
| LDA | Linear Discriminant Analysis |
| LR | Linear Logistic Regression |
| RF | Random Forest |
| C | control group |
| B | group of ALS participants diagnosed with bulbar involvement |
| NB | group of ALS participants not diagnosed with bulbar involvement |
| A | group of ALS participants with or without bulbar involvement |

## References

1. Kiernan, M.C.; Vucic, S.; Cheah, B.C.; Turner, M.R.; Eisen, A.; Hardiman, O.; Burrell, J.R.; Zoing, M.C. Amyotrophic lateral sclerosis. *Lancet* **2011**, *377*, 942–955. [CrossRef]
2. Tomik, B.; Guiloff, R. Dysarthria in amyotrophic lateral sclerosis: A review. *Amyotroph. Lateral Scler. Off. Publ. World Fed. Neurol. Res. Group Mot. Neuron Dis.* **2010**, *11*, 4–15. [CrossRef] [PubMed]
3. Shellikeri, S.; Green, J.R.; Kulkarni, M.; Rong, P.; Martino, R.; Zinman, L.; Yunusova, Y. Speech Movement Measures as Markers of Bulbar Disease in Amyotrophic Lateral Sclerosis. *J. Speech Lang. Heart Res. JSLHR* **2016**, *59*, 887–899. [CrossRef] [PubMed]
4. Lee, J.; Dickey, E.; Simmons, Z. Vowel-Specific Intelligibility and Acoustic Patterns in Individuals With Dysarthria Secondary to Amyotrophic Lateral Sclerosis. *J. Speech Lang. Heart Res.* **2019**, *62*, 1–26. [CrossRef] [PubMed]
5. Carpenter, R.; McDonald, T.; Howard, F. The Otolaryngologic Presentation of Amyotrophic Lateral Sclerosis. *Otolaryngology* **1978**, *86*, ORL-479–ORL-484. [CrossRef]

6. Tena, A.; Clarià, F.; Solsona, F.; Meister, E.; Povedano, M. Detection of Bulbar Involvement in Patients With Amyotrophic Lateral Sclerosis by Machine Learning Voice Analysis: Diagnostic Decision Support Development Study. *JMIR Med. Inform.* **2021**, *9*, e21331. [CrossRef]
7. Silbergleit, A.K.; Johnson, A.F.; Jacobson, B.H. Acoustic analysis of voice in individuals with amyotrophic lateral sclerosis and perceptually normal vocal quality. *J. Voice* **1997**, *11*, 222–231. [CrossRef]
8. Wang, J.; Kothalkar, P.V.; Heitzman, D. Towards Automatic Detection of Amyotrophic Lateral Sclerosis from Speech Acoustic and Articulatory Samples. In Proceedings of the InterSpeech, San Francisco, CA, USA, 8–12 September 2016.
9. Chiaramonte, R.; Luciano, C.; Chiaramonte, I.; Serra, A.; Bonfiglio, M. Multi-disciplinary clinical protocol for the diagnosis of bulbar amyotrophic lateral sclerosis. *Acta Otorrinolaringol.* **2019**, *70*, 25–31. [CrossRef]
10. Tomik, J.; Tomik, B.; Wiatr, M.; Skladzien, J.; Strek, P.; Szczudlik, A. The Evaluation of Abnormal Voice Qualities in Patients with Amyotrophic Lateral Sclerosis. *Neuro-Degener. Dis.* **2015**, *15*, 225–232. [CrossRef]
11. Horwitz-Martin, R.L.; Horwitz-Martin, R.L.; Quatieri, T.F.; Lammert, A.C.; Williamson, J.R.; Yunusova, Y.; Godoy, E.; Mehta, D.D.; Green, J.R. Relation of automatically extracted formant trajectories with intelligibility loss and speaking rate decline in amyotrophic lateral sclerosis. In Proceedings of the InterSpeech, San Francisco, CA, USA, 8–12 September 2016; pp. 1205–1209.
12. Spangler, T.; Vinodchandran, N.V.; Samal, A.; Green, J.R. Fractal features for automatic detection of dysarthria. In Proceedings of the 2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI), Orland, FL, USA, 16–19 February 2017; pp. 437–440. [CrossRef]
13. Suhas, B.; Patel, D.; Rao, N.; Belur, Y.; Reddy, P.; Atchayaram, N.; Yadav, R.; Gope, D.; Ghosh, P.K. Comparison of Speech Tasks and Recording Devices for Voice Based Automatic Classification of Healthy Subjects and Patients with Amyotrophic Lateral Sclerosis. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 4564–4568. [CrossRef]
14. Melia, U.; Vallverdú, M.; Jospin, M.; Jensen, E.W.; Valencia, J.F.; Clariá, F.; Gambus, P.L.; Caminal, P. Prediction of nociceptive responses during sedation by time-frequency representation. In Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, Japan, 3–7 July 2013; pp. 2547–2550. [CrossRef]
15. Melia, U.; Claria, F.; Vallverdu, M.; Caminal, P. Measuring Instantaneous and Spectral Information Entropies by Shannon Entropy of Choi-Williams Distribution in the Context of Electroencephalography. *Entropy* **2014**, *16*, 2530–2548. [CrossRef]
16. Claria, F.; Vallverdú, M.; Baranowski, R.; Chojnowska, L.; Caminal, P. Heart rate variability analysis based on time-frequency representation and entropies in hypertrophic cardiomyopathy patients. *Physiol. Meas.* **2008**, *29*, 401–416. [CrossRef]
17. Clariá, F.; Vallverdú, M.; Riba, J.; Romero, S.; Barbanoj, M.J.; Caminal, P. Characterization of the cerebral activity by time-frequency representation of evoked EEG potentials. *Physiol. Meas.* **2011**, *32*, 1327–1346. [CrossRef] [PubMed]
18. Tena, A.; Clarià, F.; Solsona, F. Automated detection of COVID-19 cough. *Biomed. Signal Process. Control* **2022**, *71*, 103175. [CrossRef] [PubMed]
19. Mika, D.; Józwik, J. Advanced Time-Frequency Representation in Voice Signal Analysis. *Adv. Sci. Technol. Res. J.* **2018**, *12*, 251–259. [CrossRef]
20. Yang, Y.; Peng, Z.; Zhang, W.; Meng, G. Parameterised time-frequency analysis methods and their engineering applications: A review of recent advances. *Mech. Syst. Signal Process.* **2019**, *119*, 182–221. [CrossRef]
21. Davies, M.; Daudet, L. Sparse audio representations using the MCLT. *Signal Process.* **2006**, *86*, 457–470. [CrossRef]
22. Makkonen, T.; Ruottinen, H.; Puhto, R.; Helminen, M.; Palmiol, J. Speech deterioration in amyotrophic lateral sclerosis (ALS) after manifestation of bulbar symptoms: Speech deterioration in ALS. *Int. J. Lang. Commun. Disord.* **2017**, *53*. [CrossRef]
23. Tomik, B.; Krupinski, J.; Glodzik-Sobanska, L.; Bala-Slodowska, M.; Wszolek, W.; Kusiak, M.; Lechwacka, A. Acoustic analysis of dysarthria profile in ALS patients. *J. Neurol. Sci.* **1999**, *169*, 35–42. [CrossRef]
24. Norel, R.; Pietrowicz, M.; Agurto, C.; Rishoni, S.; Cecchi, G. Detection of Amyotrophic Lateral Sclerosis (ALS) via Acoustic Analysis. *bioRxiv* **2018**. [CrossRef]
25. An, K.; Kim, M.; Teplansky, K.; Green, J.; Campbell, T.; Yunusova, Y.; Heitzman, D.; Wang, J. Automatic Early Detection of Amyotrophic Lateral Sclerosis from Intelligible Speech Using Convolutional Neural Networks. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 1913–1917. [CrossRef]
26. Gutz, S.E.; Wang, J.; Yunusova, Y.; Green, J.R. Early Identification of Speech Changes Due to Amyotrophic Lateral Sclerosis Using Machine Classification. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 604–608. [CrossRef]
27. Audacity Manual. Available online: https://manual.audacityteam.org (accessed on 27 January 2022).
28. Sprecher, A.; Olszewski, A.; Jiang, J.J.; Zhang, Y. Updating signal typing in voice: Addition of type 4 signals. *J. Acoust. Soc. Am.* **2010**, *127*, 3710–3716. [CrossRef]
29. Boersma, P.; Weenink, D. *Praat: Doing Phonetics by Computer, V6.1.01*; Technical Report; University of Amsterdam: Amsterdam, The Netherlands 2019.
30. MATLAB. *Version 9.9.0.1495850 (R2020b)*; The MathWorks Inc.: Natick, MA, USA, 2020.
31. Loughlin, P. What are the time-frequency moments of a signal? *Proc. SPIE Int. Soc. Opt. Eng.* **2001**, *4474*, 35–44. [CrossRef]
32. IBM Corp. *IBM SPSS Statistics for Windows*; IBM Corp.: Armonk, NY, USA, 2016.
33. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019.
34. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: Berlin/Heidelberg, Germany, 2013. [CrossRef]
35. Tharwat, A. Classification assessment methods. *Appl. Comput. Inform.* **2021**, *17*, 168–192. [CrossRef]

36. Vashkevich, M.; Petrovsky, A.; Rushkevich, Y. Bulbar ALS Detection Based on Analysis of Voice Perturbation and Vibrato. In Proceedings of the 2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), Poznan, Poland, 18–20 September 2019; pp. 267–272.
37. Plowman, E.K.; Tabor, L.C.; Wymer, J.; Pattee, G. The evaluation of bulbar dysfunction in amyotrophic lateral sclerosis: survey of clinical practice patterns in the United States. *Amyotroph. Lateral Scler. Front. Degener.* **2017**, *18*, 351–357. [CrossRef] [PubMed]
38. Nowok, B.; Raab, G.M.; Dibben, C. synthpop: Bespoke Creation of Synthetic Data in R. *J. Stat. Softw.* **2016**, *74*, 1–26. [CrossRef]

## 2.4 Paper 4: Voice Fingerprint and Machine Learning Models for Early Detection of Bulbar Dysfunction in ALS

- *Submitted to the journal Artificial Intelligence in Medicine.*

We conjectured that the diagnosis of ALS patients with bulbar dysfunction would greatly benefit from the creation of a voice fingerprint able to detect bulbar dysfunction in ALS before the first symptoms can be detected by human hearing. This could be done effectively by means of analyzing a pattern generated from the quasi-periodic waveform produced by the vocal folds when a vowel is elicited. Quasi-periodic waveform analysis has been applied to several clinical applications such as heartbeat detection, cardiopulmonary modeling and intrinsic brain activity detection [37, 38]. Furthermore, performance could be increased by correcting the bias as well as enlarging the corpus upsampling it [60], and relabeling bulbar and non-bulbar ALS patients by using semi-supervised classifiers, as pointed out in [61, 62].

Our objective (and contribution) was to create a machine-learning model obtained by applying supervised and unsupervised classifiers and upsampling to improve the corpus for diagnosing bulbar dysfunction. This will be done through the creation of a voice fingerprint consisting of a pattern generated from the quasi-periodic components of a steady portion of the five Spanish vowels, and the five principal and independent components of this pattern. This model should behave properly with small and usually badly annotated corpus, the kind associated with rare diseases (i.e. ALS without bulbar involvement).

The best model (Random Forest) obtained an accuracy of 88.3% when classifying bulbar vs. control participants. Furthermore, due to the great uncertainty found in the annotated corpus of the ALS patients without bulbar involvement, we used a safe semi-supervised support vector machine (S4VM) to relabel the ALS participants diagnosed without bulbar involvement as bulbar and no-bulbar. The performance of the results obtained increased by 5% (to 93.5%). This demonstrates that our model can improve the diagnosis of bulbar dysfunction compared not only with clinicians, but also the methods published to date.

The results obtained are very encouraging and demonstrate the efficiency and applicability of the methodology presented in this paper. It could be an appropriate tool for screening bulbar involvement in early stages of the disease.

# Voice Fingerprint and Machine Learning Models for Early Detection of Bulbar Dysfunction in ALS

Alberto Tena, Francesc Clarià, Francesc Solsona, Mónica Povedano

***Abstract*—The NEALS bulbar subcommittee released a recent statement regarding the need for objective-based approaches to diagnosing bulbar dysfunction in ALS patients. Bulbar dysfunction is a term used in ALS. It refers to motor neuron disability in the corticobulbar area of the brainstem which leads to a dysfunction of speech and swallowing. One of the earliest symptoms of bulbar dysfunction is voice deterioration characterized by grossly defective articulation, extremely slow laborious speech, marked hypernasality and severe harshness. Recently, research efforts have focused on voice analysis to capture this dysfunction.**

**The main aim of this paper is to provide a new methodology to diagnose this dysfunction automatically at early stages of the disease, earlier than clinicians can do.**

**The study focused on the creation of a voice fingerprint consisting of a pattern generated from the quasi-periodic components of a steady portion of the five Spanish vowels and the computation of the five principal and independent components of this pattern. Then, a set of statistically significant features was obtained and the outcomes of the most common supervised classification models were obtained. The best model (Random Forest) obtained an accuracy of 88.3% when classifying bulbar vs. control participants.**

**In addition, due to the great uncertainty found in the annotated corpus of the ALS patients without bulbar involvement, we used a safe semi-supervised support vector machine (S4VM) to relabel the ALS participants diagnosed without bulbar involvement as bulbar and no-bulbar. The performance of the results obtained increased by 5% (to 93.5%). This demonstrates that our model can improve the diagnosis of bulbar dysfunction compared not only with clinicians, but also the methods published to date.**

**The results obtained are very encouraging and demonstrate the efficiency and applicability of the methodology presented in this paper. It may be an appropriate tool for screening bulbar involvement in early stages of the disease.**

***Index Terms*—ALS, bulbar dysfunction, voice, diagnosis, machine learning.**

## I. Introduction

AMYOTROPHIC lateral sclerosis (ALS) is a neurodegenerative disease characterized by a progressive loss of both upper and lower motor neurons leading to muscular atrophy, paralysis and death. Currently, there is no cure for ALS, although early detection can slow progress [1].

ALS is known as spinal (80%; limb or spinal onset) and bulbar (20%; bulbar onset). The first bulbar symptoms appear early in the disease in bulbar ALS, but can also appear in later stages of spinal ALS. Early detection of bulbar dysfunction may be the key to effectively slowing down the disease. However, diagnosing this is challenging due to the limitations of human hearing [2].

