Voice as a Digital Biomarker

Machine Learning Applications for Chronic Obstructive Pulmonary Disease Assessment

Alper İdrisoğlu

Department of Health
Blekinge Institute of Technology

Doctoral Dissertation Series no. 2025:07

Blekinge Institute of Technology
Doctoral Dissertation Series No. 2025:07
ISSN 1653-2090
ISBN 978-91-7295-503-5

Voice as a Digital Biomarker

Machine Learning Applications for Chronic Obstructive Pulmonary Disease Assessment

Alper İdrisoğlu

DOCTORAL DISSERTATION
for the degree of Doctor of Philosophy at Blekinge Institute of Technology to be publicly defended on 20251015 at 10:00 in J1630
Supervisor
Johan Sanmartin-Berglund,Professor, Blekinge Institute of Technology
Andreas Jakobsson, Professor, Lund University
Peter Anderberg, Professor, Blekinge Institute of Technology
Ana Luiza Dallora Moraes, Associate Professor, Blekinge Institute of Technology
Abbas Cheddad, Associate Professor, Blekinge Institute of Technology
Faculty opponent
Julie Wall, Professor, University of West London

Abstract

Chronic Obstructive Pulmonary Disease (COPD) is a leading cause of morbidity and mortality worldwide, with high underdiagnosis rates due to limitations in current diagnostic methods such as spirometry. This doctoral thesis explores the potential of voice as a digital biomarker to support the assessment of COPD, guided by the principles of Applied Health Technology (AHT), which emphasizes interdisciplinary collaboration and real-world applicability.

The research includes four interconnected studies. Study I presents a systematic literature review of machine learning (ML) applications for voice-affecting disorders, identifying COPD as underrepresented in current research. Study II addresses this gap by collecting a new dataset of vowel [a:] recordings from Swedish-speaking COPD patients and healthy controls once a week in self-determined quiet settings. Voice features, including baseline acoustic (BLA) parameters and Mel-Frequency Cepstral Coefficients (MFCCs), were extracted and used to train three ML classifiers: CatBoost (CB), Random Forest (RF), and Support Vector Machine (SVM). CB demonstrated the highest test accuracy at 78%.

Study III investigates the effects of signal segmentation on model performance and shows that certain temporal segments of voice recordings contain more informative patterns, enhancing classification outcomes by increasing accuracy to 85%. Study IV applies statistical and practical significance tests to compare voice features between COPD and healthy groups. A total of 34 features, including shimmer measures and higher-order MFCC derivatives, were found to meaningfully differentiate the groups.

This thesis reframes the human voice as a source of clinically relevant data, demonstrating how it can be digitized, analyzed, and interpreted using ML to aid COPD assessment. The results indicate that voice-based analysis can provide an accessible, non-invasive, and scalable complement to existing diagnostic tools. By integrating technical, clinical, and ethical perspectives, the thesis contributes new knowledge and practical methodologies that align with AHT's goal of creating value-driven, user-centered healthcare solutions. The findings support future development of mobile and remote voice-based screening tools for COPD and other conditions.

Keywords: Chronic Obstructive Pulmonary Disease; Machine Learning; Non-invasive Diagnostic; Segmentation; Voice-based Analysis.

Blekinge Institute of Technology
Doctoral Dissertation Series No. 2025:07

Voice as a Digital Biomarker

Machine Learning Applications for Chronic Obstructive Pulmonary Disease Assessment

Alper İdrisoğlu

Doctoral Dissertation in Applied Health Technology

Department of Health
Blekinge Institute of Technology
SWEDEN

Blekinge Institute of Technology
Department of Health

Blekinge Institute of Technology Doctoral Dissertation Series No. 2025:07
ISBN 978-91-7295-503-5
ISSN 1653-2090
urn:nbn:se:bth-28038

Printed in Sweden by Media-Tryck, Lund University, Lund 2025

Alongside my family, I dedicate this work to Individuals who have faith in and embrace science.

Dünyada her şey için, maddiyat için, maneviyat için, başarı için en hakiki mürşit ilimdir, fendir. İlim ve fennin dışında mürşit aramak gaflettir, cehalettir, delalettir.

Mustafa Kemal Atatürk 1924

Acknowledgements

First and foremost, I would like to extend my sincere gratitude to my research supervisors: Associate Professor Ana Luiza Dallora Moraes, Associate Professor Abbas Cheddad, Professor Peter Anderberg, Professor Andreas Jakobsson, and Professor Johan Sanmartin Berglund. Their guidance, encouragement, and expertise have been fundamental to the success of this research.

I am especially thankful to Associate Professor Ana Luiza Dallora, my daily supervisor, whose consistent support, availability, and insightful feedback helped steer the project forward. I am equally grateful to Professor Johan Sanmartin Berglund, my main supervisor, for his clinical perspective and unwavering commitment to ensuring high scientific and medical standards. The collective contributions of Associate Professor Abbas Cheddad, Professor Peter Anderberg, and Professor Andreas Jakobsson through their knowledge, mentorship, and resources have been instrumental throughout this journey.

My thanks also go to the Department of Health for cultivating a positive and collaborative research environment, and to my fellow Ph.D. candidates for their collegiality and thoughtful exchange of ideas. I would especially like to acknowledge Dr. Johan Flyborg and Ulrika Isaksson for their close friendship and collaboration. I am also thankful to the dedicated team at the Blekinge Institute of Technology Research Clinic for their essential role in participant recruitment and data collection.

I extend heartfelt appreciation to all the study participants for their time and engagement, without whom this research would not have been possible.

This work was supported by the Excellence Center at Linköping – Lund in Information Technology (ELLIIT), whose funding and backing I gratefully acknowledge.

Finally, I owe a profound debt of gratitude to my family, whose love, patience, and steadfast support have carried me through every stage of this journey.

List of Papers

Study I

Idrisoglu A, Dallora AL, Anderberg P, Berglund JS. Applied Machine Learning Techniques to Diagnose Voice-Affecting Conditions and Disorders: Systematic Literature Review. Journal of Medical Internet Research. 2023;25: e46105. doi:10.2196/46105.

Study II

Idrisoglu A, Dallora AL, Cheddad A, Anderberg P, Jakobsson A, Sanmartin Berglund J. COPDVD: Automated Classification of Chronic Obstructive Pulmonary Disease on a New Developed and Evaluated Voice Dataset. Artificial Intelligence in Medicine. 2024;156. doi:10.1016/j.artmed.2024.102953.

Study III

Idrisoglu, A., Moraes, A. L. D., Cheddad, A., Anderberg, P., Jakobsson, A., & Berglund, J. S. (2025). Vowel segmentation impact on machine learning classification for chronic obstructive pulmonary disease. *Scientific Reports*, *15*(1), 9930. https://doi.org/10.1038/s41598-025-95320-3

Study IV

Idrisoglu A, Dallora AL, Cheddad A, Anderberg P, Jakobsson A, Sanmartin Berglund J. Feature Analysis of the Vowel [a:] in Individuals with Chronic Obstructive Pulmonary Disease and Healthy Controls. (Submitted to Journal)

Author contribution

Study I

AI is the primary contributor to the study and manuscript, and he is involved in all aspects. ALD assisted in the design of the study, simultaneous study selection, and manuscript revisions. PA contributed to the final revisions of the manuscript. JSB provided medical expertise in the field and assisted with manuscript revisions.

Study II

The study and manuscript were predominantly shaped by AI by writing the original draft, playing a key role in all facets. ALD and AC contributed to the study's design, provided supervision in Machine Learning, and were involved in writing - review and editing. PA played a role in supervision and the final manuscript writing - review and editing. AJ provided the resources used for data collection and writing - review and editing. JSB provided resources, supervision, and medical expertise in the field.

Study III

AI conducted the experiment and conceptualized and wrote the original draft of the manuscript. ALD, AC, PA, AJ, and JSB analyzed the results, revised the manuscript, and supervised and provided resources. All authors reviewed the manuscript.

Study IV

AI: Writing – review and editing, Writing – original draft, Methodology, Conceptualization. ALDM: Writing – review and editing. AC: Writing – review and editing, Validation. SW: Writing – review and editing, Validation. PA: Writing – review and editing. AJ: Writing – review and editing, Validation, Methodology. JSB: Writing – review and editing, Validation, Supervision.

Abbreviations

AD	:	Alzheimer's Disease
AI	:	Artificial Intelligence
ALS	:	Amyotrophic Lateral Sclerosis
AHT	:	Applied Health Technology
ANN	:	Artificial Neural Network
BLA	:	Baseline Acoustic
BTH	:	Blekinge Institute of Technology
CB	:	Cat Boost
CD	:	Cardiovascular Disease
COPD	:	Chronic Obstructive Pulmonary Disease
DT	:	Decision Tree
EMA	:	Ecological Momentary Assessment
ET	:	Essential Tremor
GB	:	Gradient Boosting
GDPR	:	General Data Protection Regulation
GMM	:	Gaussian Mixture Model
HNR	:	Harmonic to Noice Ratio
KNN	:	K-Nearest Neighbor
LG	:	Logistic Regression
MCI/CI	:	Mild Cognitive Impairment/Cognitive Impairment
MeML	:	Mixed Effect Machine Learning
MFCC	:	Mel Frequency Cepstrum Coefficient
ML	:	Machine Learning

MS	:	Multiple Sclerosis
MAX	:	Maximum Value (The Highest value)
NB	:	Naïve Bayes
PA	:	Passive Active
PERMANOVA	:	Permutational analysis of variance)
PD	:	Parkinson's Disease
PRISMA	:	Preferred Reporting Items for Systematic Reviews and Meta-Analyses)
RF	:	Random Forest
SHAP	:	Shapley Additive exPlanations
SLR	:	Systematic Literature Review
SVM	:	Support Vector Machine
SVR	:	Support Vector Regression
TNR	:	True Negative Rate
TPR	:	True Positive Rate
VAD	:	Voice activity detection
WHO	:	World Health Organization

List of Tables

List of Figures

Table of Contents

Introduction

Healthcare systems are increasingly challenged by rising patient numbers, complex care needs, escalating costs, rapid technological developments, and evolving socio-political conditions across different countries [1–3]. In response, many systems are shifting toward innovative service models that integrate new technologies and adapt to dynamic patient expectations [4,5].

In this context, the field of Applied Health Technology (AHT), as pursued at Blekinge Institute of Technology, aims to address current and future healthcare challenges through interdisciplinary collaboration, co-creation, and real-world implementation of technological solutions [6]. Among these, artificial intelligence (AI)-driven decision support systems and screening tools are being explored as promising innovations [7]. Numerous studies have investigated the integration of AI in various diagnostic and treatment domains [8], including the emerging use of voice as a digital biomarker to detect or monitor disorders that influence speech production [9–11].

One such disorder is Chronic Obstructive Pulmonary Disease (COPD), a progressive condition that primarily affects respiratory function by reducing expiratory flow and lung capacity [12,13]. Beyond its pulmonary manifestations, COPD is associated with systemic comorbidities that may also influence vocal characteristics [13]. Several studies have reported that COPD alters specific acoustic features of the voice, including baseline acoustic (BLA) parameters such as jitter and shimmer, as well as Mel-Frequency Cepstral Coefficients (MFCCs) [12,14–17]. These features are extracted through signal processing techniques that convert the voice signal into measurable digital parameters. However, the high dimensionality of these feature sets often exceeds the capacity for manual analysis. To address this, researchers have increasingly turned to machine learning (ML), a subfield of AI, capable of identifying patterns and relationships within large, complex datasets.

While the use of ML in voice analysis for COPD remains relatively novel and underexplored area, the broader integration of AI in respiratory medicine has gained significant momentum. Numerous studies have applied AI techniques for early detection, classification, and monitoring of respiratory conditions such as asthma, pneumonia, and COVID-19. For instance, deep learning models have been effectively used to analyze chest X-rays and CT scans to identify lung abnormalities with high sensitivity and specificity [18,19]. Additionally, wearable sensor data and electronic

health records have been leveraged to predict respiratory exacerbations and hospital readmissions [20,21].

Despite these advances, the majority of AI-based tools in respiratory health focus on imaging or structured clinical data, with limited exploration of non-invasive and accessible alternatives like voice analysis. This is noteworthy given that vocal alterations are recognized symptoms of several pulmonary and systemic diseases. The limited attention to voice as a diagnostic signal underscores a critical research gap, one that this thesis directly addresses.

The aforementioned developments demonstrate a high level of technological readiness and confirm that the timing is appropriate for investigating voice as a digital biomarker. In this context, the present thesis contributes to the advancement of voice-based diagnostic and decision-support tools by building upon current innovations, exploring state-of-the-art developments, and expanding the knowledge base with a specific focus on COPD. By applying ML models to voice data, this work provides a complementary or even alternative diagnostic modality, particularly valuable in low-resource or remote settings.

Scope of the Thesis

This thesis delves into the landscape of disorders affecting the voice that are not directly linked to pathological issues in the larynx (voice box). By leveraging existing evidence, it aims to channel technologies used in prior studies into the realm of COPD assessment. The dissertation includes studies that provide a broader understanding of a voice-based decision support system's state-of-the-art application. It focuses on analyzing recordings of the Swedish utterance of the vowel [a:] from individuals with COPD and healthy control (HC) groups from the perspective of AHT. The application of ML by using voice features to support the decision-making process shifts the research direction toward a rarely investigated pulmonary condition regarding the use of voice as a digital biomarker in the decision support of COPD diagnosis.

The core purpose of AHT is not primarily to develop new technologies, but to apply existing and emerging technologies within the healthcare domain to generate meaningful value for patients, professionals, and health systems. Rather than focusing on technical innovation alone, AHT emphasizes the translation of technological advancements into practical, context-aware solutions through interdisciplinary collaboration. The studies included in this thesis reflect this purpose by exploring how ML and voice analysis can be used to develop tools that support the assessment of COPD, laying the groundwork for future clinical applications.

As aligned with the purpose of AHT, this thesis demonstrates the methods and outcomes of interdisciplinary research with a specific focus on COPD assessment and the use of voice as a digital biomarker. The integration of ML reflects the use of emerging technologies, but the true strength of the work lies in how the results are analyzed, not only through a technological view but also from the perspectives of healthcare and clinical practice. This multifaceted approach exemplifies the essence of interdisciplinarity, where different fields collaborate to address specific healthcare challenges in a comprehensive and cohesive manner, as demonstrated by the studies included in this thesis.

Background

This thesis brings together the research conducted at BTH within the DiaVoc project, registered under DNR: BTH-6.1.1-0061-2023, which is funded by the Excellence Center at Linköping – Lund in Information Technology (ELLIIT) under DNR: BTH-6.1.1-0135-2020. The DiaVoc project consists of two PhD positions, one based at BTH and the other at Lund University. The project is coordinated by Principal Investigator Andreas Jakobsson and co-Principal Investigator Johan Sanmartin Berglund.

The responsibilities within the project are divided between the two institutions. Lund University is responsible for the signal processing components, while the AHT research is conducted at BTH. The work presented in this thesis is situated within the BTH branch of the project, focusing on the application of ML and voice analysis as decision-support tools in healthcare, with particular emphasis on the assessment of COPD.

Applied Health Technology

The studies discussed in the following sections are grounded in the perspective of AHT, a research field at the BTH that explores the possible application areas of emerging technologies, examines how technology influences health, and how users perceive health-related technologies.

AHT is an interdisciplinary domain that bridges various fields to address shared health challenges through the adoption of emerging technologies [6]. This approach underscores the importance of transdiciplinary collaboration, where disciplines not only work together but also integrate their methods and perspectives, which fostering the exchange of experiences and knowledge across disciplines. Such collaboration aims to enhance mutual understanding, facilitate the integration of future technologies, and address emerging health-related challenges [22]. By bringing together expertise from fields such as engineering, healthcare, psychology, information technology, and user experience, AHT creates a comprehensive framework for innovation. This holistic perspective enables the development of user-centered, effective, and accessible solutions.

Chronic Obstructive Pulmonary Disease

COPD is a progressive respiratory condition characterized by persistent airflow limitation and chronic inflammatory responses in the airways and lungs, typically caused by long-term exposure to noxious particles or gases, most commonly cigarette smoke [23,24]. It encompasses clinical entities such as chronic bronchitis and emphysema and is associated with symptoms including dyspnea, chronic cough, and sputum production [25,26]. Beyond pulmonary manifestations, COPD is increasingly recognized as a systemic disease with extrapulmonary effects, contributing to comorbidities and reduced quality of life [27,28]. The global burden of COPD is substantial, and it remains a leading cause of morbidity and mortality worldwide [29].

Challenges in COPD assessment and potential with AI-based solutions

According to the reports from the World Health Organization (WHO), COPD is a global burden that holds the third place in the list of the most common causes of death globally [29]. The common clinical practice to assess and diagnose COPD relies mainly on medical history, CT scans, and spirometry measurements [30–32]. Consequently, the global under-diagnosis rates of 70% to 90% mean that the majority of individuals who actually have COPD remain unidentified, highlighting major shortcomings in current decision-making approaches [33]. There is a vast potential for the absence of severe symptoms at the early stages of COPD to contribute to the mentioned under-diagnostic rates [34,35].

On the other hand, there is accelerated immigration and application of AI and ML techniques in every area of life, including health care [36]. Development in ML and voice-based feature extraction techniques offer the potential to apply this technique for COPD assessment as a new, innovative, and non-invasive tool to support decision-making [37–39]. Success in this area holds the potential to enable timely and personalized interventions and can revolutionize today's clinical practices toward more proactive health care. An additional potential is the decreased use of healthcare resources associated with the late diagnosis and reduced morbidity caused by COPD.

From the patient's point of view, a voice-based system can make it possible to integrate the patient into the treatment process in the form of self-monitoring. In this way, the patient can make more active choices about their health, supporting person-centered care. The adoption of this technology into mobile applications, especially offers borderless applications.

COPD and Voice Association

COPD is a progressive respiratory condition marked by persistent airflow restriction, which is frequently linked to chronic bronchitis and emphysema [40]. This disease affects millions of individuals across the globe and ranks among the leading causes of disability and mortality [40]. The breathing difficulties characteristic of COPD not only compromise respiratory function but also interfere with the mechanisms responsible for voice production [12,15,41]. As a result, patients may experience vocal fatigue, hoarseness, and diminished control over their voice, which can make it challenging for them to communicate effectively with others [12,15,41]. This additional challenge amplifies both the physical and psychological burdens associated with the disease.

Despite the substantial volume of research focused on voice-related disorders, COPD remains relatively unexplored in this context [42]. While spirometry is widely recommended as a diagnostic tool for confirming COPD, the high prevalence of underdiagnosis, overdiagnosis, and misdiagnosis highlights the need for innovative thinking [43,44]. Early diagnosis is essential for effective disease management [44]Yet the limitations of current methods point to a growing demand for more accurate and accessible diagnostic tools.

Voice, a fundamental aspect of human evolution and communication, has been utilized as a means of interpersonal connection for centuries. Recent advancements in technology and the analysis of vocal features have opened new possibilities for medical diagnostics. By leveraging vocal attributes as digital biomarkers, researchers are now exploring less invasive methods for detecting and classifying voice-impacting disorders [11]. This approach has the potential to revolutionize healthcare, offering simpler and more accessible diagnostic alternatives.

For individuals with COPD, the integration of ML with vocal feature analysis presents a promising avenue for enhancing diagnostic precision. By analyzing patterns within a patient's voice, ML models could help identify COPD-related anomalies, supporting healthcare providers in making more accurate diagnoses. Such decision-support tools have the potential to reduce reliance on conventional diagnostic methods and improve early detection rates.

Voice Features

As a product of digitization and the emergence of signal-processing technology, the quantification of voice into measurable metrics, often referred to as digital voice/vocal features, or, in the case of health-related usage, the features were identified as digital biomarkers [11,45]. This development allowed researchers to leverage computational tools for processing and analyzing vocal signals, facilitating more objective and data-driven assessments of voice disorders [46].

These features serve as key indicators of vocal characteristics and are critical for the evaluation of voice quality and pathology. A wide array of feature extraction techniques has been developed to derive useful information from voice recordings, with different methods tailored to suit diverse research and clinical purposes [47,48]. Moreover, the types of voice recordings collected can vary depending on the objective, as certain pathological voice conditions require specialized recording methods for effective analysis [11].

The most frequently used vocal features for voice pathology assessment are BLA and MFCC features [49]. BLA features capture essential properties such as jitter, shimmer pitch, intensity, and harmonic content, while MFCCs provide a compact representation of the advanced spectral properties of the voice. Together, these features play an important role in the detection, classification, and monitoring of pathological voice conditions [42,50,51], supporting advancements in both clinical diagnosis and voice research.

Machine Learning

ML is a fundamental branch of AI that empowers computers to learn from experience and improve their performance without being explicitly programmed. The origins of ML can be traced back to the mid-20th century, when significant theoretical and practical milestones laid the groundwork for modern AI systems.

In 1943, Warren McCulloch and Walter Pitts proposed the McCulloch-Pitts neuron, the first mathematical model of an artificial neuron [52]. Their work demonstrated that a network of interconnected neurons could perform logical operations such as AND, OR, and NOT, establishing a theoretical foundation for neural networks and highlighting the computational potential of biological systems. A decade later, in the 1950s, Arthur Samuel introduced the concept of ML, defining it as "a field of study that gives computers the ability to learn without being explicitly programmed." He developed a checkers-playing program that learned and improved its performance through past experiences, marking one of the first examples of a ML system in action [53]. Building upon this foundation, Frank Rosenblatt introduced the Perceptron in 1958, which was one of the earliest practical implementations of an artificial neural network [54]. The Perceptron could classify data points into categories based on training examples and demonstrated the ability of machines to learn simple patterns. This model laid the groundwork for future developments in neural networks and deep learning.

At its core, ML operates by learning from examples, capturing patterns in data, and using these patterns to improve performance on specific tasks. The primary learning paradigms include supervised learning, unsupervised learning, and end-to-end learning. Supervised learning involves training models on labeled datasets to learn

mappings from inputs to defined outputs and is widely applied in classification tasks [55]. Unsupervised learning, in contrast, works with unlabeled data, allowing models to discover inherent structures, such as clusters or patterns. Common techniques include clustering algorithms and dimensionality reduction methods like Principal Component Analysis (PCA) [56]. End-to-end learning, while generally a form of supervised learning, simplifies complex processing pipelines by enabling models to learn directly from raw inputs to target outputs. This approach is commonly employed in deep learning architectures, including convolutional neural networks (CNNs) for image recognition and sequence-to-sequence models for natural language processing tasks [57].

Several key concepts are essential for understanding ML workflows, including the training set, validation set, and test set, used for model training, fine-tuning, and evaluation, respectively. Other important terminologies include features (input data characteristics), parameters (internal variables within the model), hyperparameter tuning (optimization processes), cross-validation (preventing overfitting), and metrics for performance evaluation, such as accuracy, precision, recall, and the F1-score [58,59].

The experimental studies, Study II and Study III, included in this thesis primarily employ supervised learning techniques. In supervised learning, the algorithm is trained on a dataset that contains both input data and corresponding target labels. This approach allows the model to learn a mapping from inputs to outputs by minimizing the error between the predicted and actual labels during the training process (training set data). The central assumption is that the patterns learned from the labelled training data will generalize well to unseen data (validation set and/or test set data). This approach offers several advantages. First, it provides a transparent framework for performance evaluation using well-established metrics like accuracy, precision, recall, and the F1-score. Second, it allows for model interpretability and feature importance analysis, which are crucial for understanding the physiological or clinical relevance of specific voice features.

Use of Machine Learning in Healthcare

Over the decades, ML has experienced substantial growth and paradigm shifts, driving innovations in numerous fields. ML techniques now enable self-driving cars, where image recognition systems identify road signs, pedestrians, and obstacles [60]. They also power human-machine interaction through advanced voice recognition systems, such as virtual assistants and speech-to-text technologies [61]. In healthcare, ML facilitates decision support systems by identifying complex patterns within medical data, enabling early diagnosis, predictive analytics, and personalized treatment strategies [62,63].

The purpose of utilizing ML in the healthcare environment is not only to reduce the use of healthcare resources but also to improve the speed and accuracy of physicians [64]. The use of ML approaches in healthcare has shown increasing promise for a variety of objectives, including risk assessment [65]. The flexibility and scalability of ML algorithms, compared to standard biostatistical methods, are key advantages that make them suitable for a wide range of tasks such as risk stratification, diagnosis and classification, and survival predictions [66]. ML has emerged as a transformative force in many domains, and healthcare is no exception [67]. Increased access to medical data, in conjunction with new developments in AI methodologies, opens up the possibility of creating advanced systems that can support medical professionals [68,69]. The ability to process complex datasets is a core strength of ML, making it especially useful in healthcare, where enormous amounts of data are generated on a regular basis [70–72]. Voice features are one example of such complex data, ranging from a few to thousands of data points per recording, often beyond human capacity to process. Therefore, ML algorithms are highly suitable for the purposes explored in studies included in this thesis, as they can uncover trends and make predictions by analyzing large and complex datasets [73–75].

Voice Affecting Disorders

Voice creation depends on harmony between multiple physiological systems, as Figure 1 illustrates, and voice health assessment spans multiple disciplines, including speech pathology and respiratory medicine [76]. While conditions that affect voice quality are generally grouped under the broad category of voice disorders, this doctoral thesis adopts a more specific definition. Here, voice-affecting disorders are considered a distinct subgroup, those that influence vocal characteristics without directly involving the larynx (voice box). This subgroup includes categories outlined in the voice disorder classification manual, such as Non-laryngeal aerodigestive disorders, Systemic conditions, and Neurological disorders, all of which can impact voice production [77]. In this thesis, Studies II, III, and IV specifically investigate the use of voice as a digital biomarker for assessing COPD, which falls under the category of Non-laryngeal aerodigestive disorders.

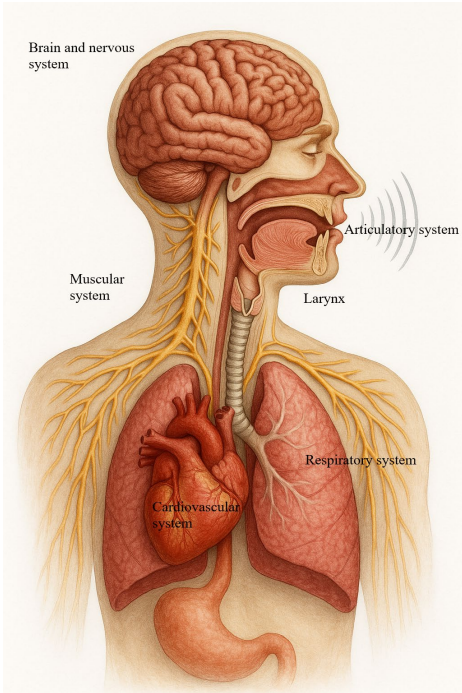


Figure 1. Several systems involved in voice production.

Using Voice as a Digital Biomarker

Voice is not just a means of communication; it is a complex phenomenon shaped by both physiological and psychological factors, offering valuable insights into an individual's emotional state and health [78,79]. An emerging field that targets to benefit from voice as a digital biomarker is focusing on exploring how vocal characteristics can reveal information about various physiological and medical conditions [80]. Using voice as a diagnostic tool provides a non-invasive and accessible approach to health monitoring, offering a potential way to assess a person's well-being and supporting the idea of person-centered care [81,82].

The foundation of this concept is the idea that changes in voice patterns can be linked to specific health issues, including mood disorders like depression, anxiety, and stress, each of which is often accompanied by marked changes in voice production [83,84]. The human voice results from the coordinated efforts of three main systems: the respiratory system (which controls airflow and pressure), the phonatory system (which produces sound through vocal fold vibrations), and the articulatory system (which shapes sounds with the mouth, tongue, and throat) [79]. When any of these systems experience disruptions due to factors such as

neurological conditions, respiratory disorders, or emotional states, these changes are often reflected in the voice's acoustic properties, including pitch, intonation, rhythm, and spectral features [83–87].

All aforementioned solutions are based on the quantification of the voice signal using different feature extraction techniques, transforming raw audio into measurable parameters that capture phonatory, spectral, and temporal characteristics. These features serve as potential digital biomarkers, enabling the use of voice as a non-invasive, cost-effective, and scalable tool for assessing health conditions, including respiratory disorders such as COPD [88,89]. In this thesis, the focus is placed on a selected set of commonly used voice features, including jitter and shimmer ,which reflect frequency and amplitude microvariations and are widely associated with vocal instability [90], HNR, which represents the degree of noise in the voice signal [91], and MFCCs, which provide a compact representation of the spectral envelope of speech and are commonly used in both clinical voice assessment and speech recognition tasks [92,93]. While other advanced approaches such as deep spectral representations, prosodic analysis, and nonlinear dynamic measures have been explored in the literature [11,51,94–96], only the BLA and MFCC feature types are utilized in the present studies due to their established relevance in voice analysis and their compatibility with the applied ML models.

Rationale

After ischemic heart disease and Stroke, COPD holds third place as the leading cause of death among non-communicable diseases (disease that is not transmissible directly from one person to another) in the list of leading causes of death in 2021 globally of WHO [97]. Regarding the high-income countries, including Sweden, according to the WHO data for the same date, COPD is ranked fifth as a non-communicable disease. Recent statistics on the cause of death from the Swedish Social Department for 2023 indicate that this trend does not have any signs of change [98]. A recent study reported a worldwide 10.3% COPD prevalence, corresponding to around 400 million individuals between 30-79 years old [99]. Even if the prevalence of COPD decreases in Sweden, it remains quite high, around 7% for the age group between 21-78 years old, and high under diagnosis and misclassification rates of 70% to 90% were reported [33]. In that regard, the application of new emerging technologies in the assessment of COPD to aid the decision-making process holds the potential of exploring new tools that may contribute to a trend change in detecting this disorder in early stages. The research field of AHT aligns well with this vision, as it focuses on exploring and adapting new technologies to address real clinical challenges, with the ultimate goal of enhancing healthcare delivery and patient outcomes through interdisciplinary collaboration.

Aims

This PhD thesis aims to contribute to the growing body of knowledge on the potential of voice as a biomarker in healthcare by leveraging ML techniques to address critical challenges in the assessment of voice-affecting conditions, particularly COPD. Through research focused on employing features such as BLA features and MFCCs, exploring the impact of segmentation strategies, and integrating demographic variability, this work provides valuable insights and methodologies that support the development of robust and scalable ML models. Grounded in the interdisciplinary domains of AHT, speech pathology, and ML, the thesis strives to address gaps in the research landscape and offer validated frameworks and practical tools that might help to build ideas for the future integration of voice-based biomarkers into healthcare systems.

Study I

Voice disorders encompass a range of conditions affecting vocal quality, pitch, and resonance, with ML algorithms increasingly utilized to enhance diagnostic accuracy and efficiency. Numerous studies have explored the integration of ML with voice features for monitoring and diagnostic purposes. Study I presents an SLR where the aim was to synthesize the current state of research on voice-affecting disorders and the ML models applied to support diagnosis and monitoring. The study sought to provide a comprehensive overview of ML algorithms, their purposes, the voice features utilized, information on the trends, and the data available. Furthermore, the SLR aims to identify gaps in the existing research, offering insights to guide future studies and advancements in this domain.

Study II

Building on the findings of Study I, COPD emerged as a voice-affecting condition that has been underexplored in research, despite its severe health implications and established impact on voice. The lack of publicly available COPD voice datasets

was identified as a major barrier to advancing research in this area. Study II aimed to address and contribute to this gap by developing a novel Swedish COPD voice dataset and evaluating the performance of several ML models, Random Forest, Support Vector Machine, and CatBoost, for binary classification of COPD versus non-COPD. The study intended to identify the best-performing model and highlights the most significant voice features, providing a foundational dataset and methodology for future COPD-related voice research.

Study III

Various methods for voice activity detection (VAD) exist, but the effects of different segmentation strategies on the performance of ML models have not been thoroughly explored. Study III aims to investigate how segmentation techniques applied to vowel [a:] recordings influence the performance of ML models. The study combines features extracted from smaller segments or combines features of segments of these recordings and evaluates how different segmentation approaches affect model accuracy and reliability. The study also intended to provide a clinical perspective based on the observed effects, using metrics such as True Positive Rate (TPR), True Negative Rate (TNR), and confusion matrix analysis to assess the diagnostic relevance and effectiveness of the segmentation strategies in identifying voice-related conditions. The goal is to determine if some segments contain more information for the ML models that give a better outcome while offering clinical insights for diagnosis.

Study IV

In Study IV, the aim was to conduct a statistical and practical significance analysis of voice features in COPD patients compared to HC. The intent to examine significant differences in voice characteristics using baseline acoustic features, MFCCs, and vowel quadrilateral analysis to identify differences in vowel articulation between groups. The analysis goes beyond comparing COPD and HC, also considering gender differences within each group to assess how utterance patterns may vary. Statistical methods such as the Mann-Whitney U and PERMANOVA (Permutational analysis of variance) test aimed to determine significant differences between groups, with p-value analysis to evaluate the significance of these differences. Additionally, effect size is calculated using Clift Delta, and confidence intervals for effect sizes are included aiming to assess the strength and reliability of the findings. The study strived to provide clinically actionable insights by linking these statistical and practical findings to real-world

healthcare applications, offering evidence to support the use of voice features as biomarkers for COPD decision support.

Materials and Methods

Study design

The present thesis comprises four quantitative studies that explore the application of ML and voice analysis to support decision-making in the assessment of COPD. Study I investigates the state of the art through a systematic literature review (SLR), while the remaining studies focus on dataset and model development, methodological optimization, and model interpretation, as outlined in Table 1.

Table 1. Methodological overview and chronological phase of the four studies included in this thesis.

Description	Study I	Study II	Study III	Study IV
Design	Quantitative	Quantitative	Quantitative	Quantitative
Approach	SLR	Exploratory	Observational	Statistical
Data	Peer-reviewed publications included in the SLR	Voice recordings and health data from individuals with COPD and HCs	Voice recordings collected during Study II	Dataset extracted during Study II
ML use	No	Yes	Yes	No
Ethical Approval	Not Applicable	Yes	Yes	Yes
Main contribution	Foundation for future direction and identified research needs	Collection and creation of a voice dataset and experimentation using ML for classification	Highlighting the importance and effects of signal processing steps to optimize ML outcomes	A deeper understanding of digital voice biomarkers

Exploring the state of the art in the context

To conduct rigorous and transparent research, a comprehensive understanding of the existing body of knowledge is essential. In study I, a SLR methodology was applied to identify, evaluate, and synthesize relevant research in a structured and reproducible manner. The review was based on a predefined protocol, developed in accordance with the guidelines proposed by Kitchenham et al. [100], which are widely adopted for evidence-based reviews. The protocol defined the research objectives, inclusion and exclusion criteria, search strategy, and procedures for screening and data extraction[1]. To ensure transparency and methodological rigor, the review process followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [101].

A comprehensive search strategy was designed and applied across three major databases: Scopus, Web of Science, and PubMed. Search strings were constructed using predefined keywords and Boolean logic to ensure broad and relevant coverage of the literature. To ensure scientific quality and reliability, the review focused exclusively on peer-reviewed journal publications, excluding conference proceedings, dissertations, and other forms of grey literature. This decision was based on the aim to include studies that had undergone rigorous academic scrutiny. However, this exclusion was applied with caution, as prior literature has highlighted the risk of omitting valuable early-stage or domain-specific research, particularly in fast-developing or interdisciplinary fields. The approach is consistent with systematic review standards while acknowledging the potential trade-off between rigor and inclusiveness. Screening and eligibility assessments were performed in parallel by the authors, with consultation from domain experts to resolve any uncertainties. While the methodology was carefully designed to ensure rigor and reproducibility, certain limitations were acknowledged, including the number of databases used, language, and the exclusion of grey literature. Data from the included studies were extracted using structured extraction templates, ensuring consistency across all extracted variables. The full review process is illustrated in Figure 2, which presents a high-level workflow.

[1] The protocol is available at https://github.com/AIITPlanet/Protocol.git. Accessed: 2025-06-02

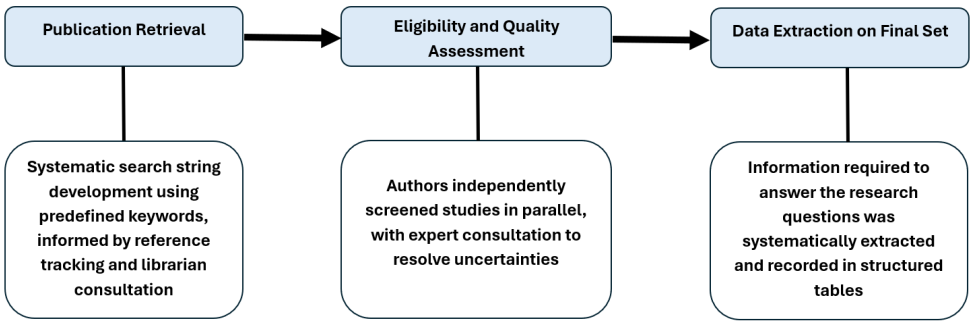


Figure 2. High-level workflow of the systematic literature review.

Data collection and experimentation

Study II was designed based on the results of Study I and the compliance with the ethical approval, where COPD was included as one of the disorders to be investigated in regard to the effects on voice and the possible use of ML for the assessment and the classification of disease to support the decision-making process.

Since the ethical approval was already in place, the data collection process was initiated in parallel to Study I, targeting a broad spectrum of participants such as individuals with Parkinson's disease, COPD, and myasthenia gravis, and participants without a diagnosed voice affecting disorder as a control group. Inclusion of other disorders than COPD was based on the fact that the outcome of study I was unknown. However, at that time, the participant portfolio of the research clinic was leaning more towards individuals with COPD diagnoses and control groups, which allowed for the collection of more voice data than other targeted groups and the experiment to be conducted in Study II. The inclusion and exclusion criteria are described below:

Inclusion criteria

COPD group:

- Being 18 years old or older.
- Having a COPD diagnosis given by a physician.
- Having a smartphone.

HC group:

- Being 18 years old or older.
- Not having a diagnosed voice-affecting disorder.
- Having a smartphone.

Exclusion criteria

COPD group:

- Being younger than 18 years old.
- Having a voice-affecting disorder other than COPD.
- Declaring no access or proficiency to use a smartphone.

HC group:

- Being younger than 18 years old.
- Having any voice-affecting disorder.
- Declaring no access or proficiency to use a smartphone.

Based on the results and identified gaps in Study I, a project plan was prepared and registered both at BTH and clinic trials, with registration numbers "BTH-6.1.1-0074-2023" and "NCT05897944", respectively, for Study II. Based on the objectives in the project plan, in Study II, a COPD voice dataset was developed from the participants' voice recordings who recorded their voices using a mobile application over six months. The dataset included BLA, MFCCs, and some demographic parameters registered per recording for classification purposes. These features were among the most used ones in Study I. ML classifiers (RF, SVM, and CatBoost (CB)) were employed to perform binary classification of COPD vs. HC. Even this choice was based on their frequent use and strong performance in Study I. However, the inclusion of CB was based on a suggestion in the author group and its promising results in the literature, where CB was presented as a novel approach [102,102–104]. The performance of the classifiers was evaluated using accuracy, precision, recall, F1-score, and ROC analysis. Feature importance was assessed using Shapley Additive exPlanations (SHAP) values to interpret the contribution of individual voice features. SHAP is a unified approach to explain the output of ML models based on cooperative game theory, providing consistent and locally accurate attributions for each feature's contribution to the model's prediction [105]. It has been widely adopted in biomedical and voice analysis applications due to its ability to enhance model interpretability and support transparent decision-making in clinical contexts [106–108].

Investigating the effects of signal processing on the ML performance

One of the gaps identified in the study I was the absence of a voice dataset with a focus on COPD. Study II has addressed this gap by creating a new COPD voice

dataset for classification purposes. During the creation of the dataset, many different signal processing techniques and research papers regarding the voice creation theory were scanned [109–115]. This process raised questions about methodologies used to construct features and their relationship with the voice as a signal. Many articles acknowledged the voice as a dynamic signal [116,117]. However, no paper has investigated the effects of signal processing. For example, using windowing with or without overlap, which might affect the informative quality of a dynamic signal, in turn, the ML outcome [118].

Study III focused on the effect of segmentation on ML performance for classifying COPD-related voice recordings. Segmentation was applied to vowel [a:] recordings collected in Study II to create smaller segments for feature extraction. The performance of ML algorithms was evaluated using metrics such as accuracy, precision, recall, and F1-score. The study also examined the clinical significance of segmentation effects by analyzing TPR, TNR, and confusion matrix results, providing insights into the impact of segmentation on diagnostic accuracy. This study utilized the same methodology as Study II with some adjustments. The nCV was set to max 5X5 inner and outer loops to minimize the computational cost of increased feature vectors per recording. Applying different segment combinations allowed the simulation of different techniques, such as windowing with or without overlapping. However, a key difference from Study II was the evaluation of the effect of various segments on ML performance. Therefore, the analyses were based on the traditional ML performance metrics and their reflections on clinical relevance measures for each combination of segments.

Even Study III was registered separately on Clinical trials and locally at BTH with the registration numbers "NCT06160674" and "BTH-6.1.1-0169-2023" to respect the transparent research expectations and best practice requirements for research involving human subjects in clinical research.

Association of voice features with COPD

While Studies II and III incorporated a clinical perspective in their analyses, they did not examine the associations between input parameters and the investigated disorders. Understanding the relationship between input features and disease status is essential for developing healthcare technologies intended as decision-support tools in clinical assessment [119–122]. To establish such associations, statistical techniques such as non-parametric tests, correlation analysis, and effect size estimation play a critical role [123–125]. These methods allow researchers to identify features that differ significantly between diagnostic groups and to quantify the magnitude and direction of these differences.

In Study IV, biostatistical methods were applied to compare voice features between individuals with COPD and HCs. The dataset included BLA features and MFCCs. Statistical analysis was conducted using the Mann-Whitney U test for group comparisons, PERMANOVA for multivariate analysis, and Cliff's Delta to estimate effect sizes. Key metrics, U-values, p-values, δ-values, and confidence intervals for effect sizes, were used to assess both statistical significance and clinical relevance.

Furthermore, vowel quadrilateral analysis was conducted to assess articulatory differences in vowel production, comparing not only COPD and HC groups but also examining gender differences within each group. The dataset for these analyses was derived from the recordings collected during Study II.

Given that the study involved data from individual participants and addressed clinical implications, it was registered in official trial registries. The project was registered at BTH under the identifier "BTH-6.1.1-0198-2024" and in ClinicalTrials.gov under the identifier "NCT06705647".

Summary of Findings

The four studies together provide complementary insights. Study I showed that ML has been applied to various voice-affecting disorders, with COPD receiving limited attention. Study II addressed this gap by collecting a new COPD voice dataset and demonstrated that ML classifiers, particularly CatBoost, achieved the highest performance in distinguishing COPD from healthy controls. Study III further revealed that segmenting vowel recordings improved model accuracy, indicating that certain parts of the signal contained more discriminative information. Study IV identified 34 statistically and practically significant acoustic features, with shimmer measures and higher-order MFCC derivatives showing the strongest group differences. More specific findings for each study is presented bllow:

Study I

One of the research questions aimed to be answered in Study 1 was about the ML techniques employed in studies aiming to classify or monitor voice-affecting disorders. Figure 3 depicts the decomposition of ML techniques across the included studies where the best precision was achieved. More than 1/3 (51) studies reported SVM as the high-performance model. Artificial neural networks (ANN) followed 39 studies in second place, and Random forest (RF) and K-nearest neighbors (KNN) took third and fourth place with 21 and 13 studies, respectively.

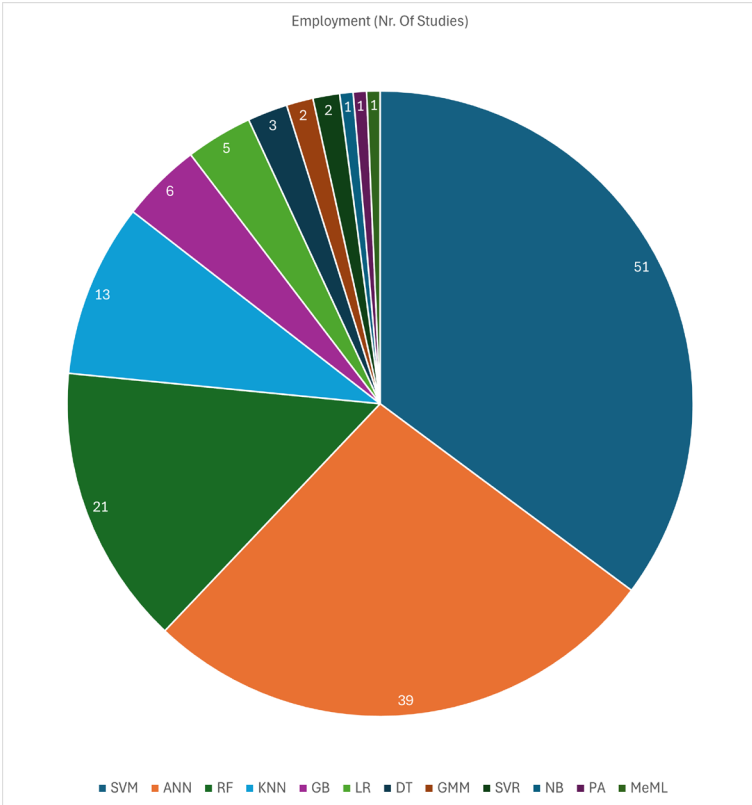


Figure 3. Distribution of employed ML techniques among included studies.

A look at the trends in Figure 4 shows a clear increased trajectory of interest in using ML to analyze voice data for clinical decision support in studies. While the overall trajectory seems to grow at a low speed up to 2020, starting in 2021, the speed of interest in using ML in studies seems to accelerate. An interesting observation is that SVM usage has dominated studies up to 2021, but in 2022, ANN seems more popular than SVM. Additionally, starting in 2017, a spread of using different ML models is observable.

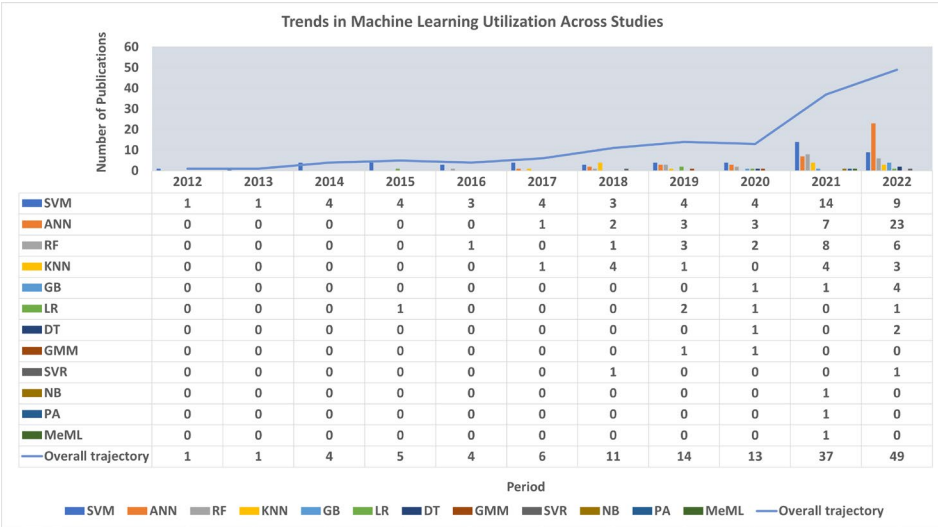


Figure 4. Annual decomposition of ML techniques in studies.

When the focus is shifted to the trajectory of voice-affecting disorders in interest across all included studies in Figure 5, a quite similar trend is observable as in Figure 2. The number of studies on voice-affecting disorders has increased since 2017, and the speed of interest in studies has accelerated after 2020. However, most of the focus in studies seems to be concentrated on Parkinson's disease (PD) during the whole period of interest from 2012 to 2022.