Several studies demonstrated that the voice is one of the most important aspects for detecting bulbar dysfunction. R. Norel et al. [3] developed machine models for recognizing the presence and severity of ALS using a variety of frequency, spectral, and voice quality features. An et al. [4] used Convolutional Neural Networks (CNNs) to classify the intelligible speech produced by patients with ALS compared with healthy individuals. Wang et al. [5] used Support Vector Machines (SVM) and Neuronal Networks (NN) employing acoustic features and adding articulatory motion information (from the tongue and lips). In a recent study [6], we demonstrated the feasibility of automatic detection of bulbar dysfunction through phonatory features obtained from vowel utterance even before it becomes perceptible to human hearing. Great uncertainty in the annotated corpus of the ALS patients without bulbar involvement was found. Although all methods performed well in general, this performance dropped significantly when diagnosing bulbar involvement among ALS patients. The aforementioned study argued that the main causes of this uncertainty was a small and wrongly annotated corpus of the ALS patients without bulbar involvement. This suggests that subjective methods employed by clinicians could lead them to misdiagnose this dysfunction. This is coherent with the NEALS bulbar subcommittee, which calls for objective-based approaches.

We conjecture that the diagnosis of ALS patients with bulbar dysfunction would greatly benefit from the creation of a voice fingerprint able to detect bulbar dysfunction in ALS before the first symptoms can be detected by human hearing. This could be done effectively by means of analyzing a pattern generated from the quasi-periodic waveform produced by the vocal folds when a vowel is elicited. Quasi-periodic waveform analysis has been applied to several clinical applications such as heartbeat detection, cardiopulmonary modeling and intrinsic brain activity detection [7, 8]. Furthermore, performance could be increased by correcting the bias as well as enlarging the corpus upsampling it [9], and relabeling bulbar and non-bulbar ALS patients by using semi-supervised classifiers, as pointed out in [10, 11].

Our objective (and contribution) is to create a machine-learning model obtained by applying supervised and unsupervised classifiers and upsampling to improve the corpus for

A. Tena (corresponding author) is with CIMNE. Building C1, North Campus, UPC. Gran Capità, 08034 Barcelona, Spain. email: atena@cimne.upc.edu. F. Clarià and F. Solsona are with the Department of Computer Science and Industrial Engineering, University of Lleida, Lleida, 25001 Spain. e-mail: francesc.claria,francesc.solsona@udl.cat. M. Povedano is with the Neurology Department, Hospital Universitari de Bellvitge, L'Hospitalet, Barcelona, Spain. email: mpovedano@bellvitgehospital.cat.

This work was supported by the Intelligent Energy Europe (IEE) programme and the Ministerio de Economía y Competitividad under contract TIN2017-84553-C2-2-R. FS is member of the research group 2017-SGR363, funded by the Generalitat de Catalunya.

—

diagnosing bulbar dysfunction. This will be done through the creation of a voice fingerprint consisting of a pattern generated from the quasi-periodic components of a steady portion of the five Spanish vowels, and the five principal and independent components of this pattern. This model should behave properly with small and usually badly annotated corpus, the kind associated with rare diseases (i.e. ALS without bulbar involvement).

## II. Methods

The study was approved by the Research Ethics Committee for Biomedical Research Projects (CEIm) at the Bellvitge University Hospital in Barcelona.

The methods presented in this section were implemented and a synthetic dataset based on a random sample of the corpus is freely available online [12].

### A. Participants

Forty-five ALS participants (26 males and 19 females) aged from 37 to 84 (M = 57.8 years, SD = 11.8 years) and 18 control subjects (9 males and 9 females) aged from 21 to 68 (M = 45.2 years, SD = 12.2 years) took part in this study. All the ALS participants were diagnosed by a neurologist.

Bulbar dysfunction was diagnosed by following subjective clinical approaches [13] and the neurologist made the diagnosis of whether an ALS patient had bulbar dysfunction. Among all the ALS participants, 5 reported bulbar onset and 40 spinal onset. However, at the time of the study, 14 of them presented bulbar symptoms.

To summarize, 14 of the 63 participants were ALS patients diagnosed with bulbar dysfunction (3 males and 11 females) aged from 38 to 84 (M = 56.8 years, SD = 12.3), 31 were ALS patients that did not present this dysfunction (23 males and 8 females) aged from 37 to 81 (M = 58.3 years, SD = 11.7) and 18 were control subjects (9 males and 9 females) aged from 21 to 68 (M = 45.2 years, SD = 12.2 years).

### B. Vowel Recording

Sustained samples of the Spanish vowels, a, e, i, o and u, were elicited under medium vocal loudness conditions for 3-4 seconds. The recordings were made in a regular hospital room using an USB EMITA Streaming GXT 252 microphone connected to a laptop at a sampling rate of 44,100 Hz. 32-bit quantization was done using *Audicity*, an open-source application. Each individual phonation was cut out and anonymously labeled. The boundaries of the speech segments were determined with an oscillogram and a spectrogram using the Praat manual [14], and were audibly checked. The starting point of the boundaries was established at the onset of the periodic energy in the waveform observed in the oscillogram and checked by the appearance of the formants in the spectrogram. The endpoint was established at the end of the periodic oscillation when a marked decrease in amplitude in the periodic energy was observed. It was also identified by the disappearance of the waveform in the oscillogram and the formants in the spectrogram.

### C. Generating the pattern of the quasi-periodic components of the five Spanish vowels

A sample of 250 ms of each vowel was considered for analysis by taking the middle point at the center of the phonation. This fragment of the signal was normalized by centering each sample to have a mean of 0 and scaled to have a standard deviation of 1 ($x(n)$). A pattern generator was developed to obtain a pattern sequence of the quasi-periodic components of the fundamental frequency of $x(n)$ inspired by [15]. This process consisted of 3 steps.

*1) Detrending Method:* The baseline wandering of $x(n)$, which is a low-frequency artefact present in signal recordings, was removed by implementing a detrending method. To obtain the trend, a six-order low-pass Butterworth filter [16] with a cutoff frequency of 0.0035 Hz was applied twice (forward and backward) to $x(n)$ [17, 18]. The combined filter had zero phase distortion, a filter transfer function equal to the squared magnitude of the implemented Butterworth filter transfer function, and a filter order that was double the order of the Butterworth filter. Then, the detrending signal $x_d(n)$ was obtained by removing the trend from $x(n)$. Finally, each sample of $x_d(n)$ was centered to have a mean of 0. Fig. 1a shows $x(n)$ and the trend of $x(n)$ and Fig. 1b shows $x(n)$ and $x_d(n)$.

*2) Marking the quasi-periodic components of $x(n)$:* The spectral density $|X_d(f)|^2$ of $x_d(n)$ (Fig. 2) was obtained by means of the discrete Fourier transform (DFT) implementing the fast Fourier transform (FFT) algorithm. To identify the quasi-periods, the samples of $|X_d(f)|^2$ whose frequency was $\geq 300$ Hz were considered to identify the peaks of the spectral density. To avoid noise, only the three highest peaks were selected. Finally, the quasi-period of $x_d(n)$ was defined as the lower spectral component of these three peaks ($f_r$). The number of samples of each quasi-period ($n_{rep}$) was calculated as the nearest integer of ($f_s/f_r$), $f_s$ being the recording sampling rate.

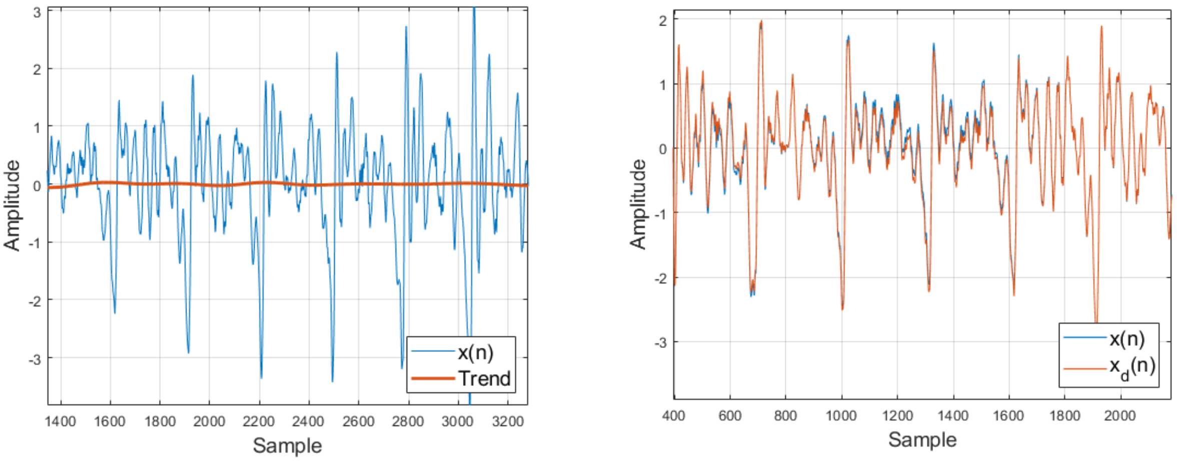
The signal envelop $x_e(n)$ was obtained by computing the cumulative sum of $x_d(n)$ and then calculating the envelope of the analytical signal [19]. Fig. 3 shows $x_e(n)$ and $x(n)$.

To detect the starting and ending point of each quasi-period, a quasi-sinusoidal signal, $s(n)$, synchronized with the period of $x(n)$ was computed. It was obtained by applying a second-order Butterworth pass-band filter forward and backward to $x_e(n)$ with a cut-off frequency $f_c = f_s/n_{rep}$ Hz. From $s(n)$, a quadratic-bipolar signal ($q(n)$) was generated assigning a constant -A in those samples where $s(n) < 0$, and A in those where $s(n) > 0$. Thus, by differentiating $q(n)$, the zero crossings of the synchronized signal $s(n)$ were obtained, which represent the beginning and end of each quasi period of $x(n)$. Fig. 4 illustrates this process. Fig. 4a represents $x(n)$, $s(n)$ and $q(n)$ and Fig. 4b depicts the starting and ending points detected of each quasi-period of $x(n)$.

Finally, the pattern function $p(T)$ was obtained as the average of the quasi-periods of $x(n)$, $T$ being the average of the number of samples of the quasi-period of $x(n)$.

*3) Pattern refinement:* $p(T)$ was compared with $x_d(n)$ to improve the boundaries of each quasi-period. First, as an adapted filter, the pattern $p(T)$ was inverted and the resulting

Fig. 1: Detrending method: Removing the trend from $x(n)$.



(a) Obtaining the trend of $x(n)$

(b) $x(n)$ and $x_d(n)$

Fig. 2: The spectral density of $x_d(n)$.

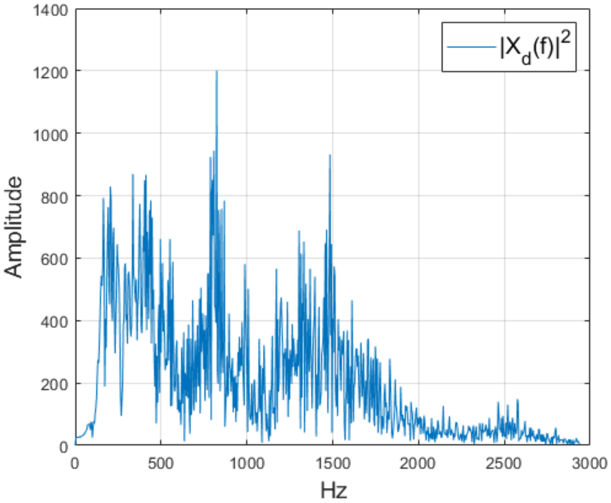
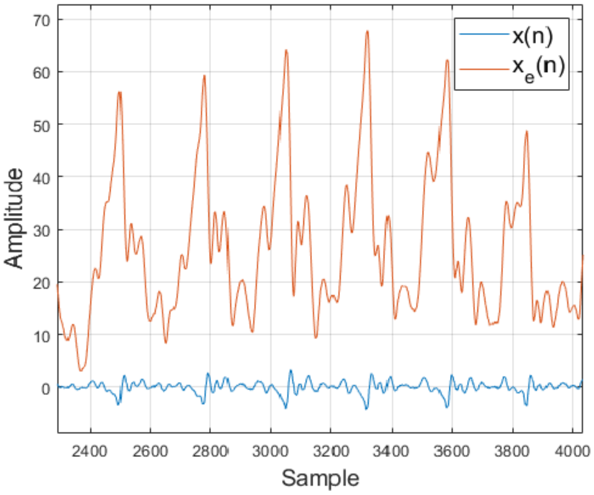


Fig. 3: The signal envelope of $x(n)$.



signal was convolved with $x_d(n)$ to detect the positions of $p(T)$ in $x_d(n)$. The positive values of the resulting signal were taken and the negative values were set at 0.

Each quasi-period detected previously was centered in the position where the maximums values of the convolution were found. The refined pattern, $p_{ref}(T)$ (Fig. 5a) was computed as the average of the quasi-periods of $x_d(n)$ with their new boundaries established.

Finally, $p_{ref}(T)$ was normalized to 550 samples and then decimated to 110 samples to obtain patterns, $p_N(T)$ (Fig, 5b), with the same length.

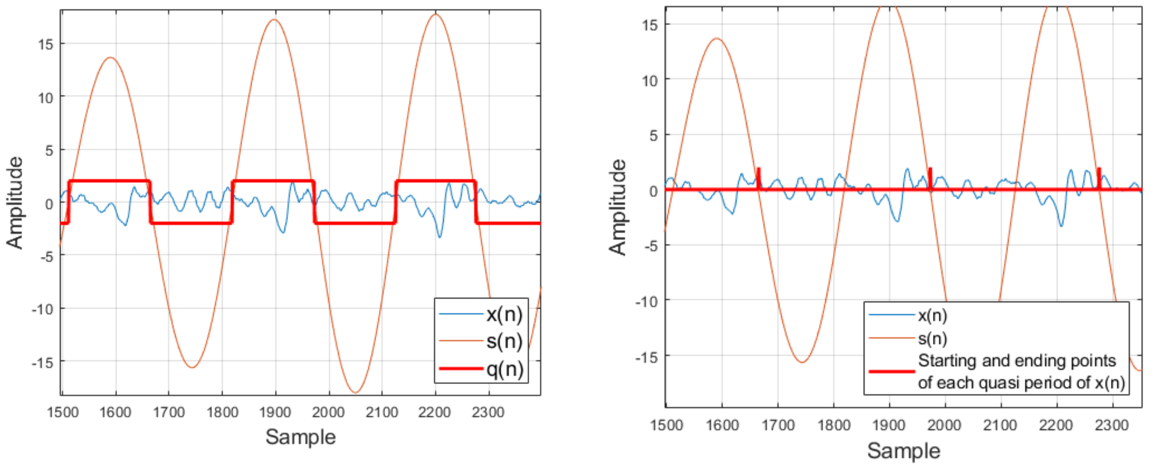
### *D. Principal and Independent Component Analysis*

Principal Component Analysis (PCA) and Independent Component Analysis (ICA) have great potential in the treatment of medical signals [20, 21]. PCA is a classical technique in statistical data analysis, feature extraction and data reduction, aiming at explaining observed signals as a linear combination of orthogonal principal components. ICA is a technique for array processing and data analysis, aiming at recovering unobserved signals from observed mixtures, exploiting only the assumption of mutual independence between the signals.