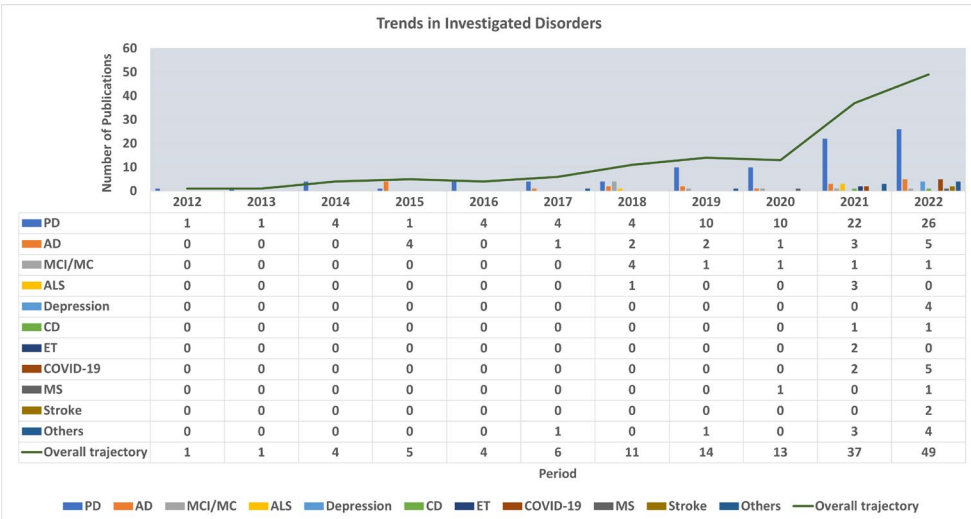


Figure 5. Annual decomposition of considered voice-affecting disorders in studies.

Almost half of the studies (66) combined different feature extraction techniques, such as baseline acoustic (BLA) and MFCC features, together as the input into ML models studies, and 37 studies considered utilizing only BLA features as input.

An analysis of the origin of the publication reveals the countries from which most of the publication comes. In that regard, the most populated countries, China, India, and the USA, dominate the first three places, respectively.

Study II

Based on the results in Study I, Study II addressed and contributed to the underrepresentation of COPD in the scope of using ML and voice for the decision support of COPD. In that regard, Study II contributes to the collection of a COPD classification dataset consisting of features extracted from Swedish utterances of the vowel [a:]. A total of 1246 recordings were collected over six months from 68 participants (30 COPD, 38 HC), recruited through a research clinic at BTH. A comprehensive set of 107 parameters, encompassing MFCC, BLA, and demographic data, were collected from each recording and participant, as described in Table 2.

Table 2. The description of the features used in the study.

Feature Group	Feature	Description
Demographic	Age	Age of participant
	Gender	Biological gender
Health	Cold (Cold/Flu)	Binary answer to the question "Cold/Flu?"
	Pain (Sore throat)	Binary answer to the question: "Sore throat?"
	Slimy (Mucus in throat)	Binary answer to the question: "Mucus in throat?"
	Other	If the participant is not Cold, has no pain, or is slimy.
BLA	Duration	Duration of the voice part
	Mean_F0	Mean fundamental frequency
	Std_F0	The standard deviation of the fundamental frequency
	HNR	Harmonic to Noice Ratio
	Local_Jitter	The average absolute frequency difference between two consecutive periods, divided by the average period
	Absolute_Jitter	The measure of the absolute difference between a clock edge as specified and its observed position
	Rap_Jitter	Relative average perturbation
	PPQ5_Jitter	Five-point period perturbation quotient
	DDP_Jitter	Divided difference between consecutive periods
	Local_Shimmer	The average absolute amplitude difference between two consecutive periods, divided by the average period
	LocaldB_Shimmer	Decibel representation of Local Shimmer
	APQ3_Shimmer	Three-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average
	APQ5_Shimmer	Five-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average

	APQ11_Shimmer	Eleven-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average
	DDA_Shimmer	The average absolute difference between consecutive differences between the amplitudes of consecutive periods
	F1_mean-F4_mean	The mean of the four lowest resonant frequencies of the vocal tract.
	F1_median- F4_median	The median of the four lowest resonant frequencies of the vocal tract.
MFCC	MFCC1_mean-MFCC13_mean	The mean of the 13 Mel Frequency Cepstral Coefficients
	MFCC1_std-MFCC13_std	The standard deviation of the 13 Mel Frequency Cepstral Coefficients
	DMFCC1_mean-DMFCC13_mean	The mean first derivative of the 13 Mel Frequency Cepstral Coefficients
	DMFCC1_std-DMFCC13_std	The standard deviation of the first derivative of the 13 Mel Frequency Cepstral Coefficients
	DDMFCC1_mean-DDMFCC13_mean	The mean second derivative of the 13 Mel Frequency Cepstral Coefficients
	DDMFCC1_std-DDMFCC13_std	The standard deviation of the second derivative of the 13 Mel Frequency Cepstral Coefficients

A subset of data (1058 recordings) was balanced on a gender and age basis and has been subject to tests on three ML models: SVM, RF, and CB. The evaluation of the result demonstrated CB's superiority over other models for training, validation, and test sets with the highest accuracy of 100%, 97%, and 78%, respectively. Table 3 represents the results of all metrics.

Table 3. The highest, average, and standard deviation (STD) scores for each ML model are presented for the training, validation, and test sets. The best-performing values are highlighted in bold.

Model	Precision (%)		Recall (%)		Accuracy (%)		F1_Score (%)	
	AVG/(STD)	MAX	AVG/(STD)	MAX	AVG/(STD)	MAX	AVG/(STD)	MAX
Training set								
CB	**99/(0.5)**	**100**	**99/(0.5)**	**100**	**99/(0.5)**	**100**	**99/(0.5)**	**100**
RF	98/(0.7)	100	100/(0.2)	100	99/(0.5)	100	99/(0.5)	100
SVM	81/(2.0)	100	84/(1.7)	100	86/(1.8)	100	82/(1.9)	100
Validation set								
CB	**95/(1.3)**	**96**	**95/(1.2)**	**97**	**95/(1.3)**	**97**	**95/(1.4)**	**97**
RF	90/(1.6)	93	90/(1.5)	93	90/(1.5)	93	90/(1.5)	93
SVM	74/(2.5)	77	74/(2.4)	77	75/(2.4)	78	74/(2.5)	77
Test set								
CB	**74/(2.8)**	**79**	73/**(2.7)**	**78**	**73/(2.9)**	**78**	**73/(2.29)**	**78**
RF	72/(3.0)	79	71/(3.3)	78	70/(3.4)	77	70/(3.6)	77
SVM	66/(3.1)	70	65/(3.2)	70	65/(3.0)	69	65/(2.8)	69

The results also exhibit promising results from a clinical perspective when the confusion matrix results in Table 4 are considered. Low numbers of false positives

and false negatives revealed CB's superiority versus SVM and RF classifiers from the clinical point of view.

Table 4. The Confusion Matrix results for each ML classifier are provided for both the validation and test sets.

	CB		RF		SVM	
	+	-	+	-	+	-
Validation set						
Positive (+)	101	7	98	10	85	23
Negative (-)	2	73	6	69	20	55
Test set						
Positive (+)	50	25	48	27	45	30
Negative (-)	7	61	6	62	15	53

Figure 6 shows the order of features that contribute to the correct output of the CB classifier, which showed the best performance regarding the discrimination of COPD and HC voices for validation and test sets. In comparison between these two plots associated with two data sets, most of the features remain the most important features, with some shifts in order between the validation and test sets. In the test set, Duration and MFCC4_mean, and in the validation set, MFCC3_mean and MFCC5_mean are not represented among the top ten features of the two plots. However, 8/10 features are still represented in both plots in shifted order. While many of the MFCC features and BLA features seem to play a crucial role in the

prediction, the model appears to be less affected by the health input values since they remain mostly at the end of the feature importance order.

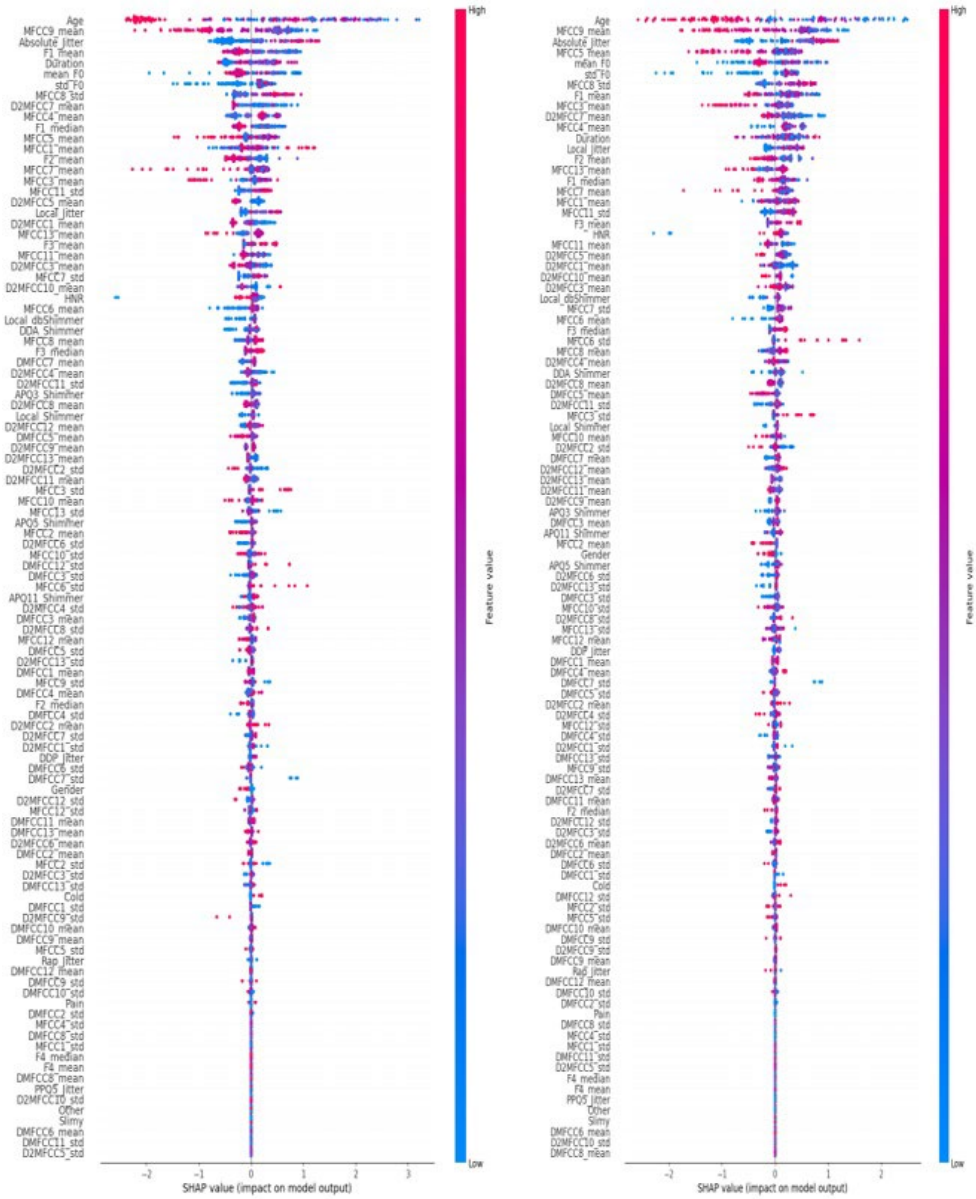


Figure 6. Order of feature importance based on SHAP plots for (a) validation set and (b) test set.

Study III

Study III, which aimed to reveal the deterministic properties of smaller segments in recordings that affect the prediction performance of ML classifiers, indicates that there are more informative parts in the recordings. The features extracted from these time frames exhibit more information, which leads to enhanced performance of ML classifiers in discriminating COPD and HC voices. In Figure 7, the effect of feature segments can be observed for CB, RF, and SVM classifiers. The highest accuracies are achieved for both validation and test sets in four segment group. Another important observation is the increased variation in accuracy metrics for test set results. However, this increase is observed only on ensembling-based classifiers CB and RF, which does not apply to the SVM classifier. An analogy between the full segment and other segments reveals that the increased variation is the effect of segmentation. Furthermore, the differences in accuracy performance between the validation and test sets give an intuition of overfitting. Nevertheless, the difference in size is lower among segmented parts compared to full recording.

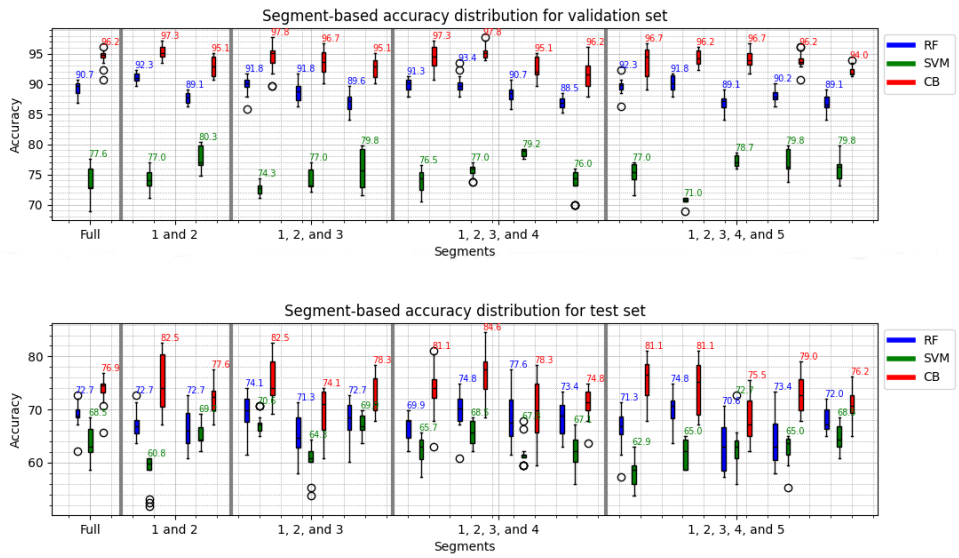


Figure 7. Validation and test set results of each segment decomposition group.

The analysis of group-wise results shown in Figure 8. However, the variation range of accuracy seems to be smaller for the group-wise results than for segment-based results for both variation and test sets. In addition to that, The high accuracy levels of segment-based results are not seen in group-wise results in cooperation between test sets.

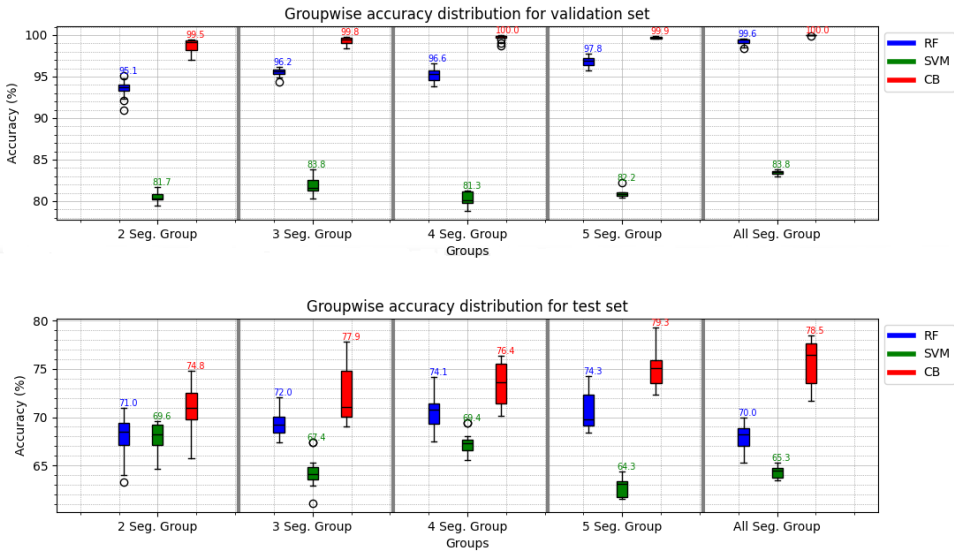


Figure 8. Validation and test results for the data set expanded by merging segment groups.

When the given results are considered from the clinical perspective, the superiority of the CB classifier for all data sets is not negligible. On the other hand, while the validation set seems to benefit from grouping data, the real-world environment represented by the test set gives better results on segments. See Table 5.

Table 5. Confusion matrices showing the best performance on the validation and test datasets.

	CB		RF		SVM	
	+	-	+	-	+	-
Validation set						
Best Segment	All Seg. groups		All Seg. group		3 Seg. group	
Positive (+)	1124	0	1119	5	177	36
Negative (-)	0	1616	5	1616	53	283
Test set						
Best Segment	4 Seg.		Full		2 Seg.	
Positive (+)	60	8	61	7	57	11
Negative (-)	14	61	32	43	33	42

Study IV

Considering the variability in the utterance of the vowel [a:] among participants, as illustrated through vowel quadrilateral plots, no sharp distinction between the COPD and HC groups was evident. Nonetheless, a consistent downward shift of

approximately 100 Hz in both formant 1 (F1) and formant 2 (F2) frequencies was observed in the COPD group, particularly among females. While this trend was visually notable, it was not statistically validated at the formant level.

The statistical evaluation of 101 extracted features showed clear group differences. The Mann–Whitney U test identified 38 features with p-values below 0.05, supporting the alternative hypothesis. Among these, 26 features had p-values under 0.01, and 32 remained under 0.02, indicating a high likelihood of meaningful distinction. After BH correction, 29 features shown in Figure 9 remained significant; D2MFCC8_std retained significance even after the more conservative Bonferroni correction. A multivariate PERMANOVA test applied to the 29 BH-corrected features yielded a significant result ($p = 0.019$), confirming the multivariate separation between groups.

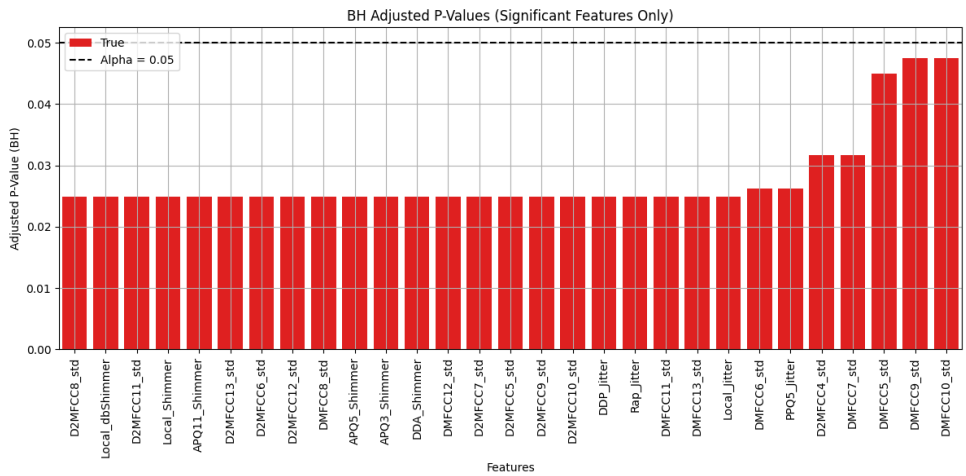


Figure 9. Features remain significant after BH and Bonferroni corrections.

Feature-wise, the most significant differences were found in shimmer-related BLA parameters, which exhibited the highest U-values, followed by several MFCC derivatives. Cliff's Delta analysis revealed that 34 features listed in Table 6 had large effect sizes ($\delta \geq 0.47$), 11 had medium effect sizes, and five non-significant features still demonstrated large effect sizes with wide confidence intervals. Notably, D2MFCC3_mean and HNR exhibited large negative δ values, reflecting systematically higher values in the HC group. Features with larger effect sizes typically showed narrower confidence intervals, supporting their reliability. In contrast, features with negligible effect sizes displayed wide variability and limited discriminative value. These findings indicate that a subset of acoustic features, particularly shimmer measures and higher-order MFCC derivatives, are both statistically and practically meaningful in distinguishing COPD from HC voices.

Table 6. The summary list of 34 statistically and practically meaningful features for discriminating individuals with COPD from HC using voices.

Feature Group (nr. of)	Feature	Description
BLA (12)	HNR	Harmonic to Noise Ratio
	Local_Jitter	The average absolute frequency difference between two consecutive periods, divided by the average period
	Absolute_Jitter	The measure of the absolute difference between a clock edge as specified and its observed position
	Rap_Jitter	Relative average perturbation
	PPQ5_Jitter	Five-point period perturbation quotient
	DDP_Jitter	Divided difference between consecutive periods
	Local_Shimmer	The average absolute amplitude difference between two consecutive periods, divided by the average period
	LocaldB_Shimmer	Decibel representation of Local Shimmer
	APQ3_Shimmer	Three-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average
	APQ5_Shimmer	Five-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average
	APQ11_Shimmer	Eleven-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average
	DDA_Shimmer	The average absolute difference between consecutive differences in the amplitudes of consecutive periods
MFCC (22)	DD2MFCC3_mean	The mean of the second derivative of the 3rd Mel Frequency Cepstral Coefficient
	DMFCC3	The standard deviation of the first derivative of the 3 and 5-13 Mel Frequency Cepstral Coefficients
	DMFCC5_std- DMFCC13_std	
	D2MFCC3_std-D2MFCC13_std	The standard deviation of the second derivative of the 3 and 5-13 Mel Frequency Cepstral Coefficients

Discussion

Discussion on results and contributions

Through four included studies, this thesis contributes to the growing body of knowledge on how voice can benefit as a potential digital biomarker in healthcare by addressing challenges in diagnosing, monitoring, and analyzing voice-affecting conditions, with a specific focus on COPD, by leveraging ML techniques and employing features such as BLA features and MFCCs. The investigation into segmentation strategies has revealed the role of features extracted from the time frames of actual recordings and grouping these recordings to expand the small datasets in improving model performance, emphasizing the importance of selecting appropriate preprocessing methods when dealing with voice data. Additionally, this work has demonstrated the utility of voice data for identifying meaningful patterns that distinguish individuals with COPD from healthy controls.

This thesis builds upon the foundation established by four studies. Study I consolidates the knowledge from prior studies reviewed in the systematic literature review 'Applied machine learning techniques to diagnose voice-affecting conditions and disorders.' While the review identified a diverse range of ML applications for diagnosing voice-affecting conditions and employed various voice features, this thesis extends the knowledge by not only summarizing the state-of-the-art information but also specifically addressing gaps in the research context. Of course, one study cannot provide a comprehensive map of all possible conditions that affect the voice; however, it serves as complementary evidence to studies focused on other conditions, such as dementia, Parkinson's disease, and many others [80,126–129]. Together, these studies build a foundation for future research in this area.

The remaining studies, II, III, and IV, contribute to the field by building upon the information gained through Study I and addressing the gaps identified in the study. In this regard, COPD was identified as a landscape worth exploring, not only due to its effects on voice production but also because of its significant contribution to the global mortality and morbidity burden [24,130,131]. Study II presents the outcomes of using ML to classify COPD and HC voices based on features extracted from the Swedish utterance of the vowel [a:]. It also includes an analysis of feature importance for the best-performing classifier CB, demonstrating the potential of ML technology for decision support in diagnosing COPD. An additional contribution of

Study II is the collection and creation of a dataset for the research community. Participants recorded their voices once per week over a six-month period in self-determined quiet environments, typically at home. This naturalistic data collection strategy enhanced ecological validity by capturing real-world variation in voice production under minimally constrained conditions. The data collection approach aligns with Self-Determination Theory, which highlights the importance of autonomy in sustaining engagement in health-related behaviors [132]. Furthermore, the method reflects principles of Ecological Momentary Assessment (EMA), a framework designed to collect data in real time and natural settings to improve contextual accuracy and reduce recall bias [133,134]. While some variation in background noise and recording quality was expected, recent evidence confirms that acoustic voice features such as jitter, shimmer, and harmonics-to-noise ratio can be reliably extracted even in nonoptimized environments using smartphone microphones [135]. The longitudinal design also enabled exploration of intra-individual variability and potential disease-related voice progression—features essential for future remote monitoring tools. In this context, the collection strategy supports a scalable and patient-centered approach for voice-based diagnostics in chronic disease management [136,137].

This dataset builds on previous research showing changes in the voice of individuals with COPD [12,15], while incorporating new technologies to support the clinical decision-making process in identifying COPD patients, which is aligned with recent efforts of utilizing the benefits of AI to solve healthcare challenges [128,138–142].

The Study III was inspired by the studies in the SLR of Study I, where different VAD and feature extraction techniques were described. However, while these techniques were useful, the dynamic nature of voice signals and the stationary behavior of short time series [143,144], both of which are well-known in the literature, raise questions about how these factors may influence the performance of ML models, particularly when using signal processing methods like segmentation with overlap and windowing techniques [92,110–113,115,145–148]. These techniques are commonly employed in voice processing, yet their effectiveness can vary depending on the characteristics of the voice signal. In that regard, to examine the potential effects of different VAD and data expansion techniques on the performance of ML outcomes, Study III examined the effects of segmentation in the binary classification of COPD and HC voices. This research showed that the features extracted from different segments affect the precision of the outcome, which can be used as a basis in the signal processing and preparation of the data before the feature extraction. These results show alignment with the results from other studies [144,149], which show the effects of preprocessed data on the outcomes of ML techniques.

In order to build reliable healthcare solutions, it is crucial to understand the underlying rationale behind the new technologies [150,151]. Study IV aimed to add meaning to the idea of harvesting voice as a biomarker for a COPD decision support

tool. Based on that, the voice features utilized in studies have undergone tests reflecting the statistical and practical significance of the features extracted from COPD and HC voices. The results pointed out 34 features out of 101, and the 32 measured values are higher for COPD group, while 2 of 34 were higher for HCs. When the statistical results combined with the practical results suggest that the differences observed in the features are not only measurable but also meaningful in a real-world context. These findings provide further evidence supporting the distinct acoustic patterns associated with COPD, even if such patterns are not uniformly apparent in all visual or descriptive analyses of the vowel quadrilateral. These results provide more transparency, reliability, and evidence on the underlying technology, an important aspect of technology acceptance in healthcare [152].

Methodological considerations

The methodological design across the four studies in this thesis reflects an iterative and structured research process grounded in evidence-based practices and adapted to the interdisciplinary domain of applied health technology. All studies followed a quantitative research design, emphasizing reproducibility, structured data analysis, and statistical validation to ensure objective and measurable outcomes across the research process. Each study was shaped not only by prior results but also by practical, clinical, and ethical constraints, resulting in a coherent progression from evidence synthesis to experimental implementation and interpretation.

Study I employed a systematic literature review guided by established protocols [100]and PRISMA guidelines [153], ensuring transparency, reproducibility, and methodological rigor. The exclusion of grey literature, including conference proceedings, dissertations, and non–peer-reviewed articles, was intended to enhance the quality and credibility of the evidence base. Recognizing that this decision might exclude early-stage or domain-specific innovations, a novelty assessment was conducted during the screening process. This involved reviewing the titles and abstracts of excluded sources to identify potentially unique contributions not yet reflected in peer-reviewed literature [154,155]. Additionally, a sensitivity check was performed by randomly removing a subset of included articles to evaluate the stability of key themes and conclusions [156]. These measures suggested that the results of the review were robust and methodologically stable.

Study II translated the findings of the review into practical data collection and model development. The absence of available voice datasets for COPD, identified in Study I, was addressed by constructing a new dataset through mobile-based voice recordings over a six-month period. Feature selection and classifier choices were informed by the literature and clinical context [157,158]. The inclusion of CatBoost complemented the established use of RF and SVM, offering improved handling of

structured data without extensive preprocessing [159]. The evaluation metrics were selected to capture both statistical and clinical relevance.

Study III extended this work by examining the effect of signal segmentation on ML performance. The use of segmentation techniques simulated different feature extraction strategies, such as overlapping and non-overlapping windowing, which are often used in voice signal processing [160–162]. This study emphasized the role of preprocessing decisions in model performance and interpretability. Nested cross-validation was employed to reduce the risk of overfitting while balancing computational cost, particularly given the expanded feature sets generated by segmentation [163,164].

Study IV focused on the statistical interpretation of voice features by comparing groups using non-parametric tests and effect size estimations. These methods provided insights into which features differed significantly between groups and how those differences might be clinically relevant [165,166]. The use of Cliff's Delta and vowel quadrilateral analysis allowed for an articulation-level interpretation of vocal differences, complementing the classification results from Studies II and III [167–169]. SHAP values were applied to enhance the interpretability of feature contributions within ML models, aligning with the increasing demand for transparent and explainable decision support systems in clinical contexts [170,171].

All observational studies, Study II, Study III, and Study IV, were registered through appropriate institutional and clinical trial platforms to ensure procedural transparency and adherence to ethical research standards [172,173]. While the methodological pipeline was carefully designed, limitations remain, such as the language constraint of the recordings, cohort size, and the demographic characteristics of the study population. Despite these, the methodological framework offers a robust and transparent approach to developing voice-based diagnostic tools for clinical use.

Threats to Validity

This thesis acknowledges several potential threats to the validity of its findings. Internal validity may be affected by selection bias if COPD patients and healthy controls differ in uncontrolled ways, such as medication use or lifestyle. Additionally, confounding variables like depression or reflux could influence vocal features independently of COPD.

External validity is another concern, as the study sample is limited to Swedish-speaking individuals living in southern Sweden. This restricts generalizability to broader populations, especially those with different linguistic or cultural backgrounds. Real-world applications may also introduce variability if the model is used in controlled environments or with different recording devices.

Construct validity hinges on whether the chosen vocal features truly capture COPD-related changes. Some features might reflect non-specific voice alterations rather than disease-specific effects. Labeling errors, due to incomplete or imprecise diagnostic information.

Lastly, statistical conclusion validity could be challenged by risks of overfitting, especially in smaller datasets, or by insufficient statistical power to detect meaningful effects. These issues were mitigated through cross-validation, careful model selection, and performance metrics, but they remain important considerations when interpreting results.

Conclusions

This thesis contributes to the evolving field of applied health technology by investigating how voice, when combined with ML, can serve as a digital biomarker to support the assessment of COPD. Set against the backdrop of increasing demands on healthcare systems and the persistent underdiagnosis of COPD, the research explores an interdisciplinary, technology-enabled response to a clinically and socially significant problem.

Through four interlinked studies, the thesis presents a cohesive research trajectory that spans from identifying knowledge gaps in the current literature to developing, evaluating, and interpreting voice-based ML models. Voice, traditionally viewed as a medium for communication, is here reframed as a source of clinically relevant data. The acoustic changes associated with COPD, though subtle, were shown to be quantifiable through signal processing and classifiable using supervised learning models.

The work is grounded in the perspective of AHT, where the focus is not on technological novelty alone but on meaningful integration into healthcare practice. Each study reflects this orientation by addressing practical aspects of data acquisition, algorithm development, and clinical interpretability. Rather than treating voice analysis or ML as isolated technical domains, the research considers them within their real-world application contexts, contributing to person-centered and accessible care models.

The value of this thesis lies not only in the empirical findings but also in its methodological positioning. The studies illustrate how interdisciplinary research can move from problem identification to artifact development and clinical interpretation in a structured, reflective manner. This mirrors the process-oriented logic of design science, where knowledge emerges from iterative experimentation and context-aware evaluation.

Beyond its contributions to COPD assessment, the thesis also positions voice as a broader diagnostic modality that could extend to other systemic or neurological conditions. The findings open the door for further exploration into longitudinal voice monitoring, remote screening, and mobile health integration, offering new avenues for both research and practical deployment. In that regard, this thesis exemplifies how emerging technologies, when thoughtfully applied, can contribute

to the transformation of healthcare. It shows that voice, supported by ML and grounded in interdisciplinary collaboration, can become a powerful tool in enhancing clinical decision-making. The work demonstrates how AHT can translate technical capability into meaningful value for patients, clinicians, and health systems alike.

Ethical Considerations

Since Paper I involved a systematic literature review (SLR) and did not include human participants, ethical considerations were not directly applicable. However, during the eligibility assessment of the studies, individual studies were examined to ensure they had obtained the necessary ethical approvals. This was part of the quality evaluation, where the methodology used in each study was ranked, and special attention was given to ethical practices, ensuring that any voice data collected was done in accordance with ethical guidelines.

Paper II, Paper III, and Paper IV all utilized the same ethical approval granted by the Swedish ethics board in Umeå (DNR: 2020-01045). All studies were conducted in accordance with the Helsinki Declaration, ensuring that the rights and safety of participants were upheld throughout the research process. In each study, written informed consent was obtained from all participants before data collection began. The data was stored on secured servers in compliance with European General Data Protection Regulation (GDPR) guidelines, with access restricted to authorized project staff. Anonymity was maintained for all participants by using unique ID numbers, and written consents were securely stored in a fireproof cabinet. The participants were recruited voluntarily, with the option to withdraw from the study at any time without consequences. Additionally, all participants were insured for any potential damage caused by participation, ensuring the protection of their integrity and minimizing the risk of unauthorized data access.

Future Research

Building on the findings of this thesis, several avenues for future research emerge that can further enhance the clinical utility, technical robustness, and societal relevance of voice-based decision support systems for COPD. A primary direction involves clinical validation in broader and more diverse populations. The current work has demonstrated feasibility within a research-controlled environment and a specific linguistic group. Future studies should expand to larger datasets including speakers of different languages and dialects, and encompass a wider range of demographic and clinical profiles. Longitudinal studies are especially needed to examine how vocal biomarkers evolve over time and whether they can serve as indicators of disease progression, exacerbation risk, or response to therapy.

Another important step involves the integration of voice data with other health-related parameters such as spirometry, oxygen saturation, and patient-reported outcomes. Combining multimodal data could lead to more robust and informative models, strengthening both diagnostic accuracy and clinical decision-making. Alongside this, technical enhancements in feature extraction and modeling should be pursued. While this thesis has focused on widely used feature types like BLA and MFCCs, future work could explore more advanced or task-specific features, including prosodic dynamics or deep spectral representations. End-to-end learning approaches, where models learn directly from raw audio without intermediate handcrafted features, may offer further improvements in generalizability and adaptability.

The translation of these findings into real-world applications will require attention to practical deployment, particularly through mobile and remote-access technologies. Implementing voice-based models in mobile apps or telehealth platforms would enable real-time, low-cost screening and monitoring, particularly useful in under-resourced or rural areas. This development should go hand-in-hand with user-centered design research to ensure that such tools are accessible, interpretable, and meaningful for both patients and clinicians. Ethical and regulatory aspects, such as data privacy, algorithmic transparency, and bias mitigation, must also be addressed to support safe and trustworthy integration into healthcare systems.

Last but not least, future studies may expand beyond COPD to investigate the broader applicability of voice as a biomarker across other respiratory, neurological,

and systemic diseases. Cross-condition generalization, supported by shared datasets and benchmarking initiatives, could pave the way for voice to become a scalable and non-invasive component of digital health infrastructure. By continuing to combine technological advancement with clinical relevance and interdisciplinary insight, this research direction holds strong potential to support earlier detection, better disease management, and more personalized care in the evolving landscape of healthcare.

References

1. Shortell SM, Gillies R, Wu F. United States Innovations in Healthcare Delivery. Public Health Rev. 2010;32: 190–212. doi:10.1007/BF03391598

2. Williams JS, Walker RJ, Egede LE. Achieving Equity in an Evolving Healthcare System: Opportunities and Challenges. The American Journal of the Medical Sciences. 2016;351: 33–43. doi:10.1016/j.amjms.2015.10.012

3. Farmanova E, Baker GR, Cohen D. Combining Integration of Care and a Population Health Approach: A Scoping Review of Redesign Strategies and Interventions, and their Impact. International Journal of Integrated Care. 2019;19. doi:10.5334/ijic.4197

4. Milella F, Minelli EA, Strozzi F, Croce D. Change and Innovation in Healthcare: Findings from Literature . CEOR. 2021;13: 395–408. doi:10.2147/CEOR.S301169

5. Jessup RL, O'Connor DA, Putrik P, Rischin K, Nezon J, Cyril S, et al. Alternative service models for delivery of healthcare services in high-income countries: a scoping review of systematic reviews. BMJ Open. 2019;9: e024385. doi:10.1136/bmjopen-2018-024385

6. Olander E, Nilsson L. Applied Health Technology – a New Research Discipline at Blekinge Institute of Technology. 2009. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:bth-6246

7. Palaniappan K, Lin EYT, Vogel S. Global Regulatory Frameworks for the Use of Artificial Intelligence (AI) in the Healthcare Services Sector. Healthcare. 2024;12: 562. doi:10.3390/healthcare12050562

8. Aldwean A, Tenney D. Artificial Intelligence in Healthcare Sector: A Literature Review of the Adoption Challenges. Open Journal of Business and Management. 2023;12: 129–147. doi:10.4236/ojbm.2024.121009

9. Evangelista EG, Bélisle-Pipon J-C, Naunheim MR, Powell M, Gallois H, Consortium B-V, et al. Voice as a Biomarker in Health-Tech: Mapping the Evolving Landscape of Voice Biomarkers in the Start-Up World.

Otolaryngology–Head and Neck Surgery. 2024;171: 340–352. doi:10.1002/ohn.830

10. Sara JDS, Orbelo D, Maor E, Lerman LO, Lerman A. Guess What We Can Hear—Novel Voice Biomarkers for the Remote Detection of Disease. Mayo Clinic Proceedings. 2023;98: 1353–1375. doi:10.1016/j.mayocp.2023.03.007

11. Fagherazzi G, Fischer A, Ismael M, Despotovic V. Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice. Digit Biomark. 2021;5: 78–88. doi:10.1159/000515346

12. Shastry A, Balasubramanium RK, Acharya PR. Voice Analysis in Individuals with Chronic Obstructive Pulmonary Disease. International Journal of Phonosurgery & Laryngology. 2014;4: 45–49. doi:10.5005/jp-journals-10023-1081

13. Kahnert K, A. Jörres R, Behr J, Welte T. The Diagnosis and Treatment of COPD and Its Comorbidities. Dtsch Arztebl Int. 2023;120: 434–444. doi:10.3238/arztebl.m2023.027

14. Hassan MM, Hussein MT, Emam AM, Rashad UM, Rezk I, Awad AH. Is insufficient pulmonary air support the cause of dysphonia in chronic obstructive pulmonary disease? Auris Nasus Larynx. 2018;45: 807–814. doi:10.1016/j.anl.2017.12.002

15. Mohamed EE, El maghraby RA. Voice changes in patients with chronic obstructive pulmonary disease. Egyptian Journal of Chest Diseases and Tuberculosis. 2014;63: 561–567. doi:10.1016/j.ejcdt.2014.03.006

16. Saeed AM, Riad NM, Osman NM, Khattab AN, Mohammed SE. Study of voice disorders in patients with bronchial asthma and chronic obstructive pulmonary disease. Egypt J Bronchol. 2018;12: 20–26. doi:10.4103/ejb.ejb_34_17

17. Węglarz K, Szczygieł E, Masłoń A, Blaut J. Assessment of breathing patterns and voice of patients with COPD and dysphonia. Respiratory Medicine. 2025;240: 108012. doi:10.1016/j.rmed.2025.108012

18. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med. 2019;25: 954–961. doi:10.1038/s41591-019-0447-x

19. Huang P, Lin CT, Li Y, Tammemagi MC, Brock MV, Atkar-Khattra S, et al. Prediction of lung cancer risk at follow-up screening with low-dose CT: a

training and validation study of a deep learning method. The Lancet Digital Health. 2019;1: e353–e362. doi:10.1016/S2589-7500(19)30159-1

20. Wu C-T, Li G-H, Huang C-T, Cheng Y-C, Chen C-H, Chien J-Y, et al. Acute Exacerbation of a Chronic Obstructive Pulmonary Disease Prediction System Using Wearable Device Data, Machine Learning, and Deep Learning: Development and Cohort Study. JMIR mHealth and uHealth. 2021;9: e22591. doi:10.2196/22591

21. Li M, Cheng K, Ku K, Li J, Hu H, Ung COL. Modelling 30-day hospital readmission after discharge for COPD patients based on electronic health records. npj Prim Care Respir Med. 2023;33: 1–8. doi:10.1038/s41533-023-00339-6

22. Nicholson D, Artz S, Armitage A, Fagan J. Working Relationships and Outcomes in Multidisciplinary Collaborative Practice Settings. Child & Youth Care Forum. 2000;29: 39–73. doi:10.1023/A:1009472223560

23. Interpretation of Global Strategy for the Diagnosis, Treatment, Management and Prevention of Chronic Obstructive Pulmonary Disease 2025 Report. [cited 27 May 2025]. Available: https://www.chinagp.net/EN/10.12114/j.issn.1007-9572.2024.0588

24. Barnes PJ, Shapiro SD, Pauwels RA. Chronic obstructive pulmonary disease: molecular and cellularmechanisms. Eur Respir J. 2003;22: 672–688. doi:10.1183/09031936.03.00040703

25. Singh D, Agusti A, Anzueto A, Barnes PJ, Bourbeau J, Celli BR, et al. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease: the GOLD science committee report 2019. Eur Respir J. 2019;53: 1900164. doi:10.1183/13993003.00164-2019

26. Celli BR, MacNee W, Agusti A, Anzueto A, Berg B, Buist AS, et al. Standards for the diagnosis and treatment of patients with COPD: a summary of the ATS/ERS position paper. European Respiratory Journal. 2004;23: 932–946. doi:10.1183/09031936.04.00014304

27. Agustí AGN, Noguera A, Sauleda J, Sala E, Pons J, Busquets X. Systemic effects of chronic obstructive pulmonary disease. Eur Respir J. 2003;21: 347–360. doi:10.1183/09031936.03.00405703

28. Mannino DM, Thorn D, Swensen A, Holguin F. Prevalence and outcomes of diabetes, hypertension and cardiovascular disease in COPD. European Respiratory Journal. 2008;32: 962–969. doi:10.1183/09031936.00012408

29. Chronic obstructive pulmonary disease (COPD). [cited 13 May 2025]. Available: https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)

30. Zhou J, Li X, Wang X, Yu N, Wang W. Accuracy of portable spirometers in the diagnosis of chronic obstructive pulmonary disease A meta-analysis. npj Prim Care Respir Med. 2022;32: 1–12. doi:10.1038/s41533-022-00275-x

31. Amaza IP, O'Shea ,Amy M J, Fortis ,Spyridon, and Comellas AP. Discordant Quantitative and Visual CT Assessments in the Diagnosis of Emphysema. International Journal of Chronic Obstructive Pulmonary Disease. 2021;16: 1231–1242. doi:10.2147/COPD.S284477

32. Kumar S, Bhagat V, Sahu P, Chaube MK, Behera AK, Guizani M, et al. A novel multimodal framework for early diagnosis and classification of COPD based on CT scan images and multivariate pulmonary respiratory diseases. Computer Methods and Programs in Biomedicine. 2024;243: 107911. doi:10.1016/j.cmpb.2023.107911

33. Axelsson M, Backman H, Nwaru BI, Stridsman C, Vanfleteren L, Hedman L, et al. Underdiagnosis and misclassification of COPD in Sweden – A Nordic Epilung study. Respiratory Medicine. 2023;217. doi:10.1016/j.rmed.2023.107347

34. O'Donnell DE, Gebke KB. Activity restriction in mild COPD: a challenging clinical problem. COPD. 2014;9: 577–588. doi:10.2147/COPD.S62766

35. Labaki WW, Han MK. Improving Detection of Early Chronic Obstructive Pulmonary Disease. Annals ATS. 2018;15: S243–S248. doi:10.1513/AnnalsATS.201808-529MG

36. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Medical Education. 2023;23: 689. doi:10.1186/s12909-023-04698-z

37. Alam MZ, Simonetti A, Brillantino R, Tayler N, Grainge C, Siribaddana P, et al. Predicting Pulmonary Function From the Analysis of Voice: A Machine Learning Approach. Front Digit Health. 2022;4. doi:10.3389/fdgth.2022.750226

38. Mayr W, Triantafyllopoulos ,Andreas, Batliner ,Anton, Schuller ,Björn W, and Berghaus TM. Assessing the Clinical and Functional Status of COPD Patients Using Speech Analysis During and After Exacerbation. International Journal of

Chronic Obstructive Pulmonary Disease. 2025;20: 137–147. doi:10.2147/COPD.S480842

39. Bijoy MH, Mondal PK, Plabon MF, Das D. Advanced Machine Learning Techniques for Accurate Diagnosis of Respiratory Diseases Using Vocal Biomarkers. 2024 2nd International Conference on Information and Communication Technology (ICICT). 2024. pp. 95–99. doi:10.1109/ICICT64387.2024.10839668

40. Varmaghani M, Dehghani M, Heidari E, Sharifi F, Saeedi Moghaddam S, Farzadfar F. Global prevalence of chronic obstructive pulmonary disease: systematic review and meta-analysis. East Mediterr Health J. 2019;25: 47–57. doi:10.26719/emhj.18.014

41. Merkus J, Hubers F, Cucchiarini C, Strik H. Digital Eavesdropper – Acoustic Speech Characteristics as Markers of Exacerbations in COPD Patients. : 10.

42. Idrisoglu A, Dallora AL, Anderberg P, Sanmartin Berglund J. Applied machine learning techniques to diagnose voice-affecting conditions and disorders: A systematic literature review. JMIR Medical Informatics. 2023; 18.

43. Ho T, Cusack RP, Chaudhary N, Satia I, Kurmi OP. Under- and over-diagnosis of COPD: a global perspective. Breathe. 2019;15: 24–35. doi:10.1183/20734735.0346-2018

44. Di Marco F, Balbo P, De Blasio F, Cardaci V, Crimi N, Girbino G, et al. Early management of COPD: where are we now and where do we go from here? A Delphi consensus project. COPD. 2019;Volume 14: 353–360. doi:10.2147/COPD.S176662

45. Robin J, Harrison JE, Kaufman LD, Rudzicz F, Simpson W, Yancheva M. Evaluation of Speech-Based Digital Biomarkers: Review and Recommendations. Digit Biomark. 2020;4: 99–108. doi:10.1159/000510820

46. Syed SA, Rashid M, Hussain S. Meta-analysis of voice disorders databases and applied machine learning techniques. Math Biosci Eng. 2020;17: 7958–7979. doi:10.3934/mbe.2020404

47. Prabakaran D, Shyamala R. A Review On Performance Of Voice Feature Extraction Techniques. 2019 3rd International Conference on Computing and Communications Technologies (ICCCT). 2019. pp. 221–231. doi:10.1109/ICCCT2.2019.8824988

48. V A, Reddy RVS. Classification of voice pathology using different features and Bi-LSTM. 2023 International Conference on Smart Systems for applications in

Electrical Sciences (ICSSES). 2023. pp. 1–4. doi:10.1109/ICSSES58299.2023.10200529

49. Teixeira Fernandes JF, Freitas D, Teixeira JP. Voice Pathologies : The Most Comum Features and Classification Tools. 2021 16th Iberian Conference on Information Systems and Technologies (CISTI). Chaves, Portugal: IEEE; 2021. pp. 1–6. doi:10.23919/CISTI52073.2021.9476466

50. Tsanas A, Little MA, McSharry PE, Spielman J, Ramig LO. Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease. IEEE Transactions on Biomedical Engineering. 2012;59: 1264–1271. doi:10.1109/TBME.2012.2183367

51. Rogers HP, Hseu A, Kim J, Silberholz E, Jo S, Dorste A, et al. Voice as a Biomarker of Pediatric Health: A Scoping Review. Children. 2024;11: 684. doi:10.3390/children11060684

52. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics. 1943;5: 115–133. doi:10.1007/BF02478259

53. Samuel AL. Some Studies in Machine Learning Using the Game of Checkers. 1959.

54. Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review. 1958;65: 386–408. doi:10.1037/h0042519

55. Mahesh B. Machine Learning Algorithms - A Review. International Journal of Science and Research. 2018;9. doi:10.21275/ART20203995

56. Flach P. Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press; 2012.

57. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nature Medicine. 2019;25. doi:10.1038/s41591-018-0316-z

58. Flach P. Performance Evaluation in Machine Learning: The Good, the Bad, the Ugly, and the Way Forward. Proceedings of the AAAI Conference on Artificial Intelligence. 2019;33: 9808–9814. doi:10.1609/aaai.v33i01.33019808

59. Artrith N. Best practices in machine learning for chemistry. Nature Chemistry. 2021;13.

60. Li J, Cheng H, Guo H, Qiu S. Survey on Artificial Intelligence for Vehicles. Automot Innov. 2018;1: 2–14. doi:10.1007/s42154-018-0009-9

61. Peng C-Y, Chen R-C. Voice recognition by Google Home and Raspberry Pi for smart socket control. 2018 Tenth International Conference on Advanced Computational Intelligence (ICACI). 2018. pp. 324–329. doi:10.1109/ICACI.2018.8377477

62. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2. doi:10.1136/svn-2017-000101