In the PCA of $p_N(T)$, five Principal Components (PCs) were computed [22]. The decomposition was obtained as $X = USV^\top$ where $X$ is $p_N(T)$ standardized, $U$ is a unitary matrix and $S$ is the diagonal matrix of singular values $s_i$. PCs were given by $US$ and $V$ contained the directions in this space that capture the maximal variance of the matrix $X$.

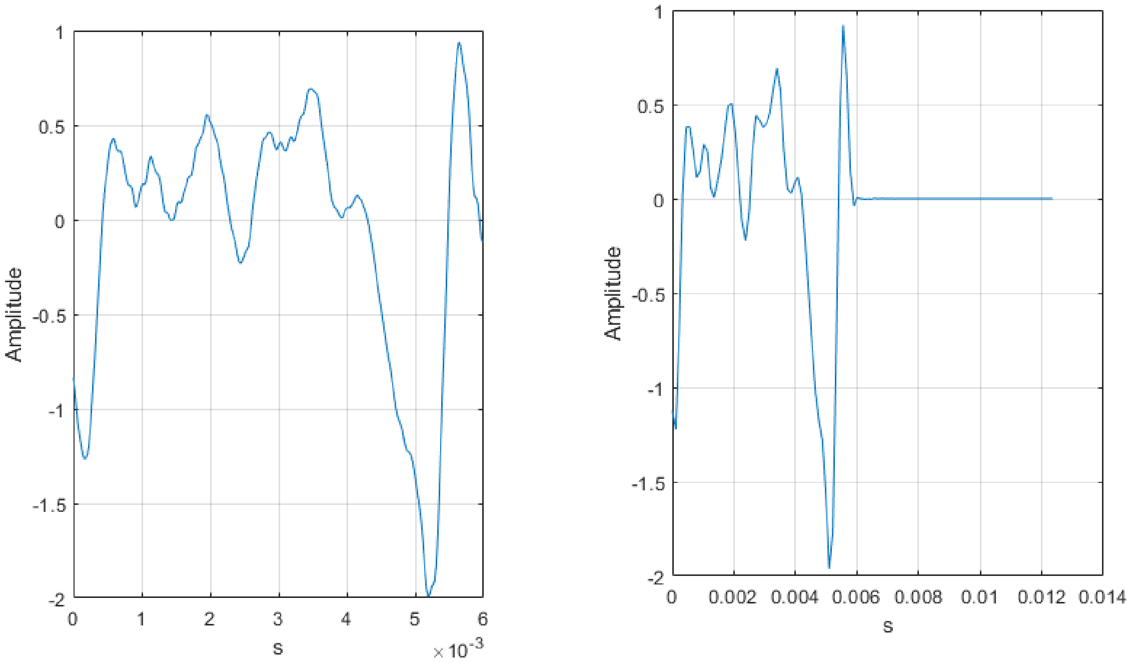
In the ICA of $p_N(T)$, five independent components (ICs) were extracted by means of a reconstruction independent component analysis (RICA) algorithm [23]. $p_N(T)$ was standardized to have zero mean and identity co-variance. The model $x = \mu + As$ is made up of the five rows of matrix $x$

Fig. 4: Detecting the starting and ending point of each quasi-period of $x(n)$.


(a) $x(n)$, $s(n)$ and $q(n)$.

(b) Starting and ending point of each quasi-period of $x(n)$.

Fig. 5: Pattern refinement and normalization.


(a) $p_{ref}(T)$

(b) $p_N(T)$

representing the patterns of the five vowels with 110 samples for each pattern. $\mu$ is a constant represented by a column vector of five rows and $s$ is a matrix in which each row (5) is an independent component with 110 samples. A(5x5) is the mixing matrix. Once the model has been obtained, the five independent components computed were used for the analysis.

### *E. Features obtained for analysis*

A total of 70 features were obtained as follow:

- Three entropy measurements per each vowel were obtained by means of the Shannon entropy:
  - From the probability density function of a signal, formed by the pattern $p_N(T)$ repeated the number of periods of $x(n)$ and quantified with $N = 2^q$ levels and $q = 8$. This measurement was coded as entPat1 . . . entPat5.
  - The density function of the five PCs was normalized and quantified with $N = 2^q$ levels and $q = 6$. Then, the Shannon entropy of the probability density function of the five PCs was obtained and coded as entPC1. . . entPC5.
  - Similarly, to compute the Shannon entropy of the five ICs, the density function of the five ICs was normalized and quantified with $N = 2^q$ levels and $q = 6$ and the Shannon entropy of each IC was obtained. The results were coded as entIC1 . . . entIC5.
- The variance of a signal formed by the pattern $p_N(T)$ repeated the number of periods of $x(n)$ was computed and coded for each vowel as var1 . . . var5.
- The Kurtosis is defined as a measure of outlier-prone. It is calculated from the distribution of a signal formed by the pattern $p_N(T)$ repeating the number of periods of $x(n)$, and coded as kurt1. . . kurt5. The bias-corrected equation defined in [23] was applied to obtain the Kurtosis.
- The rhythm variability, $RR(n)$, of $x(n)$ was computed by firstly calculating the differences (in seconds) between the quasi-periodic elements of $x(n)$ and dividing the result by the sampling frequency (44.100 Hz). Finally, $RR(n)$ was obtained by reducing the sampling to $fr = 350$. Thus, $RR(n)$ is a signal resampled to $fr = 350$ ($Tr = 0.0029$ s) with a bandwidth of 175 Hz. The mean and the standard deviation of $RR$ ($mean\_RR$ and $std\_RR$) were then computed and coded for each vowel as medRR1. . . medRR5 and dsvRR1 . . . dsvRR5 respectively.
- The spectrum of $P\_N(f)$ was obtained from the positive and normalized part of the FFT of the autocorrelation of $p_N(T)$. The mean frequency of $P\_N(f)$ ($fmEsPat$) was computed according to Eq. 1 in the frequency band 0 Hz to 2205 Hz and coded for each vowel as fmEsPat1 . . . fmEsPat5.

$$fmEsPat = \int f P\_N(f)\, df \tag{1}$$

- Similarly, the mean frequency of the probability density function of the five PCs was computed and coded as fmEsPC1 . . . fmEsPC5.
- Finally, the average spectral energy was calculated as the integral of $P\_N(f)$. The average spectral energy was computed and normalized to 1 for 5 frequency bands of the total spectrum (0-4,410 Hz): 1, 0-250 Hz; 2, 250-750 Hz; 3, 750-1500 Hz; 4, 1500-2500 Hz; 5, 2500-4,410 Hz. These measurements for the five patterns and the five bands of each pattern were coded for each vowel as enBnEs_a1 . . . enBnEs_a5 and enBnEs_u1 . . . enBnEs_u1, respectively.

### *F. Classification Models*

Five supervised classification models were implemented in R to measure the classification performance. These models are Random Forest (RF), Logistic Regression (LR), Linear Discriminant Analysis (LDA), Neural Networks (NN) and Support Vector Machine (SVM). The classification models were fitted with the features selected. These were standardized by subtracting the mean and centered at 0. 10-fold cross-validation was implemented in R using the caret package [24] to draw suitable conclusions. The upsampling technique with replacement was applied to the training data by making the group distributions equal to deal with the unbalanced dataset that could bias the classification models [9]. Supervised models with classification thresholds of 50% were built. The classification threshold is a value that converts the result of a quantitative test into a simple binary decision by treating the values above or equal to the threshold as positive, and those below as negative.

In addition, the semi-supervised classification model S4VM was implemented using the RSSL package [25]. S4VM returns predicted labels for unlabeled instances. It randomly generates multiple low-density separators and merges their predictions by solving a linear programming problem meant to penalize the cost of decreasing the performance of the classifier, compared to the supervised SVM [26]. As for SVM, a linear kernel was used, and the regularization parameter C for labeled and unlabeled data was set at 0.05.

### *G. Feature Selection*

To select a subset of relevant features for use in the construction of the classification model, the Multivariant Analysis of Variance (MANOVA), which uses the covariance between the features in testing the statistical significance of the mean differences, was implemented in IBM SPSS Statistics. By using this procedure, it was possible to contrast the null hypothesis in the features obtained.

To perform this statistical analysis, it was assumed that the features had a multivariable normal distribution and no assumptions were made regarding the homogeneity of the variance or the correlation between the features. A significance value of $p < 0.05$ was considered sufficient to assume the existence of feature differences between the four groups analyzed.

### *H. Experiments*

The participants in this study belonged to three different groups: the control group with 18 participants, labeled as C, the group with 14 ALS participants diagnosed with bulbar dysfunction, labeled B, and the group with 31 ALS participants not diagnosed with bulbar dysfunction, labeled NB. In addition, the A label was added to every ALS participant, with or without bulbar dysfunction.

Three experiments were performed with these groups:

1) Performance evaluation of the supervised models for 4 cases (C vs. B, C vs. NB, B vs. NB and C vs. A) by using the original corpus.
2) Re-labeling of the NB participants as B' and C' by applying the semi-supervised S4VM algorithm. Thus, the NB group was removed.
3) Re-evaluation of the model performance with four new groups of participants: C vs. B+ (B + B'), C vs. NB- (NB - B'), B+ vs. NB- and C+ (C + C') vs. B+.

The first experiment obtained the outcomes of the models using the original corpus. Next, due to the great uncertainty found in the ALS participants diagnosed without bulbar involvement [2, 6], it was intended to re-label the participants of the NB group as B and C using the semi-supervised S4VM model. We tried to obtain a new corpus that contains elements classified as bulbar (B') or control (C') by S4VM among those who were previously diagnosed as non-bulbar by a clinician (NB). In the third experiment, the models outcomes were again obtained by changing the composition of the B and NB groups by adding B' to B (B+) and removing C' from NB (NB-).

### *I. Performance Metrics*

There are several metrics for evaluating classification algorithms [27]. The Accuracy, Sensitivity and Specificity metrics, the most popular ones, were used to evaluate the performance of the classification models.

## III. Results

Firstly, the voice fingerprint representations and the features selected in relation to the four cases (C vs B, C vs NB, B vs NB and C vs A) are presented. Then, the performance of the classification models is evaluated.

### *A. Voice fingerprint representations to detect bulbar dysfunction in ALS*

The voice fingerprint for detecting bulbar dysfunction in ALS consisted of the computation of $p_N(T)$, the 5 PCs of $p_N(T)$ and the 5 ICs of $p_N(T)$ for the 5 Spanish vowels.

Fig. 6 shows the voice fingerprint computed of a ALS patient. Figure 6(a) shows the $p_N(T)$ of the five Spanish Vowels (a, e, i, o, u). Figure 6(b) shows the 5 PCs of $p_N(T)$ of the five Spanish Vowels. Figure 6(c) shows the 5 ICs of $p_N(T)$ of the five Spanish Vowels. Figure 6(d) shows the spectrum of $p_N(T)$ of the five Spanish Vowels. Figure 6(e) shows the spectrum of the 5 PCs of $p_N(T)$ of the five Spanish Vowels. Figure 6(f) shows the probability density function of the 5 ICs of $p_N(T)$ of the five Spanish Vowels.

TABLE I: Model performance for the first experiment.

| | | **C vs B** | **C vs NB** | **B vs NB** | **C vs A** |
|---|---|---|---|---|---|
| RF | Accuracy | **88.3** | 48.7 | 66.7 | 68.6 |
| | Sensitivity | 85.0 | 46.7 | 40.0 | **84.0** |
| | Specificity | **95.0** | 50.0 | 75.8 | 30.0 |
| LR | Accuracy | 56.7 | 62.0 | 60.2 | 65.0 |
| | Sensitivity | 45.0 | 68.3 | 60.0 | 66.0 |
| | Specificity | 65.0 | 50.0 | 59.2 | **65.0** |
| LDA | Accuracy | 54.2 | 62.0 | 54.0 | 65.2 |
| | Sensitivity | 45.0 | 68.3 | 60.0 | 68.5 |
| | Specificity | 60.0 | 50.0 | 51.7 | 60.0 |
| NN | Accuracy | 66.7 | 59.2 | 66.3 | 60.5 |
| | Sensitivity | 70.0 | 63.3 | 60.0 | 64.0 |
| | Specificity | 65.0 | 50.0 | 68.3 | 55.0 |
| SVM | Accuracy | 86.5 | **68.0** | **78.7** | **73.1** |
| | Sensitivity | **88.3** | **71.7** | **80.0** | 79.0 |
| | Specificity | 85.0 | **60.0** | **77.5** | 60.0 |

### *B. Features Selected*

A total of 75 features were obtained. The MANOVA analysis was applied to select the statistically significant features (p-value$<$0.05) for the four comparisons analyzed: C vs. B, C vs. NB, B vs. NB and C vs. A. The features not showing statistical significance (p-value$\geq$0.05) were discarded. The box plots of the statistically significant features are depicted in Figure 7.

In the case C vs B, a set of 19 statistically significant features (p-value$<$0.05) were obtained. These were medRR2, medRR3, medRR5, fmEsPat1, enBnEs_a3, enBnEs_e4, enBnEs_e5, enBnEs_i5, enBnEs_o5, entPat2, entPat3, entPat4, entPC1, entPC2, entPC4, entPC5, entIC2, entIC3 and entIC4.

In the case C vs NB, a set of 2 statistically significant features were obtained. These were enBnEs_o4 and enBnEs_o5.