63. Dallora AL, Berglund JS, Brogren M, Kvist O, Ruiz SD, Dübbel A, et al. Age Assessment of Youth and Young Adults Using Magnetic Resonance Imaging of the Knee: A Deep Learning Approach. JMIR Medical Informatics. 2019;7: e16291. doi:10.2196/16291

64. Al Kuwaiti A, Nazer K, Al-Reedy A, Al-Shehri S, Al-Muhanna A, Subbarayalu AV, et al. A Review of the Role of Artificial Intelligence in Healthcare. Journal of Personalized Medicine. 2023;13: 951. doi:10.3390/jpm13060951

65. Perveen S, Shahbaz M, Ansari MS, Keshavjee K, Guergachi A. A Hybrid Approach for Modeling Type 2 Diabetes Mellitus Progression. Front Genet. 2020;10. doi:10.3389/fgene.2019.01076

66. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. The Lancet Oncology. 2019;20: e262–e273. doi:10.1016/S1470-2045(19)30149-4

67. An Q, Rahman S, Zhou J, Kang JJ. A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges. Sensors. 2023;23: 4178. doi:10.3390/s23094178

68. McCoy LG, Banja JD, Ghassemi M, Celi LA. Ensuring machine learning for healthcare works for all. BMJ Health Care Inform. 2020;27. doi:10.1136/bmjhci-2020-100237

69. Elvas LB, Almeida A, Ferreira JC. The Role of AI in Cardiovascular Event Monitoring and Early Detection: Scoping Literature Review. JMIR Medical Informatics. 2025;13: e64349. doi:10.2196/64349

70. Álvarez JD, Matias-Guiu JA, Cabrera-Martín MN, Risco-Martín JL, Ayala JL. An application of machine learning with feature selection to improve diagnosis and classification of neurodegenerative disorders. BMC Bioinformatics. 2019;20: 491. doi:10.1186/s12859-019-3027-7

71. Pao JJ, Biggs M, Duncan D, Lin DI, Davis R, Huang RSP, et al. Predicting EGFR mutational status from pathology images using a real-world dataset. Sci Rep. 2023;13: 4404. doi:10.1038/s41598-023-31284-6

72. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Medical Research Methodology. 2019;19: 64. doi:10.1186/s12874-019-0681-4

73. Hegde S, Shetty S, Rai S, Dodderi T. A Survey on Machine Learning Approaches for Automatic Detection of Voice Disorders. Journal of Voice. 2019;33: 947.e11-947.e33. doi:10.1016/j.jvoice.2018.07.014

74. Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S. Human Decisions and Machine Predictions*. The Quarterly Journal of Economics. 2018;133: 237–293. doi:10.1093/qje/qjx032

75. Purwins H, Li B, Virtanen T, Schlüter J, Chang S-Y, Sainath T. Deep Learning for Audio Signal Processing. IEEE Journal of Selected Topics in Signal Processing. 2019;13: 206–219. doi:10.1109/JSTSP.2019.2908700

76. Roy N, Barkmeier-Kraemer J, Eadie T, Preeti Sivasankar M, Mehta D, Paul D, et al. Evidence-Based Clinical Voice Assessment: A Systematic Review. AM J SPEECH LANG PATHOL. 2013;22: 212–226. doi:10.1044/1058-0360(2012/12-0014)

77. Verdolini K, Rosen CA, Branski RC, editors. Classification Manual for Voice Disorders–I. New York: Psychology Press; 2005. doi:10.4324/9781410617293

78. Bachorowski J-A. Vocal Expression and Perception of Emotion. Current Directions in Psychological Science. 1999 [cited 27 May 2025]. Available: https://journals.sagepub.com/doi/10.1111/1467-8721.00013

79. McCabe DJ. Prosody: An Overview and Applications to Voice Therapy. GJO. 2017;7. doi:10.19080/GJO.2017.07.555719

80. Sindhu I, Sainin MS. Automatic Speech and Voice Disorder Detection Using Deep Learning—A Systematic Literature Review. IEEE Access. 2024;12: 49667–49681. doi:10.1109/ACCESS.2024.3371713

81. Barlow J, Sragi Z, Rivera-Rivera G, Al-Awady A, Daşdöğen Ü, Courey MS, et al. The Use of Deep Learning Software in the Detection of Voice Disorders: A Systematic Review. Otolaryngology–Head and Neck Surgery. 2024;170: 1531–1543. doi:10.1002/ohn.636

82. Luo J, Wu Y, Liu M, Li Z, Wang Z, Zheng Y, et al. Differentiation between depression and bipolar disorder in child and adolescents by voice features. Child and Adolescent Psychiatry and Mental Health. 2024;18: 19. doi:10.1186/s13034-024-00708-0

83. Larsen E, Murton O, Song X, Joachim D, Watts D, Kapczinski F, et al. Validating the efficacy and value proposition of mental fitness vocal biomarkers in a psychiatric population: prospective cohort study. Front Psychiatry. 2024;15. doi:10.3389/fpsyt.2024.1342835

84. Albuquerque L, Valente ARS, Teixeira A, Figueiredo D, Sa-Couto P, Oliveira C. Association between acoustic speech features and non-severe levels of anxiety and depression symptoms across lifespan. PLOS ONE. 2021;16: e0248842. doi:10.1371/journal.pone.0248842

85. Myers BR, Lense MD, Gordon RL. Pushing the Envelope: Developments in Neural Entrainment to Speech and the Biological Underpinnings of Prosody Perception. Brain Sciences. 2019;9: 70. doi:10.3390/brainsci9030070

86. Kieling MLM, Finkelsztejn A, Konzen VR, dos Santos VB, Ayres A, Klein I, et al. Articulatory speech measures can be related to the severity of multiple sclerosis. Front Neurol. 2023;14. doi:10.3389/fneur.2023.1075736

87. Ueha R, Miura C, Matsumoto N, Sato T, Goto T, Kondo K. Vocal Fold Motion Impairment in Neurodegenerative Diseases. Journal of Clinical Medicine. 2024;13: 2507. doi:10.3390/jcm13092507

88. Verde L, De Pietro G, Sannino G. Voice Disorder Identification by Using Machine Learning Techniques. IEEE Access. 2018;6: 16246–16255. doi:10.1109/ACCESS.2018.2816338

89. Chen Z, Liang N, Li H, Zhang H, Li H, Yan L, et al. Exploring explainable AI features in the vocal biomarkers of lung disease. Computers in Biology and Medicine. 2024;179: 108844. doi:10.1016/j.compbiomed.2024.108844

90. Teixeira JP, Oliveira C, Lopes C. Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters. Procedia Technology. 2013;9: 1112–1122. doi:10.1016/j.protcy.2013.12.124

91. Maryn Y, Roy N, De Bodt M, Van Cauwenberge P, Corthals P. Acoustic measurement of overall voice quality: A meta-analysis. J Acoust Soc Am. 2009;126: 2619–2634. doi:10.1121/1.3224706

92. Braga D, Madureira AM, Coelho L, Ajith R. Automatic detection of Parkinson's disease based on acoustic analysis of speech. Engineering

Applications of Artificial Intelligence. 2019;77: 148–158. doi:10.1016/j.engappai.2018.09.018

93. Rabiner L, Schafer R. Theory and Applications of Digital Speech Processing. 1st ed. USA: Prentice Hall Press; 2010.

94. Despotovic V, Elbéji A, Fünfgeld K, Pizzimenti M, Ayadi H, Nazarov PV, et al. Digital voice-based biomarker for monitoring respiratory quality of life: findings from the colive voice study. Biomedical Signal Processing and Control. 2024;96: 106555. doi:10.1016/j.bspc.2024.106555

95. MacCallum JK, Cai L, Zhou L, Zhang Y, Jiang JJ. Acoustic Analysis of Aperiodic Voice: Perturbation and Nonlinear Dynamic Properties in Esophageal Phonation. Journal of Voice. 2009;23: 283–290. doi:10.1016/j.jvoice.2007.10.004

96. Cummins N, Amiriparian S, Hagerer G, Batliner A, Steidl S, Schuller BW. An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech. Proceedings of the 25th ACM international conference on Multimedia. New York, NY, USA: Association for Computing Machinery; 2017. pp. 478–484. doi:10.1145/3123266.3123371

97. The top 10 causes of death. [cited 18 Dec 2024]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

98. Statistics on Causes of Death 2023.

99. Adeloye D, Song P, Zhu Y, Campbell H, Sheikh A, Rudan I. Global, regional, and national prevalence of, and risk factors for, chronic obstructive pulmonary disease (COPD) in 2019: a systematic review and modelling analysis. The Lancet Respiratory Medicine. 2022;10: 447–458. doi:10.1016/S2213-2600(21)00511-7

100. Kitchenham B, Charters S. Guidelines for performing Systematic Literature Reviews in Software Engineering. 2007;2.

101. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Goslashtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Journal of Clinical Epidemiology. 2009; 34.

102. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. Journal of Big Data. 2020;7: 94. doi:10.1186/s40537-020-00369-8

103. Huang G, Wu L, Ma X, Zhang W, Fan J, Yu X, et al. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. Journal of Hydrology. 2019;574: 1029–1041. doi:10.1016/j.jhydrol.2019.04.085

104. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems. 2018;31. Available: https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html

105. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2017. pp. 4768–4777.

106. Shen M, Mortezaagha P, Rahgozar A. Explainable artificial intelligence to diagnose early Parkinson's disease via voice analysis. Sci Rep. 2025;15: 11687. doi:10.1038/s41598-025-96575-6

107. Antwarg L, Miller RM, Shapira B, Rokach L. Explaining anomalies detected by autoencoders using Shapley Additive Explanations. Expert Systems with Applications. 2021;186: 115736. doi:10.1016/j.eswa.2021.115736

108. Haneesha Samudrala SS, Thambi J, Vadluri SR, Mahalingam A, Pati PB. Enhancing Parkinson's Disease Diagnosis using Speech Analysis:A Feature Subset Selection Approach with LIME and SHAP. 2024 3rd International Conference for Innovation in Technology (INOCON). 2024. pp. 1–5. doi:10.1109/INOCON60754.2024.10511805

109. Zhang Z. Mechanics of human voice production and control. The Journal of the Acoustical Society of America. 2016;140: 2614–2635. doi:10.1121/1.4964509

110. Zhang L, Qu Y, Jin B, Jing L, Gao Z, Liang Z. An intelligent mobile-enabled system for diagnosing parkinson disease: Development and validation of a speech impairment detection system. JMIR Medical Informatics. 2020;8. doi:10.2196/18689

111. Chun KS, Nathan V, Vatanparvar K, Nemati E, Rahman MM, Blackstock E, et al. Towards Passive Assessment of Pulmonary Function from Natural Speech Recorded Using a Mobile Phone. 2020 IEEE International Conference on Pervasive Computing and Communications (PerCom). 2020. pp. 1–10. doi:10.1109/PerCom45495.2020.9127380

112. Farrús M, Codina-Filbà J, Reixach E, Andrés E, Sans M, Garcia N, et al. Speech-Based Support System to Supervise Chronic Obstructive Pulmonary Disease Patient Status. Applied Sciences. 2021;11: 7999. doi:10.3390/app11177999

113. Soumaya Z, Taoufiq BD, Benayad N, Achraf B, Ammoumou A. A Hybrid Method for the Diagnosis and Classifying Parkinson's Patients based on Time–frequency Domain Properties and K-nearest Neighbor. J Med Signals Sens. 2020;10: 60–66. doi:10.4103/jmss.JMSS_61_18

114. Nathan V, Vatanparvar K, Rahman MM, Nemati E, Kuang J. Assessment of Chronic Pulmonary Disease Patients Using Biomarkers from Natural Speech Recorded by Mobile Devices. 2019 IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN). 2019. pp. 1–4. doi:10.1109/BSN.2019.8771043

115. Nathan V, Rahman MM, Vatanparvar K, Nemati E, Blackstock E, Kuang J. Extraction of Voice Parameters from Continuous Running Speech for Pulmonary Disease Monitoring. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2019. pp. 859–864. doi:10.1109/BIBM47256.2019.8983115

116. Kent RD, Rountrey C. What Acoustic Studies Tell Us About Vowels in Developing and Disordered Speech. Am J Speech Lang Pathol. 2020;29: 1749–1778. doi:10.1044/2020_AJSLP-19-00178

117. Fujisaki H. Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing. In: MacNeilage PF, editor. The Production of Speech. New York, NY: Springer; 1983. pp. 39–55. doi:10.1007/978-1-4613-8202-7_3

118. Martono NP, Ohwada H. Evaluating the Impact of Windowing Techniques on Fourier Transform-Preprocessed Signals for Deep Learning-Based ECG Classification. Hearts. 2024;5: 501–515. doi:10.3390/hearts5040037

119. Ding H, Mandapati A, Karjadi C, Ang TFA, Lu S, Miao X, et al. Association Between Acoustic Features and Neuropsychological Test Performance in the Framingham Heart Study: Observational Study. Journal of Medical Internet Research. 2022;24: e42886. doi:10.2196/42886

120. Asan O, Bayrak AE, Choudhury A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. Journal of Medical Internet Research. 2020;22: e15154. doi:10.2196/15154

121. Hartskamp M van, Consoli S, Verhaegh W, Petkovic M, Stolpe A van de. Artificial Intelligence in Clinical Health Care Applications: Viewpoint. Interactive Journal of Medical Research. 2019;8: e12100. doi:10.2196/12100

122. Alonso AKM, Hirt J, Woelfle T, Janiaud P, Hemkens LG. Definitions of digital biomarkers: a systematic mapping of the biomedical literature. BMJ Health Care Inform. 2024;31. doi:10.1136/bmjhci-2023-100914

123. Ferguson CJ. An effect size primer: A guide for clinicians and researchers. Washington, DC, US: American Psychological Association; 2016. p. 310. doi:10.1037/14805-020

124. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biological Reviews. 2007;82: 591–605. doi:10.1111/j.1469-185X.2007.00027.x

125. Kim H-Y. Statistical notes for clinical researchers: effect size. Restor Dent Endod. 2015;40: 328. doi:10.5395/rde.2015.40.4.328

126. Javeed A, Dallora AL, Berglund JS, Ali A, Ali L, Anderberg P. Machine Learning for Dementia Prediction: A Systematic Review and Future Research Directions. J Med Syst. 2023;47: 17. doi:10.1007/s10916-023-01906-7

127. Khatwad R, Tiwari S, Tripathi Y, Nehra A, Sharma A. Parkinson's Disease Detection Using Voice and Speech—Systematic Literature Review. Bio-Inspired Optimization for Medical Data Mining. John Wiley & Sons, Ltd; 2024. pp. 41–74. doi:10.1002/9781394214211.ch3

128. Brahmi Z, Mahyoob M, Al-Sarem M, Algaraady J, Bousselmi K, Alblwi A. Exploring the Role of Machine Learning in Diagnosing and Treating Speech Disorders: A Systematic Literature Review. Psychology Research and Behavior Management. 2024;17: 2205–2232. doi:10.2147/PRBM.S460283

129. Shrivastava P, Tripathi N, Dewangan BK, Singh BK, Choudhury T, Kotecha K, et al. Autonomic Computing Based Respiratory Disorders Assessment Using Speech Parameters: A Systematic Review. 2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). 2023. pp. 1–6. doi:10.1109/ISMSIT58785.2023.10304972

130. Boers E, Barrett M, Su JG, Benjafield AV, Sinha S, Kaye L, et al. Global Burden of Chronic Obstructive Pulmonary Disease Through 2050.

JAMA Network Open. 2023;6: e2346598. doi:10.1001/jamanetworkopen.2023.46598

131. Pauwels RA, Buist AS, Calverley PMA, Jenkins CR, Hurd SS. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease. Am J Respir Crit Care Med. 2001;163: 1256–1276. doi:10.1164/ajrccm.163.5.2101039

132. Deci EL, and Ryan RM. The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior. Psychological Inquiry. 2000;11: 227–268. doi:10.1207/S15327965PLI1104_01

133. Shiffman S, Stone AA, Hufford MR. Ecological Momentary Assessment. Annu Rev Clin Psychol. 2008;4: 1–32. doi:10.1146/annurev.clinpsy.3.022806.091415

134. Trull TJ, Ebner-Priemer U. Ambulatory Assessment. Annual Review of Clinical Psychology. 2013;9: 151–176. doi:10.1146/annurev-clinpsy-050212-185510

135. van der Woerd B, Wu M, Parsa V, Doyle PC, Fung K. Evaluation of Acoustic Analyses of Voice in Nonoptimized Conditions. Journal of Speech, Language, and Hearing Research. 2020;63: 3991–3999. doi:10.1044/2020_JSLHR-20-00212

136. Sridhar Rao Muthineni. AI in Mobile Health Apps: Transforming Chronic Disease Management. Int J Sci Res Comput Sci Eng Inf Technol. 2025;11: 108–116. doi:10.32628/CSEIT25111212

137. Onnela J-P, Rauch SL. Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health. Neuropsychopharmacol. 2016;41: 1691–1696. doi:10.1038/npp.2016.7

138. Lee KH, Choi GH, Yun J, Choi J, Goh MJ, Sinn DH, et al. Machine learning-based clinical decision support system for treatment recommendation and overall survival prediction of hepatocellular carcinoma: a multi-center study. npj Digit Med. 2024;7: 1–8. doi:10.1038/s41746-023-00976-8

139. Ramírez DJGC, Islam MM, Even AIH. Machine Learning Applications in Healthcare: Current Trends and Future Prospects. Journal of Artificial Intelligence General science (JAIGS) ISSN:3006-4023. 2024;1. doi:10.60087/jaigs.v1i1.33

140. Levy AS, Bhatia S, Merenzon MA, Andryski AL, Rivera CA, Daggubati LC, et al. Exploring the Landscape of Machine Learning

Applications in Neurosurgery: A Bibliometric Analysis and Narrative Review of Trends and Future Directions. World Neurosurgery. 2024;181: 108–115. doi:10.1016/j.wneu.2023.10.042

141. Khalifa M, Albadawy M, Iqbal U. Advancing clinical decision support: The role of artificial intelligence across six domains. Computer Methods and Programs in Biomedicine Update. 2024;5: 100142. doi:10.1016/j.cmpbup.2024.100142

142. Gomez-Cabello CA, Borna S, Pressman S, Haider SA, Haider CR, Forte AJ. Artificial-Intelligence-Based Clinical Decision Support Systems in Primary Care: A Scoping Review of Current Clinical Implementations. European Journal of Investigation in Health, Psychology and Education. 2024;14: 685–698. doi:10.3390/ejihpe14030045

143. Vieira VJD, Costa SC, Correia SEN. Non-Stationarity-Based Adaptive Segmentation Applied to Voice Disorder Discrimination. IEEE Access. 2023;11: 54750–54759. doi:10.1109/ACCESS.2023.3281191

144. Antsiperov VE, Morozov VA, Nikitov SA. Isolated-word segmentation based on the dynamics of the parameters of short correlation functions. J Commun Technol Electron. 2006;51: 1356–1368. doi:10.1134/S1064226906120059

145. Ali L, Zhu C, Zhang Z, Liu Y. Automated Detection of Parkinson's Disease Based on Multiple Types of Sustained Phonations Using Linear Discriminant Analysis and Genetically Optimized Neural Network. IEEE Journal of Translational Engineering in Health and Medicine. 2019;7. doi:10.1109/JTEHM.2019.2940900

146. Pramanik M, Pradhan R, Nandy P, Qaisar SM, Bhoi AK. Assessment of Acoustic Features and Machine Learning for Parkinson's Detection. Ahmad SA, editor. Journal of Healthcare Engineering. 2021;2021: 1–13. doi:10.1155/2021/9957132

147. Tsanas A, Little MA, McSharry PE, Ramig LO. Accurate Telemonitoring of Parkinson's Disease Progression by Noninvasive Speech Tests. IEEE Trans Biomed Eng. 2010;57: 884–893. doi:10.1109/TBME.2009.2036000

148. Nathan V, Vatanparvar K, Rahman MM, Nemati E, Kuang J. Assessment of Chronic Pulmonary Disease Patients Using Biomarkers from Natural Speech Recorded by Mobile Devices. 2019 IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN). 2019. pp. 1–4. doi:10.1109/BSN.2019.8771043

149. Bakır H, Çayır AN, Navruz TS. A comprehensive experimental study for analyzing the effects of data augmentation techniques on voice classification. Multimed Tools Appl. 2024;83: 17601–17628. doi:10.1007/s11042-023-16200-4

150. Hyysalo S. Health Technology Development and Use: From Practice-Bound Imagination to Evolving Impacts. New York: Routledge; 2010. doi:10.4324/9780203849156

151. Yi MY, Jackson JD, Park JS, Probst JC. Understanding information technology acceptance by individual professionals: Toward an integrative view. Information & Management. 2006;43: 350–363. doi:10.1016/j.im.2005.08.006

152. AlQudah AA, Al-Emran M, Shaalan K. Technology Acceptance in Healthcare: A Systematic Review. Applied Sciences. 2021;11: 10537. doi:10.3390/app112210537

153. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. International Journal of Surgery. 2010;8: 336–341. doi:10.1016/j.ijsu.2010.02.007

154. Paez A. Gray literature: An important resource in systematic reviews. Journal of Evidence-Based Medicine. 2017;10: 233–240. doi:10.1111/jebm.12266

155. Scherer RW, Saldanha IJ. How should systematic reviewers handle conference abstracts? A view from the trenches. Syst Rev. 2019;8: 264. doi:10.1186/s13643-019-1188-0

156. Gough D, Thomas J, Oliver S. An Introduction to Systematic Reviews. 2017; 1–352.

157. Drugman T, Kane J, Gobl C. Data-driven detection and analysis of the patterns of creaky voice. Computer Speech & Language. 2014;28: 1233–1253. doi:10.1016/j.csl.2014.03.002

158. Kapetanidis P, Kalioras F, Tsakonas C, Tzamalis P, Kontogiannis G, Karamanidou T, et al. Respiratory Diseases Diagnosis Using Audio Analysis and Artificial Intelligence: A Systematic Review. Sensors. 2024;24: 1173. doi:10.3390/s24041173

159. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2018. Available:

https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html

160. Kelly AC, Gobl C. The Effects of Windowing on the Calculation of MFCCs for Different Types of Speech Sounds. Advances in Nonlinear Speech Processing. Springer, Berlin, Heidelberg; 2011. pp. 111–118. doi:10.1007/978-3-642-25020-0_15

161. Tirronen S, Kadiri SR, Alku P. The Effect of the MFCC Frame Length in Automatic Voice Pathology Detection. Journal of Voice. 2024;38: 975–982. doi:10.1016/j.jvoice.2022.03.021

162. Madanian S, Chen T, Adeleye O, Templeton JM, Poellabauer C, Parry D, et al. Speech emotion recognition using machine learning — A systematic review. Intelligent Systems with Applications. 2023;20: 200266. doi:10.1016/j.iswa.2023.200266

163. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7: 91. doi:10.1186/1471-2105-7-91

164. Cawley GC, Talbot NLC. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. The Journal of Machine Learning Research. 2010;11: 2079–2107. doi:10.5555/1756006.1859921

165. Mann HB, Whitney DR. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics. 1947;18: 50–60.

166. Cliff N. Dominance statistics: Ordinal analyses to answer ordinal questions. Psychological Bulletin. 1993;114: 494–509. doi:10.1037/0033-2909.114.3.494

167. Persson A. The acoustic characteristics of Swedish vowels. Phonetica. 2024 [cited 29 Nov 2024]. doi:10.1515/phon-2024-0011

168. Meissel K, Yao ES. Using Cliff's Delta as a Non-Parametric Effect Size Measure: An Accessible Web App and R Tutorial. Practical Assessment, Research, and Evaluation. 2024;29. doi:10.7275/pare.1977

169. Macbeth G, Razumiejczyk E, Ledesma RD. Cliff's Delta Calculator: A non-parametric effect size program for two groups of observations. Universitas Psychologica. 2011;10: 545–555.

170. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2017. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

171. Tjoa E, Guan C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Transactions on Neural Networks and Learning Systems. 2021;32: 4793–4813. doi:10.1109/TNNLS.2020.3027314

172. Garg R, Maurya I. Clinical trial registration: An essential step toward transparency in clinical research. Indian Journal of Anaesthesia. 2023;67: 321. doi:10.4103/ija.ija_234_23

173. Naudet F, Patel CJ, DeVito NJ, Goff GL, Cristea IA, Braillon A, et al. Improving the transparency and reliability of observational studies through registration. BMJ. 2024;384: e076123. doi:10.1136/bmj-2023-076123

Study I

Applied Machine Learning Techniques to Diagnose Voice-Affecting Conditions and Disorders: Systematic Literature Review

Published as:

Idrisoglu A, Dallora AL, Anderberg P, Berglund JS. Applied Machine Learning Techniques to Diagnose Voice-Affecting Conditions and Disorders: Systematic Literature Review. Journal of Medical Internet Research. 2023;25: e46105. doi:10.2196/46105.

Review

Applied Machine Learning Techniques to Diagnose Voice-Affecting Conditions and Disorders: Systematic Literature Review

Alper Idrisoglu[1], MSc; Ana Luiza Dallora[1], PhD; Peter Anderberg[1,2], PhD; Johan Sanmartin Berglund[1], MD, PhD

[1]Department of Health, Blekinge Institute of Technology, Karslkrona, Sweden

[2]School of Health Sciences, University of Skövde, Skövde, Sweden

Corresponding Author:
Alper Idrisoglu, MSc
Department of Health
Blekinge Institute of Technology
Valhallavägen 1
Karslkrona, 37141
Sweden
Phone: 46 701462619
Email: alper.idrisoglu@bth.se

Abstract

Background: Normal voice production depends on the synchronized cooperation of multiple physiological systems, which makes the voice sensitive to changes. Any systematic, neurological, and aerodigestive distortion is prone to affect voice production through reduced cognitive, pulmonary, and muscular functionality. This sensitivity inspired using voice as a biomarker to examine disorders that affect the voice. Technological improvements and emerging machine learning (ML) technologies have enabled possibilities of extracting digital vocal features from the voice for automated diagnosis and monitoring systems.

Objective: This study aims to summarize a comprehensive view of research on voice-affecting disorders that uses ML techniques for diagnosis and monitoring through voice samples where systematic conditions, nonlaryngeal aerodigestive disorders, and neurological disorders are specifically of interest.

Methods: This systematic literature review (SLR) investigated the state of the art of voice-based diagnostic and monitoring systems with ML technologies, targeting voice-affecting disorders without direct relation to the voice box from the point of view of applied health technology. Through a comprehensive search string, studies published from 2012 to 2022 from the databases Scopus, PubMed, and Web of Science were scanned and collected for assessment. To minimize bias, retrieval of the relevant references in other studies in the field was ensured, and 2 authors assessed the collected studies. Low-quality studies were removed through a quality assessment and relevant data were extracted through summary tables for analysis. The articles were checked for similarities between author groups to prevent cumulative redundancy bias during the screening process, where only 1 article was included from the same author group.

Results: In the analysis of the 145 included studies, support vector machines were the most utilized ML technique (51/145, 35.2%), with the most studied disease being Parkinson disease (PD; reported in 87/145, 60%, studies). After 2017, 16 additional voice-affecting disorders were examined, in contrast to the 3 investigated previously. Furthermore, an upsurge in the use of artificial neural network–based architectures was observed after 2017. Almost half of the included studies were published in last 2 years (2021 and 2022). A broad interest from many countries was observed. Notably, nearly one-half (n=75) of the studies relied on 10 distinct data sets, and 11/145 (7.6%) used demographic data as an input for ML models.

Conclusions: This SLR revealed considerable interest across multiple countries in using ML techniques for diagnosing and monitoring voice-affecting disorders, with PD being the most studied disorder. However, the review identified several gaps, including limited and unbalanced data set usage in studies, and a focus on diagnostic test rather than disorder-specific monitoring. Despite the limitations of being constrained by only peer-reviewed publications written in English, the SLR provides valuable insights into the current state of research on ML-based voice-affecting disorder diagnosis and monitoring and highlighting areas to address in future research.

(J Med Internet Res 2023;25:e46105) doi: 10.2196/46105

KEYWORDS

diagnosis; digital biomarkers; machine learning; monitoring; voice-affecting disorder; voice features

Introduction

Voice-Affecting Disorders

Voice and speech production relies on complex and multiorgan cooperation. The basic mechanics of speech and voice creation is that the airflow obtained by releasing the pressure in the lungs reaches the vocal folds in the larynx and vibrates the vocal cords that result in voice, and by articulating this voice speech is created [1]. The harmony between complex biological systems involved in voice and speech production leads to normal voice formation. However, at the same time, the functional dependency of several biological structures makes the voice vulnerable to being affected by diverse conditions, which may result in a pathological or disordered voice named hoarseness (ie, dysphonia).

The anomalies and the absence of vocal quality in relation to pitch, height, resonance, and duration, which are unexpected for individuals, regardless of their gender and age are characteristics of a disordered voice [2-5]. There is no globally accepted nomenclature for voice disorders. In general, structural, inflammatory, traumatic, systemic, aerodigestive, psychiatric and psychological, neurological, and functional voice disorders are substantial categories of voice problems [6]. This can be diagnosed by a health care specialist through several examinations and tests. The current approach for the diagnosis of voice disorders relies on clinical examinations consisting of interviews, perceptual voice evaluation, patient-reported outcome measures, laryngoscopy, aerodynamic assessment, voice profile, acoustic analysis, and laryngeal electromyography [7], which is time-consuming for both the patients and the clinicians and generates a high economic burden on society [8]. The appraisal based on the assessment of biomarkers gathered through clinical examinations is a crucial step that leads to a diagnosis. Here, it is necessary to point out that the clinicians do not diagnose dysphonia; instead, the target of the clinical examination is to identify the condition that leads to dysphonia, which will be addressed in this study as a voice-affecting disorder.

Voice as a Digital Biomarker

Measurable, reliable, and repeatable assets that can be correlated with a clinical outcome are defined as biomarkers. The criteria and context of use describe the category of biomarkers such as diagnostic, monitoring, pharmacodynamic/response, predictive, prognostic, and digital biomarkers [9]. Traditional biological markers (ie, biomarkers) are used to detect molecular changes associated with diseases and have been integrated with clinical practices for decades [10]. Digital biomarkers refer to measures or features collected by digital devices [11,12] and are a developing landscape that shares the same objectives as traditional biomarkers in answering health-related questions [13].

As aforementioned, voice and speech can be influenced by several conditions and disorders, which contribute to decreased quality of life. Nevertheless, being so sensitive could open possibilities for earlier diagnosis of disorders that affect the voice through the use of voice as a biomarker [14]. As the collection of voice and speech is a noninvasive process that can be performed at a low cost [11], the voice as a digital biomarker could be a diagnostic and prognostic resource with the potential to be more economically viable, in addition to being a more ecological measure than many of the currently used clinical alternatives for the assessment of cognition and function [9,15].

Machine Learning for the Assessment of Voice Signals

Most health problems could benefit from an early diagnosis for better treatment and management of outcomes. However, the growing pressure on health care systems, due to the increased life expectancy and an aging population, may hinder this early detection. Patients are usually referred to specialist care only when apparent signs of disease are present, and thus are at an already moderate advanced state. Fortunately, the existence of digital biomarkers (eg, voice), along with the trend of digitalization in health care, opens the possibility of using technologies such as machine learning (ML) to address these issues. Research on biomarkers extracted from voice and speech with ML techniques for diagnosing, prognosticating, and monitoring disease [14] has shown satisfactory outcomes for disorders such as dementia, depression, mild cognitive impairment (MCI), autism spectrum disorder, Alzheimer disease (AD), and PD [14,16-18].

ML techniques are becoming prevalent in health care for aiding decision-making in treatment and diagnosis [19]. These techniques involve extracting features from voice data and using an ML algorithm to classify the severity of disorders or to determine whether a voice is pathological. The 2 most commonly used ML techniques in this context are supervised and unsupervised learning. In supervised learning, an ML technique is trained using labeled data sets (training set) and its accuracy is evaluated using unlabeled data sets (validation and test sets). The labeled data contain the actual diagnostic information that allows the ML technique to compare its output and adjust its parameters for improved accuracy. There are also end-to-end algorithms (ie, deep learning), an ML subfield that uses artificial neural networks to model and solve complex problems. These networks are composed of multiple layers of interconnected nodes that enable them to learn hierarchical representations of data. An example of deep learning in action is the use of convolutional neural networks to classify images. Convolutional neural networks use multiple layers of convolution and pooling operations to extract features from the input image and then classify it into 1 of several categories. This approach has been very successful in tasks such as object recognition, image segmentation, and speech recognition [20]. Unsupervised learning involves applying clustering methods on training data without labels to group data through 1 or several clustering algorithms [21].

Prior studies provide comprehensive information on feature extraction and its application [22-25]. Encouraging results on

ML classifiers with voice biomarkers bring them into the focus of researchers. In a meta-analysis on voice disorders, Syed et al [26] applied ML techniques by setting the boundaries around 3 publicly available databases, namely, Saarbrucken Voice Database, Massachusetts Eye and Ear Infirmary, and Arabic voice pathology database. The systematic literature review (SLR) presented herein includes all possible data sources. Several reviews have investigated voice-based disease diagnostics with ML algorithms separately, focusing on only PD or AD [27-29], whereas in this study, multiple voice-affecting disorders are included.

This SLR investigates state of the art of clinical applications of voice-based diagnosis that make use of ML algorithms. Adapting voice-based diagnosis and prognosis into clinical practices requires solid evidence and research to clinically validate the usability and reliability of voice biomarkers and the performance of ML classifiers [12,15]. This SLR does not consider disorders directly related to the voice production mechanisms. Examples of included and excluded conditions are PD and polyps on the vocal cords, respectively, where PD is a neurodegenerative disorder that often causes voice changes [30] that are not directly related to voice box and polyps on the vocal cords occur in the voice box (larynx) [31]. More specifically, the conditions of interest in this SLR are listed as systematic conditions affecting voice, nonlaryngeal aerodigestive disorders affecting voice, and neurological disorders affecting voice in the *Classification Manual for Voice Disorders* [6]. These have a higher chance of being related to chronic conditions, which would benefit from having a scalable and noninvasive method for screening a large population. The expected outcome is a contribution, not only by summarizing the work done in the field of applied health technology that is interested in the application of the technology and its outcomes but also by pointing out the gaps in the literature and possible future research directions that could address the problems mentioned earlier of the next generations and the health care system.

Methods

Overview and Purpose of This SLR

An SLR is a summary of the results from research papers focused on a common context or a question. The summary action includes the identification, collection, assessment, and synthesizing of high-quality research evidence within the scope of the research question by following a predefined protocol. The aim of an SLR is to provide perspective on recent research so that decision makers can benefit from up-to-date knowledge and address the gaps that can be used as a basis for new research. The predefined protocol describes the methodology to follow; defines the research question; and contains information about inclusion/exclusion criteria and quality assessment [30]. This section specifies the methodology applied in this SLR to answer the research question "How is the voice as a digital biomarker being used in clinical applications that employ ML techniques for diagnosing and monitoring voice-affecting disorders?" Additionally, the main question is split into the following subquestions (SQs):

- *SQ1:* What are the aims of pathologic voice evaluation?
- *SQ2:* Which ML techniques are being used for the diagnosis and monitoring of voice-affecting disorder through voice and which voice-affecting disorders are being investigated?
- *SQ3:* What are the time and geographical trends of publications in the scope of SLR?
- *SQ4:* What are the data characteristics of the sound samples for different disorders and types of studies?
- *SQ5:* Are the studies cross-sectional or longitudinal?
- *SQ6:* How is performance being evaluated in the studies?

All the information on the methodological approach that guided the execution of this SLR is based on the prespecified SLR protocol [31].

Search Strategy

A search string was constructed by applying the *p*opulation, *i*ntervention, *c*omparison, and *o*utcome (PICO) framework [32-34]. The most used terms, suggested by authors ALD and JSB, were used to find relevant papers; the retrieved papers and their references in the field were then used to discover new adequate keywords. By adding the new keywords to the search string, a comprehensive search string was created. A customized version of the search string in Textbox 1 was used in PubMed, Scopus, and Web of Science databases to find all relevant peer-reviewed primary journal articles published between 2012 and 2022. The application of the PICO structure excludes the *comparison* due to the nature of this SLR being a characterization:

- *Population:* Disorders that affect the voice, given by the *Classification Manual for Voice Disorders* [30], referring to the systematic conditions affecting voice, nonlaryngeal aerodigestive disorders affecting voice, and neurological disorders affecting voice.
- *Intervention:* Use of ML techniques for the diagnosis or monitoring of disorders through voice samples.
- *Outcome:* Reported quantities or results such as precision and accuracy.

The search string was adapted based on the advanced search requirements of each database. The filter options were tuned to retrieve articles from January 1, 2012 to December 31, 2022. The period was chosen after consulting with experts in the medical field with regard to the development of new technologies for health care. The development of the search string was primarily based on the MeSH (Medical Subject Headings) terms, with the help of a librarian, and categories of voice disorders in the classification manual [6].

Textbox 1. Search string used in PubMed, Scopus, and Web of Science databases (search date: March 13, 2023).

((“Voice” OR “Linguistic features” OR “acoustic parameters” OR “Vocal features” OR “Vocal” OR “Vocal Cords” OR “Vocal biomarker” OR “Voice biomarkers” OR “ Speech” OR “Vowel” OR “Sound Spectrography” OR “Cepstrum Vectors”) AND (“Deep phenotyping” OR “selection” OR “extrac- tion” OR “Detection” OR “Monitoring” OR “Classification” OR “Evaluation” OR “Analysis” OR “Estimation” OR “Projection” OR “Improving” OR “Investigation” OR “Prognosis” OR “Predict*”) AND (“Sensitivity” OR “Accuracy” OR “Specificity” OR “Performance” OR “Cross-validation” OR “precision”) AND (“Voice technology” OR “Machine learning” OR “Artificial Intelligence” OR “Gaussian mixture models” OR “Support vector machines” OR “Artificial neural network” OR “Data Mining” OR “Decision Support System” OR “Clinical Support System” OR “Deep Neural Network” OR “Kernel extreme learning machine” OR “Deep Learning”) AND (“voice disorder” OR “systemic conditions” OR “aerogestive disorders” OR “neurologic disorders” OR “central nervous system disturbance” OR “Endocrine” OR “Hypothyroidism” OR “Hyperthyroidism” OR “Sexual Hormone Imbalances” OR “Hyperpituitarism” OR “Immunologic” OR “Allergic” OR “HIV” OR “Chronic Fatigue Syndrome” OR “Systemic Lupus Erythematosus” OR “Sjogren's Syndrome” OR “Scleroderma” OR “Wegener's Disease” OR “Musculo-Skeletal Conditions Affecting Voice” OR “Overuse Injury and Repetitive Strain Injury” OR “Fibromyalgia” OR “Ehler Danlos Syndrome” OR “Dehydration” OR “Respiratory Diseases Affecting Voice” OR “Asthma” OR “Chronic Obstructive Pulmonary Disease” OR “Digastric” OR “Gastroesophageal Reflux Disease” OR “Infectious Diseases of the Aerodigestive Tract” OR “Laryngotracheobronchitis” OR “Pertussis” OR “Diphtheria” OR “Pneumonia” OR “Infectious Sinusitis” OR “Tuberculosis” OR “Upper Respiratory Infection” OR “Acute Epiglottitis” OR “Syphilis” OR “Sarcoidosis” OR “Scleroma” OR “Leprosy” OR “Actinomycosis” OR “Mycotic Infections” OR “Blastomycosis” OR “Histoplasmosis” OR “Candidiasis” OR “Coccidioidomycosis” OR “Peripheral Nervous System Pathology” OR “Superior Laryngeal Nerve Pathology” OR “Unilateral Recurrent Laryngeal Nerve Paralysis” OR “Recurrent Laryngeal Nerve Paresis” OR “Bilateral Recurrent Laryngeal Nerve Paralysis--Peripheral” OR “Myasthenia Gravis” OR “Peripheral Neuropathy” OR “Enhanced Physiologic Tremor” OR “Movement Disorders” OR “Adductor Spasmodic Dysphonia” OR “Abductor Spasmodic Dysphonia” OR “Abductor Spasmodic Dysphonia” OR “Dystonic Tremor” OR “Essential Tremor” OR “Meige's Syndrome” OR “Tardive Stereotypies” OR “Tourette's Syndrome” OR “Amyotrophic Lateral Sclerosis” OR “Wallenberg Syndrome” OR “Lateral Medullary Syndrome” OR “Infarct” OR “Parkinson Disease” OR “Multiple Systems Atrophy” OR “Shy-Drager Syndrome” OR “Striatonigral Degeneration” OR “Sporadic Olivoponto- cerebellar Atrophy” OR “Progressive Supranuclear Palsy” OR “Multiple Sclerosis” OR “Cerebellar Disorders” OR “Huntington's Chorea” OR “Bilateral Recurrent Laryngeal Nerve Paralysis--Central” OR “Myoclonus” OR “Neuromuscular” OR “cardiovascular” OR “coronary artery” OR “heart attack” OR “Voice disorders” OR “Neurological disorders” OR “multiple sclerosis” OR “Myasthenia gravis ” OR “ALS” OR “Amyotrophic lateral sclerosis” OR “Parkinson's disease” OR “Multiple sclerosis” OR “Dementia” OR “Alzheimer's disease” OR “Essential tremor” OR “Major depressive disorder” OR “pathological voice” OR “voice pathology” OR “neurodegenerative” OR “Cognitive impairment” OR “Nodule” OR “Polyp” OR “Neoplasm” OR “dysphonia” OR “Hoarseness” OR “Huntington disease”))

Study Selection

The search string was used to perform an automated search on each database. The Zotero (Corporation for Digital Scholarship) bibliography software was used to collect all relevant articles from all 3 databases and to remove duplicates [35]. First, authors AI and ALD applied the inclusion and exclusion criteria in Textbox 2 to assess the titles and abstracts of the retrieved papers. The first step was to assess randomly selected 50 papers to ensure the consistency of the criteria. Then, another batch containing 50 articles was assessed. Authors AI and ALD compared the results. Upon agreeing on the consistency of the criteria, they proceeded to assess the remainder of the papers. The degree of agreement was checked statistically by comparing the results between the first and second authors with an overall agreement of 96% using the Cohen κ index. During the evaluation, the papers were categorized into 3 groups: included, excluded, and “maybe” cases that could not be assessed by the content of the title and abstract alone. At this stage, author JSB acted as the advisor and expert in the field. Furthermore, after the evaluation of all papers, the results from both authors were cross-checked, and 30 conflicts were noticed. To minimize the risk of bias, all articles marked as included, “maybe,” and conflicts were grouped for full-text reading.

All articles in the group of full-text readings underwent a quality assessment procedure to assure high-quality evidence (Textbox 3), based on guidelines proposed by Kitchenham and Charters [36]. The quality threshold was set to 11 points, which means that articles below the score of 11 points would be rejected. The threshold of 11 points was stipulated through group discussions with authors. The questionnaire was designed in 3 sections, consisting of 5 questions each, general questions, data analysis, and results. Based on the given questions, author AI performed the quality assessment by grading the studies with scores 0, 0.5, and 1 for the sections 1, 2, and 3, respectively.

Textbox 2. Inclusion and exclusion criteria for the assessment of the articles.

Inclusion criteria

- Journal study
- Primary study written in English
- Research published not earlier than 2012
- Research that uses voice as the input data
- Research that employs at least one machine learning algorithm
- Research that aims to diagnose or monitor at least one voice-affecting disorder not related to the systematic conditions affecting voice, nonlaryngeal aerodigestive disorders affecting voice, and neurological disorders affecting voice

Exclusion criteria

- A nonpeer-reviewed study
- Research written in languages other than English
- Research published before 2012 or after 2022
- Research that does not use voice as a direct input, which means research employing various nonverbal forms of data input, such as written transcriptions, digital images, videos, electroencephalogram, and signals generated during vocalization
- Research that classifies voice-affecting disease without a machine learning approach
- Research that classifies voice disorders related to conditions other than systematic conditions affecting voice, nonlaryngeal aerodigestive disorders affecting voice, and neurological disorders affecting voice

Textbox 3. Quality assessment questionnaire.

General questions

- Are the aims clearly stated?
- Is the targeted population described?
- Has it discussed the contribution of the study?
- Are gender and age considered?
- Is/are the technique(s) being implemented clearly described?

Data analysis

- Is the origin of data given?
- Is the type of data clearly described?
- Do the data consist of voice recordings?
- Is the data validation method given?
- Is there a discussion on whether the data size can be generalized for the targeted population?

Result

- Is/are the result(s) clearly discussed?
- Are all aims or questions answered?
- Was the outcome related to the target population?
- Are the limitations discussed?
- Did results compare with previous rapports?

Data Extraction

Data extraction was carried out by author AI. Table 1 shows the list of attributes, definitions, and purpose of use for data extraction.

Table 1. Collected data attributes.

Attribute	Definition
ISSN	International Standard Serial Number recorded
Title	Full title of the research
Journal	Publication venue record
Authors	All authors' names
Publication date	The publication date of the paper
Publication type	The type of publication
Origin of publication	The geographical location of the first author's institution
Targeted disorder	Investigated disorder
Database	Source of the data
Origin of data	The geographical location of data sources
Data characteristics	Type of voice recordings
Additional data	Used additional data except for voice recordings
Data sets	The number of participants
Sample size	The number of recordings
Aim of the study	Purpose of the study
Age range	The considered age range of the participants
Gender	The number of participants (by gender)
Quantitative result(s)	Presented outcome measures
Feature sets	Excluded features from voice
The proposed features	The best feature set, if exists
Applied ML[a] technique(s)	All applied ML techniques
Outcome evaluation	How the pathological voice is evaluated
Type of validation(s)	How the data set is divided
Type of study	If the study is longitudinal or cross-sectional
The proposed ML algorithm(s)	ML technique with the best outcome

[a]ML: machine learning.

Data Analysis

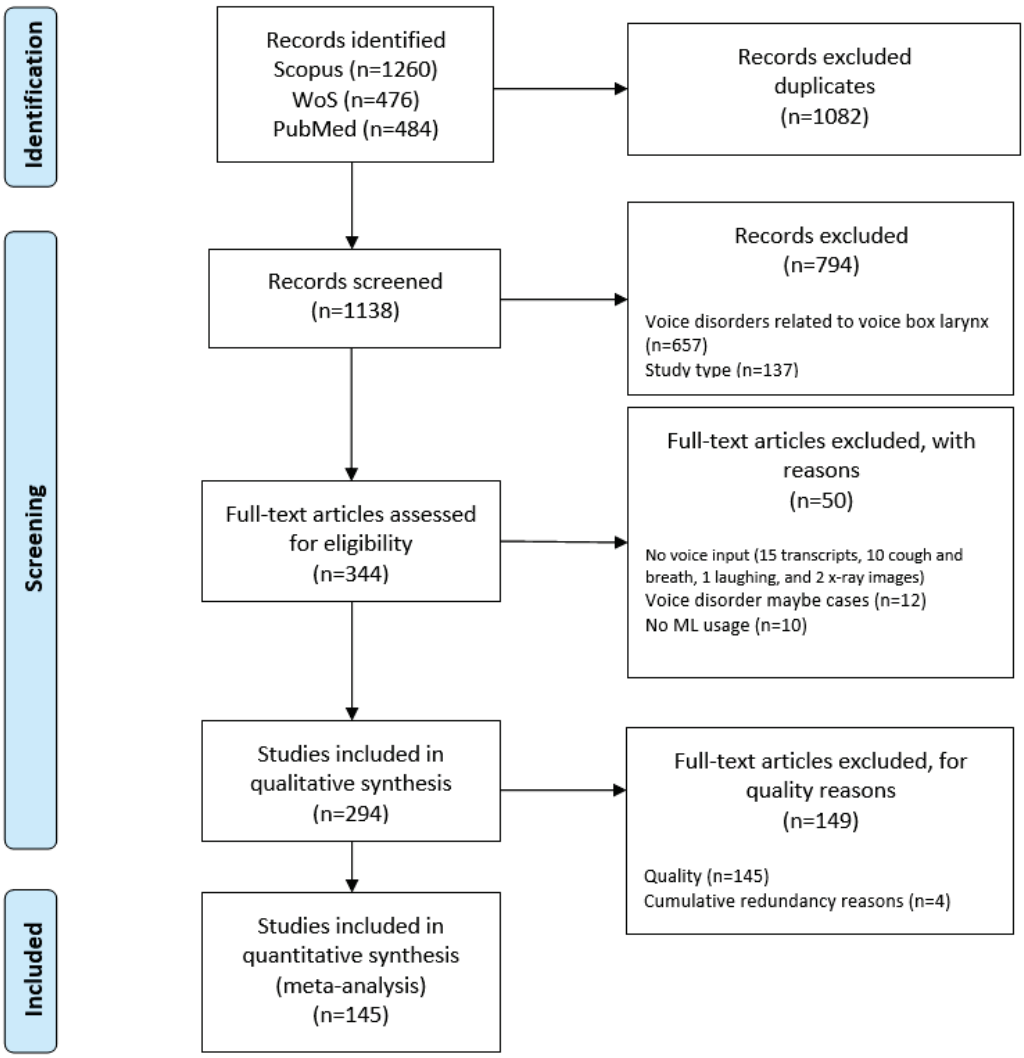
To analyze the etiology of changes over time and capture the heterogeneity, the studies were grouped into subgroup summary tables entitled with the name of disorders (see Multimedia Appendix 1). Numerical and statistical measures were used to represent the results. No assumption was made about the missing information. Microsoft Excel was used for data analysis. All the studies that successfully adhered to the inclusion and exclusion criteria and passed the quality assessment were eligible for data analysis. The results were presented in text, summary tables, and charts under a section for each research question. The robustness of the results was checked by conducting a sensitivity analysis through observations of the effect of some randomly removed data from summary tables [37,38]. The cumulative redundancy bias was checked by observing the similarity between author groups.