In the case B vs NB, a set of 20 statistically significant features were obtained. These were medRR1, medRR2, medRR3, medRR4, enBnEs_e4, enBnEs_e5, enBnEs_i3, entPat1, entPat2, entPat4, entPC1, entPC2, entPC3, entPC4, entPC5, entIC1, entIC2, entIC3, entIC4 and entIC5.

In the case C vs A, a set of 3 statistically significant features were obtained. These were fmEsPat1, enBnEs_o4 and enBnEs_o5.

### *C. Classification Model Performance and Experiments*

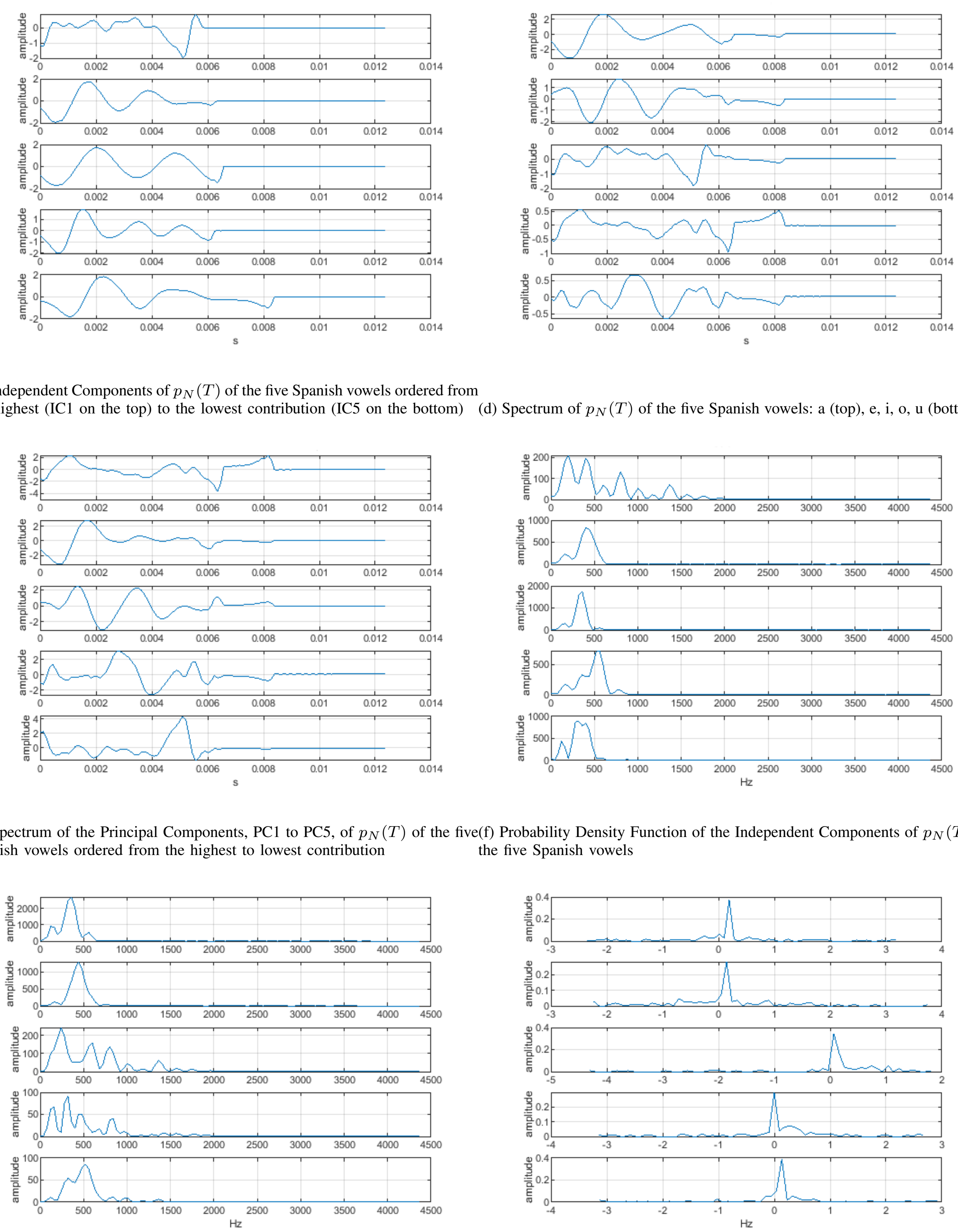
In the first experiment, the classification models were fitted with the 75 features selected. Table I shows the classification performance (*Accuracy*, *Sensitivity* and *Specificity* metrics) of the classification models tested for the four cases defined with the original labels (B, NB and C).

In the case C vs. B, the results indicate that RF and SVM have a good classification performance. RF obtained the best *Accuracy*, 88.3%, with a *Sensitivity* of 85.0% and a *Specificity* of 95.0%. SVM obtained an *Accuracy* of 86.5% with a *Sensitivity* of 88.3% and a *Specificity* of 85.0%. NN, LR and LDA showed a poorer performance, obtaining respective *Accuracies* of 66.7%, 56.7% and 54.2% respectively.

In the cases C vs. NB, B vs. NB and C vs. A, poorer results were obtained. In all these cases SVM obtained the best *Accuracy*, these being 68.0%, 78.7% and 73.1% respectively.

In the second experiment, S4VM was applied from the data labeled C and B to estimate the class of NBs which were split

Fig. 6: Voice fingerprint for a patient.



(a) $p_N(T)$ of the five Spanish vowels: a (top), e, i, o, u (bottom)

(b) Principal Components of $p_N(T)$ of the five Spanish vowels ordered from the highest (PC1 on the top) to the lowest contribution (PC5 on the bottom)

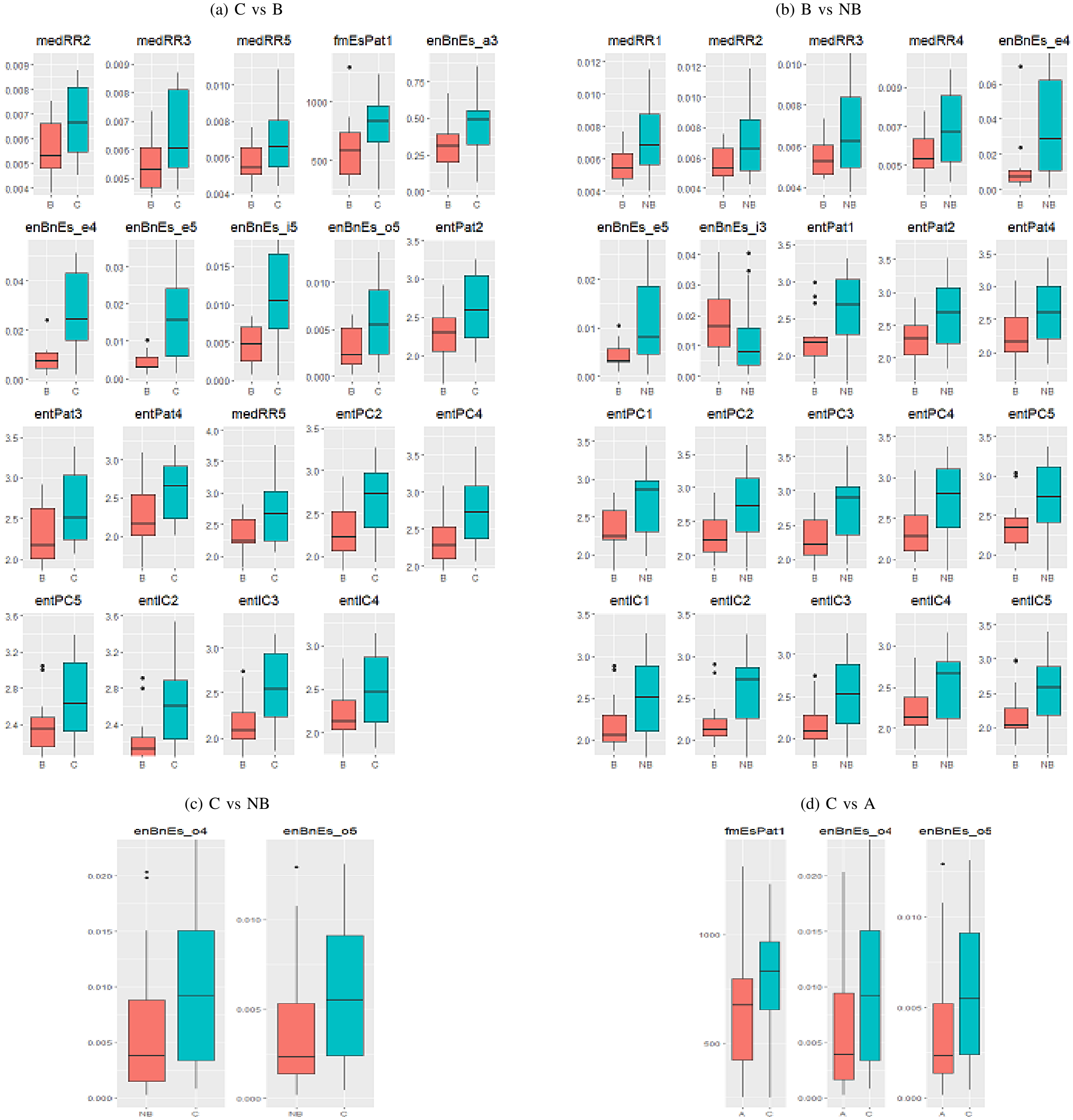
(c) Independent Components of $p_N(T)$ of the five Spanish vowels ordered from the highest (IC1 on the top) to the lowest contribution (IC5 on the bottom)

(d) Spectrum of $p_N(T)$ of the five Spanish vowels: a (top), e, i, o, u (bottom))

(e) Spectrum of the Principal Components, PC1 to PC5, of $p_N(T)$ of the five Spanish vowels ordered from the highest to lowest contribution

(f) Probability Density Function of the Independent Components of $p_N(T)$ of the five Spanish vowels

Fig. 7: Box plots of the statistically significant features per each case.



(a) C vs B

(b) B vs NB

(c) C vs NB

(d) C vs A

TABLE II: Model performance for the third experiment.

| | | C vs B+ | C vs NB- | B+ vs NB- | C+ vs B+ |
|---|---|---|---|---|---|
| RF | Accuracy | **93.5** | 69.3 | 89.7 | **92.4** |
| | Sensitivity | **96.6** | 66.7 | **83.3** | 83.3 |
| | Specificity | **90.0** | **70.0** | 96.7 | **97.5** |
| LR | Accuracy | 56.0 | 69.2 | 71.7 | 71.4 |
| | Sensitivity | 53.3 | 73.3 | 76.7 | 66.7 |
| | Specificity | 60.0 | 60.0 | 65.0 | 75.0 |
| LDA | Accuracy | 62.9 | 69.2 | 82.3 | 80.9 |
| | Sensitivity | 60.0 | 73.3 | 76.7 | 78.3 |
| | Specificity | 65.0 | 60.0 | 88.3 | 82.5 |
| NN | Accuracy | 75.9 | 68.7 | 86.9 | 82.6 |
| | Sensitivity | 80.0 | **80.0** | **83.3** | 83.3 |
| | Specificity | 70.0 | 50.0 | 91.6 | 82.5 |
| SVM | Accuracy | 89.2 | **69.6** | **91.0** | 92.1 |
| | Sensitivity | 90.1 | 73.3 | **83.3** | **86.7** |
| | Specificity | **90.0** | 60.0 | **100** | 95.0 |

into C' and B'. From the total of 31 NBs, 9 were split as B' and 22 as C'.

In the third experiment, the classification models were fitted with the features selected and tested for the four new cases (C vs. B+, C vs. NB-, B+ vs. NB- and C+ vs. B+). Table II shows the classification performance (*Accuracy*, *Sensitivity* and *Specificity* metrics) for this experiment.

In the case of C vs. B+, the results indicate that RF and SVM have good classification performance. RF obtained the best *Accuracy*, 93.5%, with a *Sensitivity* of 96.6% and a *Specificity* of 90.0%. SVM obtained an *Accuracy* of 89.2% with a *Sensitivity* of 90.1% and a *Specificity* of 90.0%. NN obtained an *Accuracy* of 75.9% with a *Sensitivity* of 80.0% and a *Specificity* of 70.0%. LR and LDA showed poorer performance, obtaining *Accuracies* of 56.0% and 62.9% respectively.

In the B+ vs. NB- and C+ vs. B+ cases, good model classification performance was also observed.

In B+ vs. NB-, SVM obtained the best *Accuracy*, 91.0%, with a *Sensitivity* of 83.3% and a *Specificity* of 100.0%. RF obtained an *Accuracy* of 89.7% with a *Sensitivity* of 83.3% and a *Specificity* of 96.7%. NN obtained an *Accuracy* of 86.9% with a *Sensitivity* of 83.3% and a *Specificity* of 91.6%. LDA obtained an *Accuracy* of 82.3% with a *Sensitivity* of 76.7% and a *Specificity* of 88.3%. Finally, LR performed the worst performance with an *Accuracy*, 71.7%, a *Sensitivity* of 76.7% and a *Specificity* of 65.0%.

In C+ vs. B+, RF obtained the best *Accuracy*, 92.4%, with a *Sensitivity* of 83.3% and a *Specificity* of 97.5%. SVM obtained an *Accuracy* of 92.1% with a *Sensitivity* of 86.7% and a *Specificity* of 95.0%. NN obtained an *Accuracy* of 82.6% with a *Sensitivity* of 83.3% and a *Specificity* of 82.5%. LDA obtained an *Accuracy* of 80.9% with a *Sensitivity* of 78.3% and a *Specificity* of 82.5%. Finally, LR had the worst performance with an *Accuracy* of 71.4%, a *Sensitivity* of 66.7% and a *Specificity* of 75.0%.

Finally, in C vs NB- poorer results were obtained. All models showed similar performance. SVM, RF, LR, LDA and NN obtained *Accuracies* of 69.6%, 69.3%, 69.2%, 69.2% and 68.7%.

# IV. Discussion

## A. Principal Findings

We have carried out a preliminary assessment of the potential for obtaining a voice fingerprint for an early detection of bulbar dysfunction in ALS patients. This was motivated by the need for a standardized diagnostic procedure for assessing bulbar dysfunction and new methodologies based on objective measurements [2].

The study demonstrated the feasibility of the methodology proposed. Its major benefit is to provide a methodology based on objective measures to identify bulbar dysfunction in early stages of the ALS disease. We suggest two new labels, C' and B', to improve the diagnosis of those patients in whom bulbar dysfunction has not yet been detected by the current subjective procedures.