Results

Study Selection

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses; also see Multimedia Appendices 2 and 3) flowchart for this study is shown in Figure 1 [39]. The automated search retrieved a total of 2220 articles from all 3 databases (Scopus, n=1260; Web of Science, n=476; and PubMed, n=484). After the removal of the duplicates, 1138 articles were assessed in the title and abstract screening. In total, 344 papers were included in the full-text reading group. During the full-text reading, 50 articles were found to be out of scope for the following reasons: related to voice box (n=12), voice was not an input (n=28; 15 transcripts, 10 coughs and breath, 1 laughing, and 2 x-ray images), and no ML technique (n=10) applied. A total of 294 articles were assessed for quality evaluation, which eliminated 145 articles and thus the final set included 149 articles that were used for data extraction.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart. ML: machine learning; WoS: Web of Science.



The included papers were assessed for cumulative redundancy bias during data extraction. The assessment showed that 8 of the included papers were published by 4 different author groups, 2 papers from Tuncer et al [40,41], 2 papers from Gunduz et al [42,43], 2 papers from Lamba et al [44,45] in the PD group, and 2 papers from Tena et al [46,47] in the amyotrophic lateral sclerosis (ALS) group. To reduce the risk of bias, only 4 of those 8 papers, one with the highest accuracy from each author, were included in the synthesis [40,42,44,46]. In total, 145 studies proceeded for analysis. Sensitivity analysis did not show any effect on trend analysis, but did show a minor effect on statistical analysis. Multimedia Appendix 1 shows the list of included studies in this SLR.

Aims of the Studies

To answer SQ1, this section sums up the aim and the assessment strategy used in studies. Many terminologies have been used to describe the aim of the studies. Generally, they can be arranged into 2 groups: diagnosis and monitoring. In the diagnosis group, 138 studies were identified. A total of 125 studies in the diagnostic group investigated ML methods to detect a pathological voice, where the participants were grouped as the healthy control (HC) group, with being "healthy" defined as people without a diagnosed disorder, and a group with known pathology [17,40,42,44,48-168]. The main idea was to deploy an ML technique for distinguishing those 2 groups from each other with high accuracy. Additionally, 13 studies [169-181] in the diagnostic group investigated ML techniques for separating several pathologies and clustered participants into several pathological groups. With the help of the ML technique, they

tried to classify each group, where the primary purpose was to investigate a system that can classify multiple disorders. A total of 7 studies [182-188] were identified in the monitoring group. The pattern was trying to predict an established clinical severity assessment with the help of an ML algorithm where only participants with diagnosed disorders were involved.

Employed ML Techniques and Voice-Affecting Disorders

Table 2 shows the results related to SQ2. A total of 19 different disorders were identified where the focus was on monitoring or diagnosis through voice or speech with ML involvement. As many as 87/145 (60%) of the studies targeted PD; 18 studies targeted dementia or AD, 8 cognitive impairment (CI)/MCI, 4 ALS, 2 cardiovascular disorders, 7 COVID-19, 2 essential tremor, 2 multiple sclerosis, 1 neurodegenerative cognitive complaint (NCC), 1 functional dysphagia/oropharyngeal dysphagia, 4 depression, 1 influenza disease, 1 neurological disease (ND), 2 stroke, 1 fatigue, 1 autism, 1 traumatic brain injury, 1 asthma, and 1 chronic obstructive pulmonary disease. NCC and ND may potentially be classified within one of either PD, AD, or CI/MCI due to their similar symptoms, but the specific underlying disorder was not provided in the studies. Therefore, these 2 disorders were grouped separately.

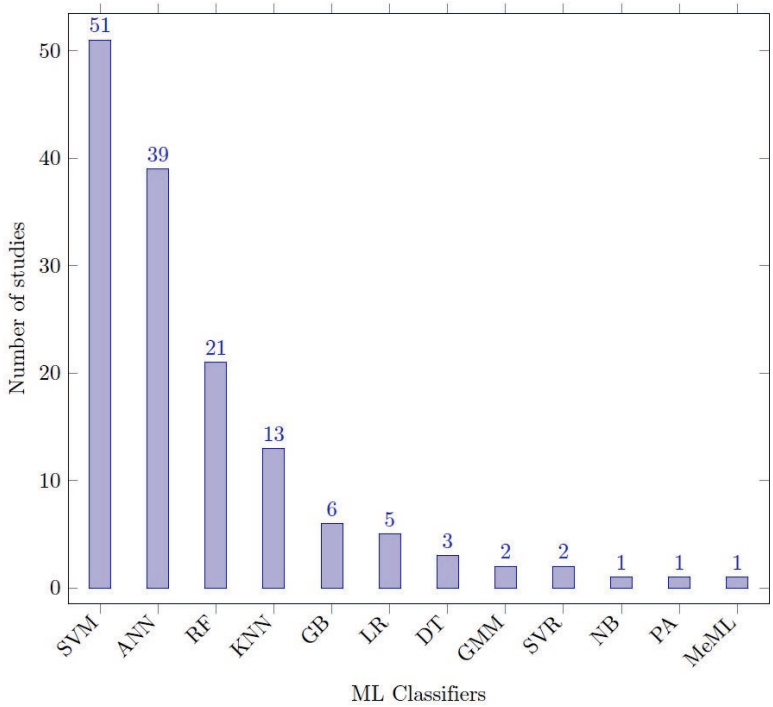
Table 2. Targeted disorders and ML[a] techniques.

Disorder	NR[b]	ML technique (NR of usage)	References
Parkinson disease	87	SVM[c] (34), ANN[d] (23), RF[e] (9), KNN[f] (6), GB[g] (5), GMM[h] (2), NB[i] (1), DT[j] (3), SVR[k] (2), LR[l] (1), PA[m] (1)	[17,40,42,44,48-102,117-135,168,170,171,180,181,183-185,188]
Dementia, Alzheimer disease	18	SVM (8), KNN (2), LR (2), RF (3), ANN (3)	[103-110,136-139,172-176,179]
Cognitive impairment/mild cognitive impairment	8	SVM (2), LR (2), RF (2), ANN (2)	[112-116,150,177,178]
COVID-19	7	ANN (3), SVM (1), RF (1), KNN (1), GB (1)	[143-147,156,157]
Amyotrophic lateral sclerosis	4	SVM (1), RF (2), MeML[n] (1)	[46,158-160]
Depression	4	ANN (3), SVM (1)	[140-142,161]
Cardiovascular disorders	2	KNN (2)	[111,186]
Essential tremor	2	SVM (2)	[162,163]
Multiple sclerosis	2	ANN (1), RF (1)	[149,164]
Stroke	2	ANN (2)	[148,153]
Asthma	1	RF	[182]
Autism	1	ANN	[151]
Fatigue	1	SVM	[187]
Chronic obstructive pulmonary disease	1	RF	[154]
Neurodegenerative cognitive complaint	1	SVM	[165]
Functional dysphagia, oropharyngeal dysphagia	1	RF	[166]
Influenza disease	1	KNN	[167]
Neurological disease	1	KNN	[169]
Traumatic brain injury	1	ANN	[152]

[a]ML: machine learning.
[b]NR: number of studies.
[c]SVM: support vector machine.
[d]ANN: artificial neural network.
[e]RF: random forest.
[f]KNN: K-nearest neighbor.
[g]GB: gradient boosting.
[h]GMM: Gaussian mixture model.
[i]NB: naïve Bayes.
[j]DT: decision tree.
[k]SVR: support vector regression.
[l]LR: logistic regression.
[m]PA: passive aggressive.
[n]MeML: mixed effect machine learning.

The usage of the 12 ML techniques is shown in Figure 2, where the support vector machine algorithm was the most used (51/145, 35.2%) and artificial neural networks were the second most utilized technique (39/145, 26.9%) among all ML techniques. Several studies have tested and compared different algorithms. Figure 2 shows the ML technique with the best results from each study. The support vector machine notation contains all different kernel combinations, and all utilized neural network architectures are grouped under artificial neural network. As many as 11 of the 12 ML techniques shown in Figure 2 have tested on PD, 5/12 on AD, 4/12 on CI/MCI, 5/12 on COVID-19, 3/12 on ALS, 1/12 on cardiovascular disorders, 1/12 on essential tremor, 2/12 on multiple sclerosis, 1/12 on stroke, 1/12 on asthma, 1/12 on autism, 1/12 on fatigue, 1/12 on chronic obstructive pulmonary disease, 1/12 on NCC, 1/12 on functional dysphagia/oropharyngeal dysphagia, 2/12 on depression, 1/12 on influenza disease, 1/12 on ND, and 1/12 on traumatic brain injury (Table 2).

Figure 2. The usage of machine learning algorithms. ANN: artificial neural network; DT: decision tree; GB: gradient boosting; GMM: Gaussian mixture model; KNN: K-nearest neighbor; LR: logic regression; MeML: mixed effect machine learning; ML: machine learning; NB: naïve Bayes; PA: passive active; RF: random forest; SVM: support vector machine; SVR: support vector regression.



Time and Geographical Trend of the Publications

Figure 3 shows the published studies by year and the investigated disorders. The results indicate that there is an upward trend in the studies involving the application of ML for voice-affecting disorder. Up to 2016, the focus of the research was solely on PD and AD. In the last 5 years, the research on voice-based diagnosis and monitoring with ML has not only increased but also diversified in terms of the investigated voice-affecting disorders with the addition of CI/MCI, ALS, cardiovascular disorder, essential tremor, COVID-19, multiple sclerosis, NCC, functional dysphagia/oropharyngeal dysphagia, depression, influenza disease, ND, stroke, asthma, autism, fatigue, chronic obstructive pulmonary disease, and traumatic brain injury. In addition, the highest publication rate occurred in 2022 (more than doubled compared with previous years); 51/145 studies included in this SLR have been published in 2022, which corresponds to 35.1% of all listed articles in Multimedia Appendix 1.

Figure 4 displays the contribution from countries for a specific disorder. Some countries tend to focus more on 1 disorder, while others investigated several voice-affecting disorders using ML techniques. In addition, PD seems to be the most investigated disorder for the majority of countries. The geographical trend described in this section reflects the country in which the study was performed and not the geographical source of the sample.

Figure 3. Usage of ML techniques and investigated disorders by year. AD: Alzheimer disease; ALS: amyotrophic lateral sclerosis; ANN: artificial neural network; CD: cardiovascular disease; CI: cognitive impairment; DT: decision tree; ET: essential tremor; GB: gradient boosting; GMM: Gaussian mixture model; KNN: K-nearest neighbor; LR: logic regression; MCI: mild cognitive impairment; MeML: mixed effect machine learning; ML: machine learning; MS: multiple sclerosis; NB: naïve Bayes; PA: passive active; PD: Parkinson disease; RF: random forest; SVM: support vector machine; SVR: support vector regression.

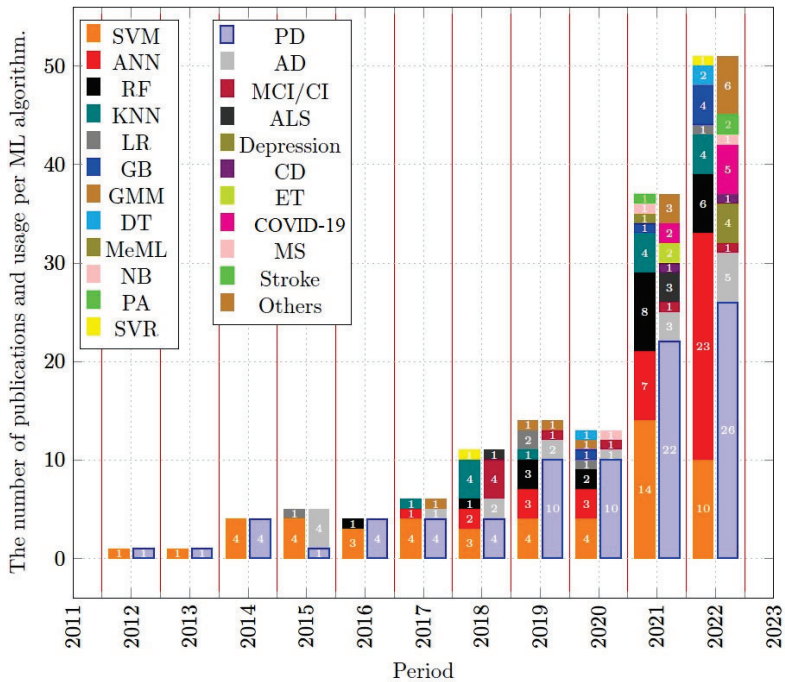
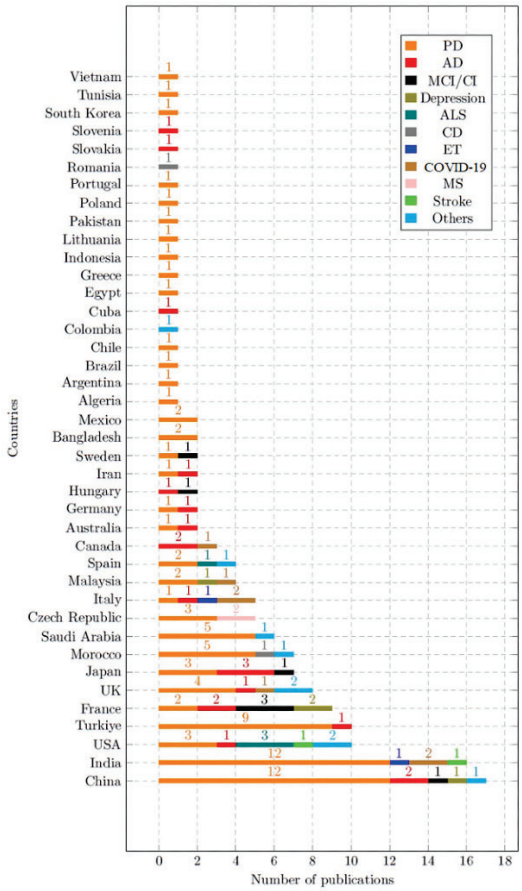


Figure 4. Investigated disorders by country. AD: Alzheimer disease; ALS: amyotrophic lateral sclerosis; CD: cardiovascular disorder; CI: cognitive impairment; ET: essential tremor; MCI: mild cognitive impairment; MS: multiple sclerosis; PD: Parkinson disease.



Data Characteristics

This section describes the characteristics of voice and nonvoice data used as input into the ML models. A total of 11/145 (7.6%) studies integrated nonvocal data in conjunction with vocal features to form the input feature sets for the ML models; 6 of these studies [54,66,96,105,149,163] incorporated demographic data, including gender, age, BMI, comorbidities, weight, height, and disease duration. Meanwhile, 2 of the studies [141,153] used video inputs, while 3 studies [83,158,159] incorporated external sensor signals such as electromyography and motion trackers.

Table 3 compiles the disorders and frequency of recorded data characteristics with the density of extracted vocal features and data source. Results indicate that vowel phonations are one of the most adopted recording types among almost all listed disorders. A total of 68 studies chose to base their analysis on vowel recordings, 33 combining different recordings, 20 free speeches, 12 scripted speeches, 9 picture descriptions, and 3 studies used syllable recordings. Cognitive disorders (eg, AD and MCI) that tend to use voice features extracted from speech and other disorders (eg, PD) lean toward features extracted from vowel phonation.

In all studies, raw data that consist of recordings underwent signal processing to extract features that were used as input data into ML techniques. Identified signal processing implementations were baseline acoustic (BLA), Mel-frequency cepstral coefficients, tunable Q-factor wavelet transform, wavelet transform, and spectrogram, which are frequency-transformed versions of the input signal that generates features in the form of digits and images. Other utilized features were linguistic and vocal features that generate statistical outputs (eg, silence rate, pause rate, duration, and ineligibility). In addition, combining several vocal features is more popular than only using BLA features, where almost 69/145 (47.6%) of the studies combined several features as input. But still, BLA features (36/145, 24.8%) are one of the most separately used feature sets. In this SLR BLA corresponds to all or a portion of acoustic, time, and frequency domain features calculated from raw recordings (eg, pitch, zero cross rate, jitter, shimmer, and formant frequencies).

Table 3. Characteristics of the input data and data source.

Disorder	Recording	Feature	Data source
Parkinson disease	• Vowel: 59 [17,40,42,44,48,50,54–59,61–64,68,69,74–85,87–91,93–97,99,100,102,119–122,124,126–128,130–132,135,170,183,184,192] • Combined: 20 [49,51,53,60,65,66,71,73,98,101,118,123,125,133,134,171,180,181,185,188] • Scripted speech: 3 [52,72,117] • Free speech: 3 [70,92,129] • Syllable: 2 [67,86]	• BLA[a]: 26 [17,40,50,53,56,58-60,64-66,71,73,77,79,82,84,87,88,92,98,118,132,133,184,188] • Combined: 45 [42,44,48,51,52,54,57,62,63,67–70,72,74,76,79,81,85,86,89–91,93,94,96,97,99,100,119,120,122,124,126–131,134,135,170,171,184,192] • MFCC[b]: 5 [49,61,80,94,101] • Spectrogram: 7 [75,117,123,125,180,181,185] • RP[c]: 1 [95] • TQWT[d]: 2 [55,121] • WT[e]: 1 [83]	• CFS[f]: 31 [48,51,52,65,67,70,72-74,78-81,83,85,86,88,92,94,96,97,101,102,117,130,131,170,171,180,185] • UCI[g]: 33 [40,42,44,50,53,55–57,60,61,63,64,68,69,71,82,84,91,93,99,100,119,120,122,124,126,127,129,133,135,183,188,192] • Multiple: 14 [49,54,58,59,66,75,87,90,98,118,123,125,132,181] • mPower: 4 [17,62,77,128] • NCVS[h]: 1 [88] • PARCZ[i]: 1 [89] • NG[j]: 3 [95,121,134]
Dementia, Alzheimer disease	• Free speech: 8 [103,105,108,172,173,175] • Picture description: 7 [109,110,136-139,174] • Scripted speech: 2 [104,176] • Combined: 1 [179]	• Combined: 11 [103,104,107,109,110,136,138,139,173-175] • Vocal: 2 [105,172] • BLA: 2 [176,179] • Spectogram: 2 [108,137] • Speech statistic: 1 [106]	• CFS: 10 [103-108,172,173,175,176] • ADBC[k]: 7 [109,110,136-139,174] • Multiple: 1 [179]
Cognitive impairment/mild cognitive impairment	• Free speech: 3 [112,114,115] • Picture description: 2 [116,178] • Scripted speech: 2 [113,177] • Combined [150]	• BLA: 3 [113,114,177] • Combined: 2 [112,116] • Linguistic: 1 [115] • Vocal: 1 [178] • Spectogram [150]	• CFS: 8 [112-116,150,177,178]
COVID-19	• Vowel: 2 [156,157] • Combined: 5 [143-147]	• Combined: 3 [144,156,157] • Spectogram: 3 [143,145,146] • MFCC [147]	• Coswara: 4 [143,144,156,157] • Multiple: 1 [146] • CFS: 1 [145] • CHRSD[l]: 1 [147]
Amyotrophic lateral sclerosis	• Vowel: 2 [46,160] • Scripted speech: 1 [159] • Combined: 1 [158]	• BLA: 2 [46,160] • Combined: 2 [158,159]	• CFS: 4 [46,158-160]
Depression	• Free speech: 2 [141,161] • Scripted speech [140] • Combined [142]	• Combined: 2 [142,161] • Spectrogram: 2 [140,141]	• Multiple: 2 [141,161] • CFS: 2 [140,142]
Cardiovascular disorders	• Vowel: 1 [111] • Combined: 1 [186]	• BLA: 1 [111] • MFCC: 1 [173	• CFS: 2 [111,186]
Essential tremor	• Vowel: 2 [162,163]	• BLA: 1 [162] • Spectrogram: 1 [163]	• CFS: 2 [162,163]
Multiple sclerosis	• Syllable [164] • Scripted speech [149]	• Spectrogram [164] • BLA [149]	• CFS: 2 [149,164]
Stroke	• Combined: 2 [148,153]	• MFCC [148] • Spectrogram [153]	• CFS: 2 [148,153]
Autism	• Free speech [151]	• Spectrogram [151]	• CFS: 1 [151]
Asthma	• Free speech [182]	• BLA [182]	• CFS: 1 [182]
Fatigue	• Scripted speech [187]	• Combined [187]	• CFS: 1 [187]

Disorder	Recording	Feature	Data source
Chronic obstructive pulmonary disease	• Scripted speech [154]	• BLA [154]	• CFS: 1 [154]
Neurodegenerative cognitive complaint	• Free speech [165]	• Combined [165]	• CFS: 1 [165]
Functional dysphagia/oropharyngeal dysphagia	• Combined [166]	• Combined [166]	• CFS: 1 [166]
Influenza disease	• Vowel [167]	• WT [167]	• CFS: 1 [167]
Neurological disease	• Vowel [169]	• Combined [169]	• CFS: 1 [169]
Traumatic brain injury	• Free speech [152]	• Spectrogram [152]	• TBIBank[m] Coelho corpus: 1 [152]

[a]BLA: baseline acoustic.

[b]MFCC: Mel-frequency cepstral coefficients.

[c]RP: recurrence plot.

[d]TQWT: tunable Q-factor wavelet transform.

[e]WT: wavelet.

[f]CFS: collected for study.

[g]UCI: University of California, Irvine.

[h]NCVS: National Center for Voice and Speech.

[i]PARCZ: Czech Parkinsonian Speech Database.

[j]NG: not given.

[k]ADBC: Alzheimer Dementia Bank blog corpus.

[l]CHRSD: Corona Hack Respiratory Sound data set.

[m]TBIBank: Traumatic Brain injury bank.

A total of 70 studies collected data for a specific study and 75 studies gathered data from an available data set. Sakar et al (2013) [73] and Sakar et al (2019) [48] are 2 different data sets donated to the UCI (University of California, Irvine), which have been used in 15 different included studies in this SLR; 5 studies [53,57,60,71,100] used the UCI data set containing 20 participants with PD and 20 HC participants, and 15 studies [40,42,44,48,55,61,64,68,99,119,121,122,124,126,127] used the UCI data set having 188 participants with PD and 64 HC participants from the same source. UCI and Coswara provide data sets that can be accessed and downloaded without any additional application [73,189]. All other data sources identified in this SLR require an application or are not publicly available. Data sets used in studies are unbalanced. Even if there is equality between the number of participants in terms of disordered and HC groups, a closer inspection of data sets reveals gender inequality. For example, Sakar et al (2013) [73] included 20 participants with PD and 20 HC participants; however, a closer inspection showed that the PD group comprised 6 females and 14 males, and the HC group consisted of 10 females and males, respectively. Another issue is the low number of participants in studies, where only 8/145 studies [17,62,77,78,81,92,113,170] based their outcome on more than 100 participants for both pathological and HC groups at the same time.

Observation Time

SQ5 aims to find out whether studies rely on longitudinal data and observation over time or observations at the same time that study was done. As the authors predefine the participants and measure the exposures and outcomes at the same time in all included studies, all studies in this SLR follow the cross-sectional study design [190].