This methodology is based on the development of a voice fingerprint which consists of a pattern generated from the quasi-periodic components of a steady portion of the five Spanish vowels, and five principal and independent components computed from this pattern. From this voice fingerprint, a total of 75 features were obtained. Then, a MANOVA analysis was applied to obtain the significant features for the four cases studied (C vs. B, C vs. NB, B vs NB and C vs. ALS). Finally, three experiments were conducted.

The first experiment showed the performance of the machine learning models used for the four cases. The best results were obtained when comparing C with B. RF and SVM achieved the best performance ($Accuracies$ of 88.3% and 86.5% respectively). LR, LDA and NN were far from these results. When comparing C with NB, poor performance was observed. SVM achieved the best accuracy (68.0%). In B vs. NB and C vs. A, SVM achieved $Accuracies$ of 78.7% and 73.1% respectively. From the good results achieved by C vs. B, it can be inferred that the methodology proposed is good at detecting bulbar dysfunction, RF and SVM being the best models for performing this task. The poor performance obtained in C vs. NB revealed a similar voice performance of Cs and NBs, as expected. Instead, the performance obtained in B vs. NB may indicate that some NBs voices could be affected in some NBs but this may yet not be perceptible to human hearing.

The second experiment revealed that 9 of the total of 31 NBs may have bulbar dysfunction. This result is consistent with the previous statement that indicated that the voices of some NBs could be affected. We suggest labelling these patients as B' if their voices show a similar performance to Bs and C' if they are similar to C.

The third experiment performed better than the first one when B' and C' labels were considered. In C vs B+, RF obtained an $Accuracy$ of 93.5% (increasing by 5% over the first experiment), with a $Sensitivity$ and $Specificity$ of 96.6% and 90.0%, outperforming the results obtained in the first experiment. Similarly, in B+ vs. NB-, the classification performance was greatly improved. SVM obtained an $Accuracy$ of 91.0% with a $Sensitivity$ and $Specificity$ of 93.3% and 100.0%. Good results were also obtained in C+ vs B+. RF obtained the best result with an $Accuracy$ of 92.4%, and in C vs. NB-, the models showed poor performance, SVM being the one that obtained the best $Accuracy = 69.6\%$.

In general, the third experiment achieved better performance than the first one. This indicates that our method can diagnose bulbar dysfunction better than clinicians with the current subjective approaches.

### B. Comparison with Prior Work

This study is consistent with Tena et al. [6], which found a great uncertainty in ALS patients in whom bulbar dysfunction was not detected yet suggesting that some of them were misdiagnosed. It is also consistent with Plowman et al. [2], which indicated the difficulties in diagnosing bulbar dysfunction by subjective approaches. In many cases, the perturbance in those subjects' voices could not be appreciated by the human ear until advanced stages of the disease. We went a step further by providing two new labels, B' and C', to achieve an earlier and more accurate diagnosis.

This study is also in line with Norel et al., An et al. and Wang et al. [3, 4, 5], who demonstrated that voice is one of the most important aspects for detecting bulbar dysfunction. Norel et al. implemented SVM classifiers to recognize the presence of voice disability in patients with ALS. They identified acoustic features in naturalistic contexts, achieving 79% accuracy (sensitivity 78%, specificity 76%) in the classification of males and 83% accuracy (sensitivity 86%, specificity 78%) in the classification of females. An et al. implemented CNNs to classify the intelligible speech produced by patients with ALS and healthy individuals. The experimental results indicated a sensitivity of 76.9% and a specificity of 92.3%. Wang et al. [5] implemented SVM and NN using acoustic features and adding articulatory motion information (from tongue and lips). When only acoustic data were used to fit the SVM, the overall accuracy was slightly higher than the level of chance (50%). Adding articulatory motion information further increased the accuracy to 80.9%. The results using NN were more promising, with accuracies of 91.7% being obtained using only acoustic features, increasing to 96.5% with the addition of both lip and tongue data. Adding motion measures increased the classifier accuracy significantly at the expense of including more invasive measurements to obtain the data. We investigated the means of optimizing accuracy in detecting ALS bulbar dysfunction by only analyzing the voices of patients. These studies only focused on B vs. C cases.

To date, only Tena et al. [6] have conducted studies considering additional cases. They used phonatory subsystem features, such as jitter, shimmer, harmonic-to-noise ratio and pitch and PCA, to analyze the performance of several machine learning models considering four scenarios (C vs. B, C vs. NB, B vs. NB and C vs. A). In C vs. B, they obtained an $Accuracy$ of 95.8% with a $Sensitivity$ of 91.4% and $Specificity$ of 99.3% using SVM. Poor performance was obtained in B vs. NB ($Accuracy$ of 75.5% with $Sensitivity$ of 55.7% and and $Specificity$ of 88.4% using RF). In C vs. NB and C vs. A, good results were also obtained, NN being the model which performanced best in both cases ($Accuracies$ of 92.5% and 92.2% respectively).

In this study, in C vs. B, an $Accuracy$ of 88.3% was obtained for RF with a $Sensitivity$ of 85.0% and a $Specificity$ of 95.0%. This performance improved when considering B' patients, C vs. B+, obtaining an $Accuracy$ of 93.5% outperforming the results of [3, 4, 5]. In B vs. NB, we obtained an $Accuracy$ of 78.7% (SVM) outperforming the results obtained by [6]. This performance was greatly improved when considering B' patients, B+ vs. NB-, obtaining an $Accuracy$ of 91.0% with a $Sensitivity$ of 83.3% and a $Specificity$ of 100.0%. This suggests that having well-annotated patients is essential for properly assessing bulbar dysfunction in B vs. NB. We demonstrated that semi-supervised classification models such as S4VM are an useful tools for performing this task.

### C. Limitations

The size and bias of this study is heavily influenced by the fact that ALS is a rare and a very heterogeneous disease where not all the patients present the same symptomatology. Although upsampling and semi-supervised classifier techniques were used to correct the bias, it would be necessary to increase the number of participants to draw definitive conclusions.

However, we proved that the method presented can be successfully applied to such a corpus. The question is what the outcomes should be when applying it to a large enough corpus. The outcomes indicate that accuracy could increase much more. A specific study should be performed to determine the extent of this increase.

## V. Conclusions

Promising outcomes in detecting bulbar dysfunction were obtained when comparing ALS patients with and without this dysfunction in early stages of the disease, or prior to being diagnosed by clinicians. This could lead to the development of a screening tool that may help to develop standardized diagnostic procedures for assessing bulbar dysfunction based on objective measures. This directly addresses a recent statement released by the NEALS bulbar subcommittee regarding the need for objective-based approaches [2].

Due to the great uncertainty of the corpus, we highlight the importance of improving the annotation of ALS patients as regards bulbar dysfunction to develop powerful machine learning models able to distinguish this dysfunction. We provide two new labels, C' and B', and demonstrate that Semi-Supervised Machine learning models could help in the early detection of this dysfunction. Yet, further analyses are needed to develop this concept fully. These include performing longitudinal studies in which patients' diagnosis are retrieved at several follow-ups.

The usefulness of this methodology is that it could be applied to the automated identification and early diagnosis of many other neurological or respiratory illnesses where obtaining a large enough and well-annotated corpus is difficult.

## Acknowledgements

This work was supported by the Ministerio de Economía y Competitividad (TIN2017-84553-C2-2-R) and the Ministerio de Ciencia e Innovacion (PID2020-113614RB-C22). AT is a member of CIMNE, a Severo Ochoa Centre of Excellence (2019-2023) under grant CEX2018-000797-S, funded by MCIN/AEI/10.13039/501100011033. The Neurology Department of the Bellvitge University Hospital in Barcelona permitted the recording of the voices of the participants in its facilities. The clinical records were provided by Carlos Augusto Salazar Talavera. Dr. Marta Fulla and Maria Carmen Majos Bellmunt contributed advice about the process of eliciting the sounds.

## References

[1] Cristina et al. Carmona-Duarte. Study of several parameters for the detection of amyotrophic lateral sclerosis from articulatory movement. *Loquens*, 4, 2017.

[2] Emily K Plowman, Lauren C Tabor, James Wymer, and Gary Pattee. The evaluation of bulbar dysfunction in amyotrophic lateral sclerosis: survey of clinical practice patterns in the United States. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 18(5-6):351–357, 2017.

[3] Raquel Norel, Mary Pietrowicz, Carla Agurto, Shay Rishoni, and Guillermo Cecchi. Detection of Amyotrophic Lateral Sclerosis (ALS) via Acoustic Analysis. *bioRxiv*, 2018.

[4] An Et al. Automatic Early Detection of Amyotrophic Lateral Sclerosis from Intelligible Speech Using Convolutional Neural Networks. *Interspeech*, pages 1913–1917, 2018.

[5] Jun Wang Et al. Towards Automatic Detection of Amyotrophic Lateral Sclerosis from Speech Acoustic and Articulatory Samples. In *INTERSPEECH*, 2016.

[6] Alberto Tena, Francec Claria, Francesc Solsona, Einar Meister, and Monica Povedano. Detection of Bulbar Involvement in Patients With Amyotrophic Lateral Sclerosis by Machine Learning Voice Analysis: Diagnostic Decision Support Development Study. *JMIR Med Inform*, 9(3):e21331, 2021.

[7] Behnaz Yousefi, Jaemin Shin, Eric H Schumacher, and Shella D Keilholz. Quasi-periodic patterns of intrinsic brain activity in individuals and their relationship to global signal. *NeuroImage*, 167:297–308, 2018.

[8] Dany Obeid, Sawsan Sadek, Gheorghe Zaharia, and Ghais El Zein. Touch-less heartbeat detection and cardiopulmonary modeling. In *2009 2nd International Symposium on Applied Sciences in Biomedical and Communication Technologies*, pages 1–5, 2009.

[9] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. Springer, 2013.

[10] Kayhan N Batmanghelich, Dong H Ye, Kilian M Pohl, Ben Taskar, Christos Davatzikos, and ADNI. Disease classification and prediction via semi-supervised dimensionality reduction. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1086–1090, 2011.

[11] Ehsan Adeli, Kim-Han Thung, Le An, Guorong Wu, Feng Shi, Tao Wang, and Dinggang Shen. Semi-Supervised Discriminative Classification Robust to Sample-Outliers and Feature-Noises. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):515–522, 2019.

[12] Alberto Tena. ALS Models and Data repository. https://github.com/atenad/ALS. Date accessed: November 9, 2021.

[13] Gary L Pattee and Others. Provisional best practices guidelines for the evaluation of bulbar dysfunction in amyotrophic lateral sclerosis. *Muscle & Nerve*, 59(5):531–536, 2019.

[14] Paul Boersma and David Weenink. Praat: doing phonetics by computer [Computer program] version 6.1.01. Technical report, of the University of Amsterdam, 2019.

[15] MATLAB and Signal Processing Toolbox Release. *The MathWorks, Inc., Natick, Massachusetts*. United States, 2012.

[16] T W Parks and C S Burrus. Digital Filter Design, John Wiley & Sons. *chapter 7, section 7. 3*, 3.(3.), 1987.

[17] Sanjit K Mitra and Digital Signal Processing. *2nd ed.* McGraw-Hill, 2001.

[18] F Gustafsson. “Determining the initial states in forward-backward filtering. ” *IEEE Transactions on Signal Processing*, 44:988–992, 1996.

[19] S L Marple. “Computing the Discrete-Time Analytic Signal via FFT. ” *IEEE® Transactions on Signal Processing*, 47:2600–2603, 1999.

[20] Irene Rodriguez-Lujan, Gonzalo Bailador, Carmen Sánchez Ávila, Ana Herrero, and Guillermo Vidal. Analysis of pattern recognition and dimensionality reduction techniques for odor biometrics. *Knowledge-Based Systems*, 52, 2013.

[21] Scott Makeig, Anthony J Bell, Tzyy-Ping Jung, and Terrence J Sejnowski. *Independent Component Analysis of Electroencephalographic Data*. Advances in Neural Information Processing Systems 8, 1996.

[22] I T Principal Component Analysis Jolliffe. *2nd ed., Springer*. 2002.

[23] MATLAB. *version 9.9.0.1495850 (R2020b Update 1)*. The MathWorks Inc., Natick, Massachusetts, 2020.

[24] Max Kuhn. The caret Package, 2009.

[25] J H Krijthe. RSSL: R package for Semi-supervised Learning. In B Kerautret, M Colom, and P Monasse, editors, *Reproducible Research in Pattern Recognition*, pages 104–115. 2016.

[26] Yu-Feng Li and Zhi-Hua Zhou. *Towards Making Unlabeled Data Never Hurt*. In: Proceedings of the 28th International Conference on Machine Learning (ICML’11), Bellevue, Washington, 2011.

[27] Alaa Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 2018.

# Chapter 3

# Discussion

This work demonstrated the feasibility of using bio-sounds for the automatic detection of certain diseases such as bulbar involvement in ALS patients or COVID-19 positive cases through the acoustic analysis of voices and coughs respectively. Four studies have been conducted. Three of them were related to detect bulbar involvement in ALS patients and one of them was related to detect COVID-19 cough.

The participants in the ALS studies belonged to three different groups: the control group (C), patients with ALS with bulbar involvement (B), and patients with ALS without bulbar involvement (NB). In addition, the ALS group (A) was composed of every participant with ALS, with or without bulbar involvement. Then, four classification experiments were performed. These consisted of C vs. B, C vs. NB, B vs. NB and C vs. ALS. Table 3.1 summarizes nomenclature employed for each group considered in the ALS studies.

Table 3.1: Groups of subjects considered in the ALS studies.

| **Group** | **Subjects** |
|---|---|
| C | Control subjects |
| B | ALS patients with bulbar involvement |
| NB | ALS patients without bulbar involvement |
| ALS | ALS patients with and without bulbar involvement |

In the COVID-19 cough study, five groups of subjects were defined for analysis. These were: the group of subjects tested COVID-19 positive (C), the group of subjects tested COVID-19 negative who had no cough as symptom (N), the group of subjects tested COVID-19 negative with non-specific-cough as a symptom (NC), the group of non-COVID-19 subjects with pertussis cough (PT), and finally, the group NNC, which merged all non-COVID-19 subjects (N, NC and PT). Four classification experiments were performed. These consisted of C vs. N, C vs. NC, C vs. PT and C vs. NNC. Table 3.2 summarizes nomenclature employed for each group considered in the COVID-19 cough study.

Table 3.2: Groups of subjects considered in the COVID-19 cough study.

| Group | Subjects |
|---|---|
| C | COVID-19 tested positive subjects |
| N | COVID-19 tested negative subjects<br>without cough as symptom |
| NC | COVID-19 tested negative subjects<br>with with non-specific-cough as a symptom |
| PT | non-COVID-19 subjects with pertussis cough |
| NCC | N, NC and PT subjects |

## 3.1 First study - ALS bulbar involvement

The first study was guided by 2 objectives:

1. to design a methodology for diagnosing bulbar involvement efficiently through the acoustic parameters of uttered vowels in Spanish, and
2. to demonstrate the better performance of automated diagnosis of bulbar involvement compared with human diagnosis.

This study was based on the accurate acoustic analysis of the five Spanish vowel segments, that were elicited by all the participants. 15 acoustic features were extracted. These were jitter(absolute), jitter(relative), jitter(rap), jitter(ppq5), shimmer(relative), shimmer(dB), shimmer(apq3), shimmer(apq5), shimmer(apq11), pitch(mean), pitch(SD), pitch(min), pitch(max), HNR(mean) and HNR(SD). Then, the PCs of these features were obtained to fit the most common supervised classification models in clinical diagnosis, SVM, NN, LDA, LR, NaB and RF. Finally, their performance was compared.

This study demonstrated the feasibility of automatic detection of bulbar involvement in ALS patients through acoustic features obtained from vowel utterance. The study also confirmed that speech impairment is one of the most important issues for diagnosing bulbar involvement as was suggested in [83]. Furthermore, bulbar involvement can be detected using automatic tools before it becomes perceivable to human hearing.

Voice features extracted from B compared with those features extracted from C showed the best classification models performance to determine bulbar involvement in ALS patients.

*Accuracy* for C vs. B revealed values of 95.8% for SVM with the classification threshold established at 50%. However, on switching it to 95%, the *Accuracy* values for SVM dropped (86.3%) and LR showed the best performance (92.8%). NN also showed a good *Accuracy* (92.6%). This implied that NN and LR were more robust for finding *Accuracy*.

For that case, the results obtained reinforced the idea that it is possible to diagnose bulbar affection of ALS patients using supervised models and objective measures. The SVM and LR models provided the best performance for the 50% and 95% thresholds respectively.

A great uncertainty was found in the analysis regarding bulbar involvement in the NB patients.

The *Accuracy* values of C vs. NB and C vs. ALS with the classification threshold at 50%, gave 92.57% and 92.2% for NN. That revealed that the features extracted from NB differed significantly from C. Lower performance was expected because participants labeled as C and NB should have had similar voice performance. This indicated that some of the NB participants probably had bulbar involvement, but they were not correctly diagnosed because the perturbance in their voices could not be appreciated by the human ear. It could alternatively be simply that a classification threshold of 50% could be very optimistic. For a 95% classification threshold, lower results were obtained in C vs. NB and C vs. ALS. NN showed the best performance with *Accuracy* = 84.8% and *Accuracy* = 86.8% respectively for the two cases. The performance between the B and C showed better results than between NB and the C. Despite this, the unexpectedly high performance of the models for C vs. NB still suggested that some NB participants could have bulbar involvement. Changing the classification threshold to 95% worsened the results, specially for *Sensitivity* although it still remained significant.

B vs. NB revealed that the classification models did not distinguish B and NB participants as well as they did with the other groups. The *Accuracy* in the 50% threshold showed the highest performance of 75.5% for RF while the models showed difficulties in identifying positive cases. That may be due to the low difference in the variation of the data among B and NB participants. The same occurred for the 95% threshold. LR obtained the highest *Accuracy* of 74.1% but obtained a *Sensitivity* of 16.7%. These values remained far from those in C vs. B. These results also reinforced the idea that NB participants were misdiagnosed.

The good model performance obtained in comparing C vs. NB supported these findings underscoring the importance of using objective measures for assessing bulbar involvement.

The projection of the NB in the PCA biplot chart required special attention. Although the projection of these subjects had a spatial proximity with respect to the C, their variability was higher, overflowing the circle corresponding to the B. This indicated that some features of some NB patients, especially shimmer and jitter, had similar projections to the B patients. This may reveal that these NB patients could have bulbar involvement but were not correctly diagnosed yet because the perturbance in their voices could still not be appreciated by human hearing. The box plots of the features also indicated that the means of the features of the NB patients were between the means of the features of C and B corroborating these assumptions.

The PCA biplot charts indicated that shimmer and jitter were the most important features for group separation in the 2-PC model for ALS classification, while they also revealed pitch and HNR parameters as good variables for this purpose. These results were consistent with Vashkevich et al. [16], who demonstrated significant differences in jitter and shimmer in ALS patients. They were also consistent with Mekyska [30] and Teixeria et al. [29] who mentioned pitch, jitter, shimmer and HNR values as the most popular features describing pathological voices. Finally, Silbergleit [26] suggested the shimmer, jitter and HNR parameters were sensitive indicators of early laryngeal deterioration in ALS.

Concerning the classification models, recently, Norel et al. [18] implemented SVM classifiers to recognize the presence of speech impairment in ALS patients. They identified acoustic speech features in naturalistic contexts, achieving 79% *accuracy* (*precision*=0.78, *recall*=0.76) for male classification and 83% *accuracy* (*precision*=0.86, *recall*=0.78) for females. The data used did not originate from a clinical trial or contrived study nor was it collected under laboratory conditions. Wang [17] implemented SVM and NN using acoustic features and adding articulatory motion information (from tongue and lips). When only acoustic data were used to fit the SVM, the overall *accuracy* was slightly above the level of chance (50%). Adding articulatory motion information further increased the *accuracy* to 80.9%. The results using NN were more promising, with accuracies of 91.7% being obtained using only acoustic features and these accuracies increasing to 96.5% with the addition of both lip and tongue data. Adding motion measures increased the classifier *accuracy* significantly at the expense of including more invasive measurements to obtain the data. We investigated the means of optimizing *accuracy* in detecting ALS bulbar involvement by only analyzing the voices of patients. An et al. [19] implemented CNNs to classify the intelligible speech produced by patients with ALS and healthy individuals. The experimental results indicated a sensitivity of 76.9% and a specificity of 92.3%. Vashkevich et al. [16] performed LDA with accuracy of 90.7% and Susha et al. [13], used DNNs based on MFCCs with a 92.2% accuracy for automatic detection of patients with ALS.

Starting from the most widely used features suggested in the literature, the classification models used in this paper to detect bulbar involvement automatically (C vs. B) performed better than the ones used by other authors. We obtained the best ever performance metrics. This suggested that decomposing the original dataset of features into PCs to obtain another dataset whose data (PCs) were linearly independent and therefore uncorrelated improves the performance of the models.

## 3.2 Second study - ALS bulbar involvement

The second study demonstrated that it is possible to diagnose bulbar involvement by using supervised gender-specific models fitted to the significant phonatory and time-frequency features.

In the case of B vs C, the *Accuracy* achieved was up to 98.1% (RF) and 96.1% (RF) for females and males respectively.

Lower performance was obtained in C vs NB but this was still higher than expected. The voice performance in C or NB should be similar. This indicated that some participants in the NB group were probably incorrectly diagnosed. This was coherent with [59]. Similarly, the excellent performance achieved in C vs. A suggested that some of the members of A (14 out of 45) had bulbar involvement.

On the whole, huge uncertainty was observed in the evaluation concerning bulbar involvement among the participants in the NB group. The case of B vs NB disclosed that the models did not differentiate between the B and NB subject groups as well as they did with the other groups. RF achieved the best overall performance (Accuracy = 91.8%) in males. However, the model presented problems for spotting positive cases (Sensitivity = 55.0%). In females, RF achieved an *Accuracy* of 84.8%. These values were still far from the ones obtained in the C vs B case. These outcomes additionally reinforced the idea that NB subjects were misdiagnosed.

The outcomes of each comparison between groups depended on the significant features chosen (between phonatory and time-frequency). In other words, the optimal results in each experiment were obtained with an ad-hoc set of features. This means the differentiation between the participants in different groups depended on different features. However, classifiers obtained very similar results for each experiment, showing a lesser influence.

The results obtained proved that combining phonatory subsystem and time-frequency features improved the ability of the machine-learning models to detect bulbar involvement. In addition, detecting bulbar involvement also depended on the ad-hoc set of significant features found for such a case.

This study was consistent with [16, 17, 26, 59] which demonstrated that such phonatory subsystem features as jitter, shimmer, pitch and HNR were sensitive indicators for describing pathological voices in ALS. It was also consistent with [59] where great uncertainty was found in the diagnosis of NBs participants.

Besides the 15 phonatory subsystem features obtained in [59], this study also provided 35 time-frequency features. The combination of phonatory subsystem and time-frequency features, after performing MANOVA for feature selection, enhanced the outcomes of [59], which achieved the best results to date for detecting bulbar involvement in ALS using only acoustic features, ahead of [13, 17, 18].

*Accuracies* of up to 98.1% (RF) and 96.1% (RF) for females and males respectively were achieved when comparing the bulbar and control participants (case B vs C). This *Accuracy* exceeded the one obtained in [18] with SVM (79.0%) by 17.1% for males and 15.1% for females. The other studies found did not distinguish the classification problems by gender. In [59], SVM obtained an Accuracy of 95.8%. In [13], NN based on Mel Frequency Cepstral Coefficients (coefficients for speech representation based on human auditory perception) obtained 90.7%. In [17], NN based on phonatory subsystem features obtained 91.7% and adding motion sensors for both lip and tongue data increased the Accuracy to 96.5% at the expense of including more invasive measurements. For females, our results outperformed those from the aforementioned studies by 2.3%, 7.4% and 6.4% respectively. For males, ours were 0.3% above those obtained in [59] and 5.4% and 4.4% above those obtained in [13, 17].

When comparing ALS patients diagnosed with bulbar involvement with those patients in whom bulbar involvement has yet to be detected (B vs NB), the outcomes outperformed the ones obtained in [59]. The respective accuracy for males and females increased by 16.3% and 9.3% with the same classifier (RF) (91.8% and 84.8% as against 75.5%). This was an important outcome which indicated that the use of time-frequency features increased the identification of bulbar involvement among patients with ALS.

The outcomes obtained in the C vs NB and C vs A cases were very similar to those in [59], reinforcing the idea that some NBs could have bulbar involvement.

The most important gains were obtained when comparing B and NB. The selection of the significant features for this comparison improved the outcomes. Thus, involvements (i.e. bulbar) could be detected through a separate, and more closely adjusted, set of features. Consequently, by increasing the identification of particular features, treatment could be better customized for each ALS patient.

In addition, only studies showing C vs. B have been presented in the literature (except in [59]). No attempts to distinguish other subjects have been made to date. We highlighted this differentiating issue, and the importance of future research into it.

## 3.3 Third study - ALS bulbar involvement

In the third study, we carried out a preliminary assessment of the potential of obtaining a voice fingerprint for an early detection of bulbar dysfunction in ALS patients. This was motivated by the need of standardised diagnostic procedure for assessing bulbar dysfunction and new methodologies based on objective measures [2].

The study demonstrated the feasibility of the methodology proposed. Its major benefit was to provide a methodology based on objective measures to identify bulbar dysfunction in early stages of the ALS disease. We suggested two new labels, C' and B', to improve the diagnosis of those patients in whom bulbar dysfunction had not been detected yet by the current subjective procedures.

This methodology was based on the development of a voice fingerprint which consisted of a pattern generated from the quasi-periodic components of a steady portion of the five Spanish vowels, and five principal and independent components computed from this pattern. From this voice fingerprint, a total of 75 features were obtained. Then, a MANOVA analysis was applied to obtain the significant features for the four cases studied (C vs. B, C vs. NB, B vs NB and C vs. ALS). Finally, three experiments were conducted.

The first experiment showed the performance of the machine learning models used for the four cases. The best results were obtained when comparing C vs. B. RF and SVM achieved the best performance (*Accuracy* of 88.3% and 86.5% respectively). LR, LDA and NN were far from these results. When comparing C vs. NB a poor performance was observed. SVM achieved the best accuracy (68.0%). In B vs. NB and C vs. A, SVM achieved an *Accuracy* of 78.7% and 73.1% respectively. From the good results achieved by C vs. B, we inferred that the methodology proposed was good to detect bulbar dysfunction, being RF and SVM the best models to perform this task. The poor performance obtained in C vs. NB revealed a similar voice performance of Cs and NBs as it was expected. Instead, the performance obtained in B vs. NB indicated that some NBs voices could be affected in some NBs buy may it was not perceptible by the human hearing yet.

The second experiment revealed that 9 from the total of 31 NBs probably had bulbar dysfunction. This result was consistent with the previous statement that indicated that some NBs could have their voices affected. We suggested to label these patients as B' if their voices showed a similar performance than Bs and C' if they were similar to C.

The third experiment showed a better performance than the first one when B' and C' labels were considered. In C vs B+, RF obtained an *Accuracy* of 93.5% with a *Sensitivity* and *Specificity* of 96.6% and 90.0% respectively outperforming the results obtained in the first experiment. Similarly, in B+ vs. NB-, the classification performance was greatly improved. SVM obtained an *Accuracy* of 91.0% with a *Sensitivity* and *Specificity* of 93.3% and 100.0% respectively. In C+ vs B+ good results were also obtained, RF obtained the best result with an *Accuracy* of 92.4%, and in C vs. NB- the models showed a poor performance being SVM which obtained the best $Accuracy = 69.6\%$.

This study was consistent with the previous ones which found a great uncertainty in ALS patients in whom bulbar dysfunction was not detected yet suggesting that some of them were misdiagnosed. It was also consistent with Plowman et al. [2] which indicated the difficulties in diagnosing bulbar dysfunction by subjective approaches. In many cases, the perturbance in those subjects' voices could not be appreciated by the human ear until advanced stages of the disease. We went a step further providing two new labels, B' and C', to achieve an earlier and more accurate diagnosis.

In C vs. B, we obtained for RF an *Accuracy* of 88.3% with a *Sensitivity* and $Specificity$ of 85.0% and 95.0% respectively. This performance improved when considering B' patients, C vs. B+, obtaining an *Accuracy* of 93.5% outperforming the results of [18, 19, 17]. In B vs. NB, we obtained an *Accuracy* of 78.7% (SVM) outperforming the results obtained by our first study [59]. This performance was greatly improved when considering B' patients, B+ vs. NB-, obtaining an *Accuracy* of 91.0% with *Sensitivity* and $Specificity$ of 83.3% and 100.0% respectively. This suggested that to have well annotated patients is essential to properly asses bulbar dysfunction in B vs. NB. We demonstrated that semi-supervised classification models such as S4VM are useful tools to perform this task.

## 3.4 Fourth study - COVID-19 cough

The four study directly addressed a recent statement released by the WHO [3] which believed in the use of rapid tests essential to control people infected with COVID-19. We demonstrated the feasibility of automatic detection of COVID-19 positives from the time-frequency analysis of coughs.

The visual appraisal of the time-frequency representations confirmed differences in the frequency distribution of the voluntary coughs of the C, N, NC and PT subjects.

The features selected by RFE to fit the models obtained better results on the overall performance of the models than those features extracted by means of the Autoencoder. Furthermore, the rank of the features selected by RFE which fitted the model that obtained the best performance depended highly on the experiment done. This means that when comparing coughs, a good selection of the features must be chosen.

The classification models performed better when comparing C vs. PT than when comparing C vs. N, C vs. NC or C vs. NNC, although a good performance was observed for all the experiments. In C vs. PT, the metrics that performed better were $Accuracy = 94.81\%$, $Sensitivity = 98.91\%$ for RF, $Precision = 97.13\%$ for LR, $F - score = 97\%$ for RF and $AUC = 97.29$ for SVM. This experiment better detected positive COVID-19 coughs but did not work so well for classifying pertussis coughs ($Specificity = 85\%$ for LR and LDA). Instead, in the other experiments, the detection of positive and negative cases was more balanced. This was specially so in the C vs. NNC experiment, which obtained the best $Specificity = 85.09$. This experiment reflected a more real case scenario where COVID-19 coughs co-exist with coughs of different patterns. In the four classification experiments done, RF showed the best overall performance.

Other existing works, such as Laguarta et al. [47], extracted MFCCs from cough recordings and input them into a pre-trained CNN. Their model achieved an AUC of 97% with a $Sensitivity = 98.5\%$ and a $Specificity$ of 94.2%. Pahar et al. [51] presented a machine-learning based COVID-19 cough classifier able to discriminate COVID-19 positive coughs from both COVID-19 negative and healthy coughs recorded on a smartphone. They obtained an AUC of 98% using the Resnet50 classifier to discriminate between COVID-19 positive and healthy coughs, while an LSTM classifier was best able to discriminate between COVID-19 positive and COVID-19 negative coughs with an AUC of 94%. Brown et al. [52] used coughs and breathing to understand how discernible COVID-19 sounds were from those in asthma or healthy controls. Their results showed that a simple binary machine-learning classifier was able to classify healthy and COVID-19 sounds correctly. Their models achieved an AUC of above 80% across all tasks.

The RF model used in thi study performed similarly to the ones used by other authors (Accuracy and AUC close to, or above 90% depending on the experiment) although automated cough detection introduced some performance penalty. Additionally, our methodology allowed coughs in samples of raw audio recordings to be detected automatically by using the YAMNet deep neuronal network [63]. We also found the set of time-frequency features that could lead to distinguishing COVID-19 coughs from other cough patterns. In addition, the high performance obtained in various sampling sources (UdL, UC, Virufy and Coswara) validated our method as a more generic proposal.

Newer machine-learning works have shown lower results. For example, an accuracy of 85.2% with RF and 70.6% with CNN, were obtained in [53] and [54] respectively. Recently [55], an accuracy of 90% was obtained with a recurrent neural network (RNN) by using the Coswara dataset. However, the accuracy dropped to 80% with Coswara and Virufy simultaneously. This fact demonstrates that obtaining good outcomes when different datasets are used is a challenge. Our proposal behaved much better even when three additional datasets (UdL, UC and Pertussis) were used.

## 3.5 Limitations

This work has some limitations. In the ALS studies, using machine learning on small sample sizes made it difficult to evaluate the significance of the findings fully. The sample size of the dataset used was, in part, determined by the low prevalence of ALS, which is considered a rare disease. The small number of samples of the B group was influenced by the heterogeneity of the ALS disease in which patients' symptomatology is very diverse. Although, upsampling techniques have been made to correct the bias it would be necessary to increase the number of participants to draw definitive conclusions.

Furthermore, hand-edit of the segments of the voice recordings could have introduced subtle and unintended selection biases. Although automatic instruments have been created, these methods they are currently not accurate enough and require manual correction.

In the COVID-19 study, although in general, high performance was obtained in RF, its Specificity was not the optimal. Overall, Specificity outcomes were lower. That means that correctly classifying negative samples is an issue. This must be due to classification mistakes in the dataset. Additional efforts must be made to curate the corpus.

Furthermore, further analyses comparing COVID-19 cough patterns with cough patterns from other conditions, such as asthma or bronchitis, are needed.

# Chapter 4

# Conclusions

This work suggests that machine learning may be an appropriate tool for the automatic detection of certain diseases by means of bio-sounds analysis. Concretely, the methods presented can help with the diagnosis of ALS of the clinical multi-disciplinary teams, in particular, to help with the diagnosis of bulbar involvement. Additionally, they could also be useful to help for an early response to further COVID-19 outbreaks or other pandemics that may arise in the future.

We demonstrated that an accurate analysis of the features extracted from an acoustic analysis of the vowels elicited from ALS patients may be used for an early detection of the bulbar involvement. This could be done automatically using supervised classification models. Better performance was achieved by applying PCA previously to the obtained features. Note that, when classifying ALS participants with bulbar involvement and controls, the SVM for a 50% classification threshold exceeded the performance obtained by other authors, and more specifically, those obtained by [17] and [13].

Furthermore, bulbar involvement can be detected using automatic tools before it becomes perceivable to human hearing. The results point to the importance of obtaining objective measures to allow an early and more accurate diagnosis, given that humans may often misdiagnose this deficiency. This directly addresses a recent statement released by the NEALS bulbar subcommittee regarding the need for objective-based approaches [2].

Moreover, we demonstrated the usefulness of properly assessing bulbar involvement by using phonatory subsystem and time-frequency features from a study of the Spanish vowels outperforming previous works, specifically [17, 18, 13, 59]. We also demonstrated that each identification between two groups depends on the significant features found for such an experiment.

One of the main contribution is the idea of differentiating the diagnose by gender. This outperformed all the results of the literature.

The creation of a voice fingerprint combined with machine learning models could lead to minimize the uncertainty found in the diagnosis of bulbar involvement which use the current subjective procedures.

Due to this great uncertainty, we highlight the importance of improving the annotation of ALS patients on regards the bulbar involvement to develop powerful machine learning models able to distinguish this dysfunction. We demonstrated that semi-supervised Machine learning models could help in the early detection of this dysfunction. Further analysis is needed to fully develop this concept. This would include longitudinal studies in which the diagnosis of patients was recovered in several follow-ups.

We obtained promising results in detecting bulbar involvement when comparing ALS patients with and without this dysfunction. SVM obtained up to 91.0% of $Accuracy$ with 100.0% of $Specificity$. These could lead to the development of a screening tool that may help to develop standardised diagnostic procedures for assessing bulbar dysfunction based on objective measures.

These studies directly address a recent statement released by the NEALS bulbar subcommittee regarding the need for objective-based approaches [2]. The results obtained reinforces the idea that machine learning may be an appropriate screening tool for helping with the diagnosis of ALS with bulbar involvement.

We also demonstrated the feasibility of the automatic detection of COVID-19 from coughs. Excellent results were achieved by fitting an RF model with the set of the time-frequency features selected by RFE for distinguishing COVID-19 coughs. This new methodology presented could lead to automatic identification of COVID-19 by using existing simple and portable devices. It could be the core of a pre-screening mobile app for use as an early response to further COVID-19 outbreaks or other pandemics that may arise in the future.

The usefulness of these methods could be applied to the automated identification and early diagnosis of many other neurological or pulmonary problems, or infectious respiratory diseases such as Parkinson, asthma, bronchitis or Chronic Obstructive Pulmonary Disease.

Future work is directed towards the improvement of the ALS voice database by increasing the sample size. Also, longitudinal studies in which patients diagnosis were retrieved at several follow-ups are envisaged.

Also, we will gather more COVID-19 quality data, especially different cough patterns from other conditions, and curate the actual corpus to further train, fine-tune, and improving performance of the models.

# Abbreviations

| | |
|---|---|
| ALS | Amyotrophic lateral sclerosis |
| COVID-19 | COronaVIrus Disease of 2019 |
| NEALS | Northeast Amyotrophic Lateral Sclerosis Consortium |
| SARS-CoV2 | Severe Acute Respiratory Syndrome |
| WHO | World Health Organisation |
| ALSFRS-R | Amyotrophic lateral sclerosis Functional Rating Scale – Revised |
| F1 | First formant |
| F2 | Second formant |
| VISC | Vowel inherent spectral change |
| ENT | Ears, nose and throat |
| MDVP | Multi-Dimensional Voice Program |
| PHON | Sustained phoneme production |
| DDK | Diadochokinetic task |
| SPON | Spontaneous speech |
| SVM | Support Vector Machines |
| DNN | Deep neural network |
| MFCCs | Mel frequency cepstral coefficients |
| AUC | Area under the curve |
| LDA | Linear discriminant analysis |
| NDI | Northern Digital Inc. |
| CNN | Convolutional Neuronal Network |
| GRBAS | Grade, roughness, breathiness, asthenia, strain scale |
| MPT | Maximum phonation time |
| HNR | Harmonics-to-noise ratio |
| TFR | Time-frequency representation |
| NN | Neuronal Network |
| PNN | Probabilistic neural network |
| HMM | Hidden Markov Model |

| | |
|---|---|
| GMCC | Gammatone Cepstral Coefficient |
| FFT | Fast Fourier Transform |
| kNN | k-Nearest Neighbor |
| LR | Logistic Regression |
| RF | Random Forest |
| SMOTE | Synthetic minority oversampling technique |
| MLP | Multilayer perceptron |
| LSTM | long short-term memory |
| Resnet50 | Residual-based neural network architecture |
| SFS | Sequential forward selection |
| GBT | Gradient Boosting Trees |
| CWD | Choi-Williams distribution |
| F0 | Fundamental Frequency |
| WD | Wigner distribution |
| DFT | Discrete Fourier transform |
| MANOVA | Multivariate analysis of variance |
| RFE | Recursive Feature Elimination |
| PCA | Principal Component Analysis |
| PC | Principal Component |
| SVD | Singular Value Decomposition |
| ICA | Independent Component Analysis |
| NaB | Naive Bayes |
| S4VM | Safe Semi-Supervised Support Vector Machine |
| CEIm | Research Ethics Committee for Biomedical Research Projects |

# Bibliography

[1] Cristina et al. Carmona-Duarte. Study of several parameters for the detection of amyotrophic lateral sclerosis from articulatory movement. *Loquens*, 4, 2017.

[2] Emily K Plowman, Lauren C Tabor, James Wymer, and Gary Pattee. The evaluation of bulbar dysfunction in amyotrophic lateral sclerosis: survey of clinical practice patterns in the United States. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 18(5-6):351–357, 2017.

[3] WHO. WHO coronavirus disease (COVID-19) dashboard. https://covid19.who.int. Date accessed: August 10, 2021.

[4] Zhanwei Du, Abhishek Pandey, Yuan Bai, and et al. Comparative cost-effectiveness of SARS-CoV-2 testing strategies in the USA: a modelling study. *The Lancet Public Health*, 6(3):e184–e191, 2021.

[5] Wannian (PRC) Aylward, Bruce (WHO); Liang. Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19), 2020.

[6] J Martinek, M Tatar, and M Javorka. Distinction between voluntary cough sound and Speech in volunteers by spectral and complexity analysis. *Journal of Physiology and Pharmacology*, 59(SUPPL. 6):433–440, 2008.

[7] Hanieh Chatrzarrin, Amaya Arcelus, Rafik Goubran, and Frank Knoefel. Feature extraction for the differentiation of dry and wet cough sounds. In *MeMeA 2011 - 2011 IEEE International Symposium on Medical Measurements and Applications, Proceedings*, pages 162–166. IEEE, 2011.

[8] Renard Xaviero Adhi Pramono, Syed Anas Imtiaz, and Esther Rodriguez-Villegas. A cough-based algorithm for automatic diagnosis of pertussis. *PLoS ONE*, 11(9):1–20, 2016.

[9] Yusuf Amrulloh, Udantha Abeyratne, Vinayak Swarnkar, and Rina Triasih. Cough Sound Analysis for Pneumonia and Asthma Classification in Pediatric Population. *Proceedings - International Conference on Intelligent Systems, Modelling and Simulation, ISMS*, 2015-Octob:127–131, 2015.

[10] Kathryn P Connaghan, Jordan R Green, Sabrina Paganoni, James Chan, Harli Weber, Ella Collins, Brian Richburg, Marziye Eshghi, J-P Onnela, and James D Berry. Use of Beiwe Smartphone App to Identify and Track Speech Decline in Amyotrophic Lateral Sclerosis (ALS). In *INTERSPEECH*, 2019.

[11] Jimin Lee, Emily Dickey, and Zachary Simmons. Vowel-Specific Intelligibility and Acoustic Patterns in Individuals With Dysarthria Secondary to Amyotrophic Lateral Sclerosis. *Journal of Speech, Language, and Hearing Research*, 62:1–26, 2019.

[12] Rita Chiaramonte, Carmela Luciano, Ignazio Chiaramonte, Agostino Serra, and Marco Bonfiglio. Multi-disciplinary clinical protocol for the diagnosis of bulbar amyotrophic lateral sclerosis. *Acta Otorrinolaringologica (English Edition)*, 70:25–31, 2019.

[13] B.N. Suhas Prasanta, Deep Patel, Nithin Rao, Yamini Belur, Pradeep Reddy, Nalini Atchayaram, Ravi Yadav, and Dipanjan Gope and. Comparison of Speech Tasks and Recording Devices for Voice Based Automatic Classification of Healthy Subjects and Patients with Amyotrophic Lateral Sclerosis. In *Proc. Interspeech 2019*, pages 4564–4568, 2019.

[14] Luis Garcia-Gancedo, Madeline L Kelly, Arseniy Lavrov, Jim Parr, Rob Hart, Rachael Marsden, Martin R Turner, Kevin Talbot, Theresa Chiwera, Christopher E Shaw, and Ammar Al-Chalabi. Objectively Monitoring Amyotrophic

Lateral Sclerosis Patient Symptoms During Clinical Trials With Sensors: Observational Study. *JMIR Mhealth Uhealth*, 7(12):e13433, 12 2019.

[15] Sarah E Gutz, Jun Wang, Yana Yunusova, and Jordan R Green. Early Identification of Speech Changes Due to Amyotrophic Lateral Sclerosis Using Machine Classification. In *Proc. Interspeech 2019*, pages 604–608, 2019.

[16] M Vashkevich, A Petrovsky, and Y Rushkevich. Bulbar ALS Detection Based on Analysis of Voice Perturbation and Vibrato. In *2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pages 267–272, 2019.

[17] Jun Wang et al. Towards Automatic Detection of Amyotrophic Lateral Sclerosis from Speech Acoustic and Articulatory Samples. In *INTERSPEECH*, 2016.

[18] Raquel Norel, Mary Pietrowicz, Carla Agurto, Shay Rishoni, and Guillermo Cecchi. Detection of Amyotrophic Lateral Sclerosis (ALS) via Acoustic Analysis. *bioRxiv*, 2018.

[19] An Kwanghoon Wang, Myungjong Kim, Kristin Teplansky, Jordan Green, Thomas Campbell, Yana Yunusova, Daragh Heitzman, and Jun. Automatic Early Detection of Amyotrophic Lateral Sclerosis from Intelligible Speech Using Convolutional Neural Networks. *Interspeech*, pages 1913–1917, 2018.

[20] Taylor Spangler, N. V. Vinodchandran, Ashok Samal, and Jordan R. Green. Fractal features for automatic detection of dysarthria. In *2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pages 437–440, 2 2017.

[21] Sanjana Shellikeri and others. Speech Movement Measures as Markers of Bulbar Disease in Amyotrophic Lateral Sclerosis. *Journal of speech, language, and hearing research : JSLHR*, 59(5):887–899, 10 2016.

[22] R L Horwitz-Martin and others. Relation of automatically extracted formant trajectories with intelligibility loss and speaking rate decline in amyotrophic lateral sclerosis. *Proceedings of InterSpeech*, pages 1205–1209, 2016.

[23] Panying Rong and others. Predicting Speech Intelligibility Decline in Amyotrophic Lateral Sclerosis Based on the Deterioration of Individual Speech Subsystems. *PLOS ONE*, 11(5):1–19, 2016.

[24] Jerzy Tomik, Barbara Tomik, Maciej Wiatr, Jacek Skladzien, Pawel Strek, and Andrzej Szczudlik. The Evaluation of Abnormal Voice Qualities in Patients with Amyotrophic Lateral Sclerosis. *Neuro-degenerative diseases*, 15, 2015.

[25] R Carpenter, T McDonald, and F Howard. The Otolaryngologic Presentation of Amyotrophic Lateral Sclerosis. *Otolaryngology*, 86:479–84, 1978.

[26] Alice K Silbergleit, Alex F Johnson, and Barbara H Jacobson. Acoustic analysis of voice in individuals with amyotrophic lateral sclerosis and perceptually normal vocal quality. *Journal of Voice*, 11(2):222–231, 6 1997.

[27] Panying Rong et al. Parameterization of articulatory pattern in speakers with ALS. In *INTERSPEECH*, 2014.

[28] A Frid et al. Diagnosis of Parkinson's disease from continuous speech using deep convolutional networks without manual selection of features. In *2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE)*, pages 1–4, 2016.

[29] João Paulo Teixeira, Paula Odete Fernandes, and Nuno Alves. Vocal Acoustic Analysis - Classification of Dysphonic Voices with Artificial Neural Networks. *Procedia Computer Science*, 121:19–26, 2017.

[30] Jiri Mekyska and others. Robust and complex approach of pathological speech signal analysis. *Neurocomputing*, 167:94–111, 2015.

[31] Umberto Melia, Montserrat Vallverdú, Mathieu Jospin, Erik W Jensen, Jose Fernando Valencia, Francesc Clariá, Pedro L Gambus, and Pere Caminal. Prediction of nociceptive responses during sedation by time-frequency representation. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2013:2547–2550, 2013.

[32] Umberto Melia, Francesc Claria, Montserrat Vallverdu, and Pere Caminal. Measuring Instantaneous and Spectral Information Entropies by Shannon Entropy of Choi-Williams Distribution in the Context of Electroencephalography. *Entropy*, 16:2530–2548, 2014.

[33] Francesc Claria, Montserrat Vallverdú, Rafał Baranowski, Lidia Chojnowska, and Pere Caminal. Heart rate variability analysis based on time-frequency representation and entropies in hypertrophic cardiomyopathy patients. *Physiological measurement*, 29(3):401–416, 2008.

[34] Francesc Clariá, Montserrat Vallverdú, Jordi Riba, Sergio Romero, Manuel J Barbanoj, and Pere Caminal. Characterization of the cerebral activity by time-frequency representation of evoked EEG potentials. *Physiological measurement*, 32(8):1327–1346, 8 2011.

[35] Dariusz Mika and Jerzy Józwik. Advanced Time-Frequency Representation in Voice Signal Analysis. *Advances in Science and Technology Research Journal*, 12(1):251–259, 2018.

[36] Yang Yang, Zhike Peng, Wenming Zhang, and Guang Meng. Parameterised time-frequency analysis methods and their engineering applications: A review of recent advances. *Mechanical Systems and Signal Processing*, 119:182–221, 2019.

[37] Behnaz Yousefi, Jaemin Shin, Eric H Schumacher, and Shella D Keilholz. Quasi-periodic patterns of intrinsic brain activity in individuals and their relationship to global signal. *NeuroImage*, 167:297–308, 2018.

[38] Dany Obeid, Sawsan Sadek, Gheorghe Zaharia, and Ghais El Zein. Touch-less heartbeat detection and cardiopulmonary modeling. In *2009 2nd International Symposium on Applied Sciences in Biomedical and Communication Technologies*, pages 1–5, 2009.

[39] Joshua S Richman and J Randall Moorman. Physiological time-series analysis using approximate and sample entropy. *American Journal of Physiology - Heart and Circulatory Physiology*, 278(6 47-6), 2000.

[40] Samantha J Barry, Adrie D Dane, Alyn H Morice, and Anthony D Walmsley. The automatic recognition and counting of cough. *Cough (London, England)*, 2:8, 2006.

[41] V Swarnkar, U R Abeyratne, Yusuf Amrulloh, Craig Hukins, Rina Triasih, and Amalia Setyati. Neural network based algorithm for automatic identification of cough sounds. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 1764–1767. IEEE, 2013.

[42] Sergio Matos, Surinder S Birring, Ian D Pavord, and David H Evans. Detection of cough signals in continuous audio recordings using hidden Markov models. *IEEE Transactions on Biomedical Engineering*, 53(6):1078–1083, 2006.

[43] Jia Ming Liu, Mingyu You, Guo Zheng Li, Zheng Wang, Xianghuai Xu, Zhongmin Qiu, Wenjia Xie, Chao An, and Sili Chen. Cough signal recognition with gammatone cepstral coefficients. *2013 IEEE China Summit and International Conference on Signal and Information Processing, ChinaSIP 2013 - Proceedings*, pages 160–164, 2013.

[44] Carlos Lucio, Cesar Teixeira, Jorge Henriques, Paulo De Carvalho, and Rui Pedro Paiva. Voluntary cough detection by internal sound analysis. In *Proceedings - 2014 7th International Conference on BioMedical Engineering and Informatics, BMEI 2014*, pages 405–409, 2014.

[45] Keegan Kosasih, Udantha R Abeyratne, Vinayak Swarnkar, and Rina Triasih. Wavelet Augmented Cough Analysis for Rapid Childhood Pneumonia Diagnosis. *IEEE Transactions on Biomedical Engineering*, 62(4):1185–1194, 2015.

[46] Danny Parker, Joseph Picone, Amir Harati, Shuang Lu, Marion H Jenkyns, and Philip M Polgreen. Detecting paroxysmal coughing from pertussis cases using voice recognition technology. *PLoS ONE*, 8(12):8–12, 2013.

[47] J Laguarta, F Hueto, and B Subirana. COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings. *IEEE Open Journal of Engineering in Medicine and Biology*, 1:275–281, 2020.

[48] Carnegie Mellon University. COVID Voice Detector. https://cvd.lti. cmu.edu/. Date accessed: August 25, 2021.

[49] Vocalis Health. COVID-19 Study. https://vocalishealth.com/. Date accessed: August 25, 2021.

[50] Ali Imran, Iryna Posokhova, Haneya N Qureshi, Usama Masood, Sajid Riaz, Kamran Ali, Charles N John, and Muhammad Nabeel. AI4COVID-19: AI Enabled Preliminary Diagnosis for COVID-19 from Cough Samples via an App. *IEEE Access*, pages 1–12, 2020.

[51] Madhurananda Pahar, Marisa Klopper, Robin Warren, and Thomas Niesler. COVID-19 Cough Classification using Machine Learning and Global Smartphone Recordings, 2020.

[52] Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, and et al. Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 7 2020.

[53] Jayavrinda Vrindavanam, Raghunandan Srinath, Hari Haran Shankar, and Gaurav Nagesh. Machine Learning based COVID-19 Cough Classification Models -

A Comparative Analysis. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 420–426, 2021.

[54] Redacción Médica. Coronavirus: síntomas 'falsos' que nada tienen que ver con el Covid-19, 2020.

[55] Ke Feng, Fengyu He, Jessica Steinmann, and Ilteris Demirkiran. Deep-learning Based Approach to Identify Covid-19. In *SoutheastCon 2021*, pages 1–4, 2021.

[56] Indian Institute of Science (IISc) Bangalore. Project Coswara. https://coswara.iisc.ac.in/. Date accessed: August 25, 2021.

[57] Amil Khanzada, Chandan Chaurasia, Nikki Perez, and Lisa Chionis. Virufy. https://virufy.org/. Date accessed: August 25, 2021., 2020.

[58] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. "librosa: Audio and music signal analysis in python.". In *14th python in science conference*, pages 18–25, 2015.

[59] Alberto Tena, Francec Claria, Francesc Solsona, Einar Meister, and Monica Povedano. Detection of Bulbar Involvement in Patients With Amyotrophic Lateral Sclerosis by Machine Learning Voice Analysis: Diagnostic Decision Support Development Study. *JMIR Med Inform*, 9(3):e21331, 2021.

[60] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. Springer, 2013.

[61] Kayhan N Batmanghelich, Dong H Ye, Kilian M Pohl, Ben Taskar, Christos Davatzikos, and ADNI. Disease classification and prediction via semi-supervised dimensionality reduction. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1086–1090, 2011.

[62] Ehsan Adeli, Kim-Han Thung, Le An, Guorong Wu, Feng Shi, Tao Wang, and Dinggang Shen. Semi-Supervised Discriminative Classification Robust to Sample-Outliers and Feature-Noises. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):515–522, 2019.

[63] Dan Ellis. YAMNet. https://github.com/tensorflow/models/tree/master/research/audioset/yamn
Date accessed: August 25 2021.

[64] C T Carpenter, P V Price, and B W Christman. Exhaled breath condensate isoprostanes are elevated in patients with acute lung injury or ARDS. *Chest*, 114(6):1653–1659, 12 1998.

[65] Paul Boersma and David Weenink. Praat: doing phonetics by computer [Computer program] version 6.1.01. Technical report, of the University of Amsterdam, 2019.

[66] F Hlawatsch and G F Boudreaux-Bartels. Linear and quadratic time-frequency signal representations. *IEEE Signal Processing Magazine*, 9(2):21–67, 1992.

[67] Leon Cohen. *Time Frequency Analysis: Theory and Applications.* Prentice-Hall, 1995.

[68] Patrick Loughlin. What are the time-frequency moments of a signal? *Proceedings of SPIE - The International Society for Optical Engineering*, 4474, 2001.

[69] T W Parks and C S Burrus. Digital Filter Design, John Wiley \& Sons. *chapter 7, section 7. 3*, 3.(3.), 1987.

[70] F Gustafsson. "Determining the initial states in forward-backward filtering. " *{IEEE} Transactions on Signal Processing*, 44:988–992, 1996.

[71] Sanjit K Mitra and Digital Signal Processing. *2nd ed.* McGraw-Hill, 2001.

[72] S L Marple. "Computing the {Discrete}-{Time} {Analytic} {Signal} via {FFT}. " *{IEEE}® Transactions on Signal Processing*, 47:2600–2603, 1999.

[73] IBM Corp. IBM SPSS Statistics for Windows.

[74] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning.* Springer New York, 2009.

[75] Seshadri Sastry Kunapuli and Praveen Chakravarthy Bhallamudi. Chapter 22 - A review of deep learning models for medical diagnosis. In Pardeep Kumar, Yugal Kumar, and Mohamed A Tawhid, editors, *Machine Learning, Big Data, and IoT for Medical Informatics*, Intelligent Data-Centric Systems, pages 389–404. Academic Press, 2021.

[76] K R Gabriel. The Biplot Graphic Display of Matrices with Application to Principal Component Analysis. *Biometrika*, 58(3):453–467, 12 1971.

[77] K R Gabriel. Analysis of meteorological data by means of canonical decomposition and biplots. *J Appl Meteor*, 11:1071–1077, 1972.

[78] David W Hosmer and Stanley Lemeshow. *Applied logistic regression*. John Wiley and Sons, 2000.

[79] D Dingen and others. RegressionExplorer: Interactive Exploration of Logistic Regression Models with Subgroup Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):246–255, 1 2019.

[80] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[81] Pablo Bermejo, José A Gámez, and José M Puerta. Speeding up Incremental Wrapper Feature Subset Selection with Naive Bayes Classifier. *Know.-Based Syst.*, 55:140–147, 1 2014.

[82] Alaa Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 17(1):168–192, 1 2021.

[83] Gary L Pattee and others. Provisional best practices guidelines for the evaluation of bulbar dysfunction in amyotrophic lateral sclerosis. *Muscle \& Nerve*, 59(5):531–536, 2019.

[84] University of Cambridge. COVID-19 Sounds App. https://www.covid-19-sounds.org/en/. Date accessed: August 25, 2021.

[85] Alberto Tena, Francesc Clarià, and Francesc Solsona. Automated detection of COVID-19 cough. *Biomedical Signal Processing and Control*, 71:103175, 2022.