Performance Evaluation

Measures presented to assess the efficiency of the ML techniques used show diversity in the included articles. Accuracy is one of the most used measures to present the outcome of almost all studies. Sensitivity, specificity, precision, Matthew's correlation coefficient, area under the curve, F_1-score, recall, mean absolute error, R^2, positive predictive value, and negative predictive value were other used measures in combination with accuracy without any standard order. Under the *performance* column in Multimedia Appendix 1, all combinations can be seen; 7 articles, 5 from the PD group [63,89,97,170,171], 1 from the AD group [172], and 1 from the ALS group [114], have

presented results discriminated by gender and only 1 study [92] paid attention to language differences.

Two groups of studies [48,73] from UCI data sets were found to be suitable for meta-analysis due to the homogeneity between studies. The first group consisted of 15 studies using a data set containing voice recordings from 188 participants with PD and 64 HC participants [40,42,44,48,55,61,64,68,99,119,121,122,124,126,127]. The second group consisted of 5 studies [53,57,60,71,100] using voice recordings from 20 participants with PD and 20 HC participants (Table 4). Studies employing the first data set achieved 0.925 average accuracy within an accuracy range of 0.790-0.997. Studies employing the second data set achieved 0.869 average accuracy within an accuracy range of 0.670-0.990.

Table 4. List of comparable studies.

Data set[a]	Classifier	Feature	Performance	Reference
CFS[b] (donator)	SVM[c]	MFCC[d] and TQWT[e]	Accuracy: 0.8600	[48]
UCI[f]	GB[g]	BLA[h] and spectrum	Accuracy: 0.9388	[44]
UCI	KNN[i]	TQWT	Accuracy: 0.9800	[55]
UCI	ANN[j]	BLA	Accuracy: 0.9921	[40]
UCI	SVM	BLA, MFCC, WT[k], and TQWT	Accuracy: 0.9160	[42]
UCI	ANN	MFCC	Accuracy: 0.9674	[61]
UCI	NB[l]	BLA	Accuracy: 0.7897	[64]
UCI	SVM	BLA, MFCC, TQWT, and WT	Accuracy: 0.9470	[68]
UCI	SVM	BLA, MFCC, WT, and TQWT	Accuracy: 0.9350	[99]
UCI	SVM	BLA, MFCC, and TQWT	Accuracy: 0.8660	[119]
UCI	KNN	TQWT	Accuracy: 0.9890	[121]
UCI	RF[m]	BLA and MFCC	Accuracy: 0.8884	[122]
UCI	ANN	BLA, MFCC, and TQWT	Accuracy: 0.9200	[124]
UCI	SVM	BLA, MFCC, TQWT, and WT	Accuracy: 0.9621	[126]
UCI	ANN	BLA, MFCC, and TQWT	Accuracy: 0.9974	[127]
UCI	SVM	BLA	Accuracy: 0.6701	[53]
UCI	RF	BLA and MFCC	Accuracy: 0.9433	[57]
UCI	ANN	BLA	Accuracy: 0.9903	[60]
UCI	ANN	BLA	Accuracy: 0.8647	[72]
UCI	SVM	BLA and MFCC	Accuracy: 0.8750	[100]

[a]Italicized data sets represent Parkinson disease data set 2 containing data on patients with Parkinson disease (n=20) and HC (n=20); all other data sets correspond to Parkinson disease data set 1 containing data on patients with Parkinson disease (n=188) and HC (n=64).

[b]CFS: collected for study.

[c]SVM: support vector machine.

[d]MFCC: Mel-frequency cepstral coefficients.

[e]TQWT: tunable Q-factor wavelet transform.

[f]UCI: University of California, Irvine.

[g]GB: gradient boosting.

[h]BLA: baseline acoustic.

[i]KNN: K-nearest neighbor.

[j]ANN: artificial neural network.

[k]WT: wavelet.

[l]NB: naïve Bayes.

[m]RF: random forest.

Discussion

Principal Findings

In this SLR, 10 years of research on ML techniques applied for diagnosing and monitoring voice-affecting disorders indicates an extended interest from many countries. It seems that researchers have focused mostly on the detection of 19 identified disorders with low number of individuals in data sets that lead to gaps identified as the main findings of this SLR. These are summarized below:

- Most studies aimed to perform a diagnostic test through the detection or classification of disorders, and only a few studies aimed to monitor a specific disorder.
- PD was the most investigated disorder among all 19 voice-affecting disorders.
- There was a broad interest from many counties.
- Data sets used in studies were unbalanced, and most studies collected their data without providing open access. Additionally, only 11/145 (7.6%) included studies considered using additional data in conjunction with voice features.
- All studies were cross-sectional.
- Accuracy was the most common metric for the overall performance evaluation.

The majority of the studies focused on the detection or classification of the 19 identified voice-affecting disorders through emerging ML techniques. However, it is important to also consider the need for continuous monitoring of these disorders to improve the quality of life for those affected. Another consequence of focusing solely on detection is that it may not provide enough information about the severity of the disorder, which is a vital measure for decision-making on treatment or determining correct dosage for medication. Therefore, to improve the applicability of findings in clinical practices, it may be beneficial to navigate the focus of research toward methods for monitoring the progression, which involve severity measures of voice-affecting disorders.

Verdolini et al [6] give an intuition that the 19 disorders identified in this SLR correspond only to a small number of voice-affecting disorders that have been studied in research. This small correspondence makes it troublesome to highlight the digital biomarkers that are specifically related to a single disorder, which is essential for distinguishing underlying conditions that lead to altered voice quality. To address this issue, it is worth extending the research to other voice-affecting disorders that have been underrepresented in previous studies. This would not only extend the number of disorders being studied but also allow for the identification of differences and similarities in terms of digital biomarkers or other features across a wider range of disorders. Exploring the differences and similarities between disorders, syndromes, and symptoms is also beneficial because some disorders can function as symptoms of other underlying conditions affecting voice production, that is, while depression can be considered a disorder in and of itself, it can also manifest as a symptom of PD [6].

Based on the origin of the publication and the origin of the data sets, a wide interest from many countries was observed. However, many countries conduct research on the same data sets, which can lead to both positive and negative results regarding the clinical applicability of outcomes. Concentration on a group of data sets may increase the performance of the ML technique for the represented input data attributes. By contrast, it may also introduce limitations for the nonrepresented or underrepresented data. For example, the UCI data set in Sakar et al [73] contains several voice recordings in Turkish; using this data set may give satisfying results for recordings in the same language, but using it on English recordings could be problematic. However, interest from many countries shows enormous potential for collecting more available data sets and generalized ML techniques.

A balanced data set means the numbers of samples are relatively equal between classes, giving equivalent contributions from all classes during training, which eventually improves the performance of the ML technique on new data. By contrast, imbalanced data can lead to bias. The results of our SLR show that using balanced data has not been considered in studies. As the voice is used as a medium to detect a disorder, it is important to consider the effect of linguistic diversity, gender, age, and other sociodemographic differences on the generalizability of a system. Training and testing an ML technique on balanced data offer higher reliability for use in clinical practices. Balancing data based on different characteristics may be another option for higher reliability (eg, only male or only female). The studies included in the analysis provided demographic information about their respective data sets. However, only a limited number of studies incorporated this information into the vocal feature set that use additional nonvoice data for training the ML models. Integrating the demographic data into the automated process of data set preparation could prove beneficial, as opposed to the manual preparation of data sets based on disparate attributes. Additionally, combining multiple sensory inputs along with vocal features may further enhance the performance of the ML algorithms. However, this practice appears to be infrequently observed in recent studies.

Results of this SLR showed that 70/145 (48.3%) studies collected data specific to the research without making them publicly available. It is observable that PD is one of the most investigated disorders. That might be a result of publicly available data obtained from the UCI Parkinson data set repository. It is worthwhile to extend publicly available data sources with varied voice-affecting disorders and features to preserve research reliability and homogeneity in the scope. Another aspect that would influence clinical applicability is the small number of participants being considered in the research. Increasing the number of participants might increase the reliability of ML techniques.

In SLRs, “longitudinal study” refers to a recurrent sample taken from the same participant over time, which is a way of following the progression and trend of a disorder that helps to identify the patterns and causal relationships. Conducting a longitudinal study may even help to reduce the confounding variables [191].

The absence of longitudinal studies makes it difficult to conduct an epidemiological trend analysis of time effects on digital features extracted for ML techniques to diagnose and monitor voice-affecting disorders. All included studies in this SLR considered cross-sectional analysis, which does not represent the possible divergence tied to the progression of a specific disorder and individual. Therefore, longitudinal studies are essential to discover the voice changes over time [191,192].

The majority of the studies chose to represent the performance in accuracy, specificity, and sensitivity metrics, which were tied to the overall classification performance in research on voice-affecting disorder diagnosis and monitoring. In this SLR, only 8 results indicated that gender and language diversity may affect the performance in terms of accuracy [63,89,92,97,114,170-172]. As none of the studies address the effect of unbalanced data on performance evaluation, it is noteworthy that the divergence in accuracy in those outcomes could be the effect of unbalanced data. However, different accuracy results may be due to many other aspects. Regardless of the employed ML technique, used features; number of features; the proportion between training, validation, test sets; and feature extraction techniques may also be a factor in deviating accuracy results.

Limitations

The decision to include studies published only in English is a risk of missing important evidence in other languages, which at the same time is an unavoidable limitation for the generalization of this SLR. Another factor that can be considered as a limitation is only including peer-reviewed studies, which do not consider conference papers. Relying on the conference papers can be problematic due to the limitations including the potential for incomplete or preliminary results [193]. Additionally, including low-quality studies may introduce a risk of bias, as these studies may have suffered from selective reporting bias. In the SLR presented herein, this risk was mitigated by checking beyond what is presented in the paper, that is, when the methodological information was referenced elsewhere, the authors checked and considered the referenced material when conducting the quality assessment. Additionally, during the screening phase, when the abstract did not contain the full information to fulfill the inclusion criteria, these pieces were marked as “maybe” cases that were checked further before being fully read.

Future Work

Underrepresented monitoring purposes, research on a low number of voice-affecting disorders, unbalanced data, limited public voice data, lack of longitudinal research, and performance evaluation without paying attention to diversities were 6 gaps addressed in this SLR, which may be considered in future research. We suggest the following:

- One research direction may be to include disorders that were underrepresented in the state of the art. It is essential to take the gaps into consideration, such as working with balanced and extended data sets, to generate more reliable results.
- Conducting cross-sectional and longitudinal studies to identify specific digital features that are associated with voice-affecting disorders can be beneficial for determining the severity of the disorder and monitoring it over time.
- Studying the effects of demographic characteristics, such as gender, age, linguistic factors, and other relevant additional data on the classification models may also provide insights for building more accurate ML techniques for specific disorders.

Conclusions

Through the methodology of an SLR, we identified 145 studies on the use of voice for diagnosing or prognosing disorders, by the means of ML algorithms. These studies were summarized in terms of many aspects, including disorders and conditions that affect the voice, characteristics of the input data, ML techniques used for voice-based diagnosis, and research interests from countries. The findings of this SLR indicated that most of the studies are concerned with the detection and classification of investigated disorders and conditions based on cross-sectional studies. This study also found gaps in the literature, such as the usage of unbalanced data sets, lack of longitudinal studies, research not addressing nonvoice data in the voice studies, and most voice-affecting disorders in the interest of this study being underrepresented in research. Research in the field of voice-based diagnostics with the utilization of ML is making the practical application of this technology in health care more achievable. The use of voice as a digital biomarker could open the possibilities to large population screening of many disorders in a low-cost, noninvasive, and scalable way. To implement such a system in a clinical setting, the exploration of unknown aspects is an essential process to proceed with. To do that, it is necessary to extend the research on all possible voice-affecting disorders and identify the nuances between all different voice-affecting disorders and their effect on vocal features. Currently, research in this field primarily focuses on detection using a limited number of participants. However, for more generalizable results in the future, research may not only consider increasing the participant numbers but also maintaining a balance among them and identifying the measures that can be used for monitoring purposes.

There is a broad research interest from many countries, which creates a potential for observing the effects of cultural and language differences on ML algorithms. However, contribution to data collection and increasing the size of available data with diverse characteristics are crucial steps that each country might consider.

Acknowledgments

The authors thank the Excellence Center at Linköping – Lund in Information Technology (ELLIIT) for funding and supporting this project.

Data Availability

A summary table of all included studies and extracted data is available as Multimedia Appendix 1.

Authors' Contributions

AI is the primary contributor to the study and manuscript, with involvement in all aspects. ALD assisted in the design of the study, simultaneous study selection, and revisions of the manuscript. PA contributed to the final revisions of the manuscript. JSB provided medical expertise in the field and assisted with revisions of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Summary of included studies.
[PDF File (Adobe PDF File), 225 KB-Multimedia Appendix 1]

Multimedia Appendix 2

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[DOCX File , 33 KB-Multimedia Appendix 2]

Multimedia Appendix 3

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 for Abstract checklist.
[DOCX File , 27 KB-Multimedia Appendix 3]

References

1. Zhang Z. Mechanics of human voice production and control. J Acoust Soc Am 2016 Oct;140(4):2614 [FREE Full text] [doi: 10.1121/1.4964509] [Medline: 27794319]
2. Merrill RM, Roy N, Lowe J. Voice-related symptoms and their effects on quality of life. Ann Otol Rhinol Laryngol 2013 Jun;122(6):404-411 [doi: 10.1177/000348941312200610] [Medline: 23837394]
3. Payten CL, Chiapello G, Weir KA, Madill CJ. Frameworks, Terminology and Definitions Used for the Classification of Voice Disorders: A Scoping Review. J Voice 2022 Mar 19:S0892-1997(22)00039-X [FREE Full text] [doi: 10.1016/j.jvoice.2022.02.009] [Medline: 35317970]
4. Payten C, Chiapello G, Weir K, Madill C. Terminology and frameworks used for the classification of voice disorders: a scoping review protocol. JBI Evid Synth 2021;19:454-462 [doi: 10.11124/jbies-20-00066]
5. Andrea M, Dias Ó, Andrea M, Figueira ML. Functional Voice Disorders: The Importance of the Psychologist in Clinical Voice Assessment. J Voice 2017 Jul;31(4):507.e13-507.e22 [doi: 10.1016/j.jvoice.2016.10.013] [Medline: 27876300]
6. Verdolini K, Rosen C, Branski R, editors. Classification Manual for Voice Disorders. New York, NY: Psychology Press; 2005.
7. Umeno H, Hyodo M, Haji T, Hara H, Imaizumi M, Ishige M, et al. A summary of the Clinical Practice Guideline for the Diagnosis and Management of Voice Disorders, 2018 in Japan. Auris Nasus Larynx 2020 Feb;47(1):7-17 [doi: 10.1016/j.anl.2019.09.004] [Medline: 31587820]
8. Cohen SM, Kim J, Roy N, Asche C, Courey M. Direct health care costs of laryngeal diseases and disorders. Laryngoscope 2012 Jul;122(7):1582-1588 [doi: 10.1002/lary.23189] [Medline: 22544473]
9. Califf RM. Biomarker definitions and their applications. Exp Biol Med (Maywood) 2018 Feb;243(3):213-221 [FREE Full text] [doi: 10.1177/1535370217750088] [Medline: 29405771]
10. Mayeux R. Biomarkers: potential uses and limitations. NeuroRx 2004 Apr;1(2):182-188 [FREE Full text] [doi: 10.1602/neurorx.1.2.182] [Medline: 15717018]
11. Babrak L, Menetski J, Rebhan M, Nisato G, Zinggeler M, Brasier N, et al. Traditional and Digital Biomarkers: Two Worlds Apart? Digit Biomark 2019;3(2):92-102 [FREE Full text] [doi: 10.1159/000502000] [Medline: 32095769]
12. Coravos A, Khozin S, Mandl KD. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. NPJ Digit Med 2019;2(1):14 [FREE Full text] [doi: 10.1038/s41746-019-0090-4] [Medline: 30868107]
13. Dorsey ER, Papapetropoulos S, Xiong M, Kieburtz K. The First Frontier: Digital Biomarkers for Neurodegenerative Disorders. Digit Biomark 2017;1(1):6-13 [FREE Full text] [doi: 10.1159/000477383] [Medline: 32095743]
14. Robin J, Harrison JE, Kaufman LD, Rudzicz F, Simpson W, Yancheva M. Evaluation of Speech-Based Digital Biomarkers: Review and Recommendations. Digit Biomark 2020;4(3):99-108 [FREE Full text] [doi: 10.1159/000510820] [Medline: 33251474]

15. Fagherazzi G, Fischer A, Ismael M, Despotovic V. Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice. Digit Biomark 2021;5(1):78-88 [FREE Full text] [doi: 10.1159/000515346] [Medline: 34056518]
16. Lin H, Karjadi C, Ang T, Prajakta J, McManus C, Alhanai T, et al. Identification of digital voice biomarkers for cognitive health. Explor Med 2020;1:406-417 [FREE Full text] [doi: 10.37349/emed.2020.00028] [Medline: 33665648]
17. Tracy JM, Özkanca Y, Atkins DC, Hosseini Ghomi R. Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson's disease. J Biomed Inform 2020 Apr;104:103362 [FREE Full text] [doi: 10.1016/j.jbi.2019.103362] [Medline: 31866434]
18. López-de-Ipiña K, Alonso JB, Travieso CM, Solé-Casals J, Egiraun H, Faundez-Zanuy M, et al. On the selection of non-invasive methods based on speech analysis oriented to automatic Alzheimer disease diagnosis. Sensors (Basel) 2013 May 21;13(5):6730-6745 [FREE Full text] [doi: 10.3390/s130506730] [Medline: 23698268]
19. Van Stan J, Mehta DD, Hillman RE. Recent Innovations in Voice Assessment Expected to Impact the Clinical Management of Voice Disorders. Perspect ASHA SIGs 2017 Jan;2(3):4-13 [doi: 10.1044/persp2.SIG3.4]
20. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019 Jan;25(1):24-29 [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
21. Shailaja K, Seetharamulu B, Jabbar M. Machine Learning in Healthcare: A Review. 2018 Presented at: Second International Conference on Electronics, Communication and Aerospace Technology (ICECA); March 29-31, 2018; Coimbatore, Tamil Nadu, India p. 910-914 [doi: 10.1109/ICECA.2018.8474918]
22. Garg A, Sharma P. Survey on acoustic modeling and feature extraction for speech recognition. 2016 Presented at: 3rd International Conference on Computing for Sustainable Global Development (INDIACom); March 16-18, 2016; New Delhi, India p. 2291-2295
23. Department OEE. G. Balekundri Institute of Technology, Belagavi-India, Kanabur V, Harakannanavar SS, Torse D. An Extensive Review of Feature Extraction Techniques, Challenges and Trends in Automatic Speech Recognition. Int J Image Graph Signal Process. 2019;11:12 [doi: 10.5815/ijigsp.2019.05.01]
24. Abdallah B, Abdallah A, Ratte S. Detecting depression in Alzheimer and MCI using artificial neural networks (ANN). 2021 Apr Presented at: International Conference on Data Science, E-learning and Information Systems 2021; April 5-7, 2021; Ma'an, Jordan p. 250-253 [doi: 10.1145/3460620.3460765]
25. Kurzekar PK, Deshmukh RR, Waghmare VB, Shrishrimal PP. A Comparative Study of Feature Extraction Techniques for Speech Recognition System. International Journal of Innovative Research in Science, Engineering and Technology 2014 Dec;3(12):18006-18016 [FREE Full text] [doi: 10.15680/IJIRSET.2014.0312034]
26. Syed SA, Rashid M, Hussain S. Meta-analysis of voice disorders databases and applied machine learning techniques. Math Biosci Eng 2020 Nov 11;17(6):7958-7979 [FREE Full text] [doi: 10.3934/mbe.2020404] [Medline: 33378928]
27. Pahuja G, Nagabhushan TN. A Comparative Study of Existing Machine Learning Approaches for Parkinson's Disease Detection. IETE Journal of Research 2018 Oct 22;67(1):4-14 [doi: 10.1080/03772063.2018.1531730]
28. Petti U, Baker S, Korhonen A. A systematic literature review of automatic Alzheimer's disease detection from speech and language. J Am Med Inform Assoc 2020 Nov 01;27(11):1784-1797 [FREE Full text] [doi: 10.1093/jamia/ocaa174] [Medline: 32929494]
29. Saravanan S, Ramkumar K, Adalarasu K, Sivanandam V, Kumar SR, Stalin S, et al. A Systematic Review of Artificial Intelligence (AI) Based Approaches for the Diagnosis of Parkinson's Disease. Arch Computat Methods Eng 2022 Jan 20;29(6):3639-3653 [doi: 10.1007/s11831-022-09710-1]
30. Bettany-Saltikov J. How to do a Systematic Literature Review in Nursing: A Step-by-Step Guide. London, UK: McGraw-Hill Education; 2016.
31. Idrisoglu A. Protocol of the Systematic Literature Review. GitHub. 2023 Jan 28. URL: https://github.com/AIITPlanet/Protocol [accessed 2023-07-11]
32. Considine J, Shaban RZ, Fry M, Curtis K. Evidence based emergency nursing: Designing a research question and searching the literature. Int Emerg Nurs 2017 May;32:78-82 [doi: 10.1016/j.ienj.2017.02.001] [Medline: 28233626]
33. Butler A, Hall H, Copnell B. A Guide to Writing a Qualitative Systematic Review Protocol to Enhance Evidence-Based Practice in Nursing and Health Care. Worldviews Evid Based Nurs 2016 Jun 20;13(3):241-249 [doi: 10.1111/wvn.12134] [Medline: 26790142]
34. Schiavenato M, Chu F. PICO: What it is and what it is not. Nurse Educ Pract 2021 Oct;56:103194 [doi: 10.1016/j.nepr.2021.103194] [Medline: 34534728]
35. Mueen Ahmed KK, Dhubaib BEA. Zotero: A bibliographic assistant to researcher. Journal of Pharmacology and Pharmacotherapeutics 2022 Apr 11;2(4):304-305 [doi: 10.4103/0976-500x.85940]
36. Kitchenham B, Charters S. Guidelines for Performing Systematic Literature Reviews in Software Engineering (Technical Report EBSE-2007-01). University of Auckland. 2007. URL: https://www.cs.auckland.ac.nz/~norsaremah/2007%20Guidelines%20for%20performing%20SLR%20in%20SE%20v2.3.pdf [accessed 2023-07-11]
37. Pianosi F, Beven K, Freer J, Hall JW, Rougier J, Stephenson DB, et al. Sensitivity analysis of environmental models: A systematic review with practical workflow. Environmental Modelling & Software 2016 May;79:214-232 [doi: 10.1016/j.envsoft.2016.02.008]

38. Christopher Frey H, Patil SR. Identification and Review of Sensitivity Analysis Methods. Risk Analysis 2002 Jun;22(3):553-578 [doi: 10.1111/0272-4332.00039]
39. PRISMA 2020 Checklist. PRISMA. 2020. URL: http://www.prisma-statement.org/PRISMAStatement/Checklist [accessed 2023-03-28]
40. Tuncer T, Dogan S. A novel octopus based Parkinson's disease and gender recognition method using vowels. Applied Acoustics 2019 Dec;155:75-83 [doi: 10.1016/j.apacoust.2019.05.019]
41. Tuncer T, Dogan S, Acharya UR. Automated detection of Parkinson's disease using minimum average maximum tree and singular value decomposition method with vowels. Biocybernetics and Biomedical Engineering 2020 Jan;40(1):211-220 [doi: 10.1016/j.bbe.2019.05.006]
42. Gunduz H. Deep Learning-Based Parkinson's Disease Classification Using Vocal Feature Sets. IEEE Access 2019;7:115540-115551 [doi: 10.1109/access.2019.2936564]
43. Gunduz H. An efficient dimensionality reduction method using filter-based feature selection and variational autoencoders on Parkinson's disease classification. Biomedical Signal Processing and Control 2021 Apr;66:102452 [doi: 10.1016/j.bspc.2021.102452]
44. Lamba R, Gulati T, Jain A. A Hybrid Feature Selection Approach for Parkinson's Detection Based on Mutual Information Gain and Recursive Feature Elimination. Arab J Sci Eng 2022 Jan 18;47(8):10263-10276 [doi: 10.1007/s13369-021-06544-0]
45. Lamba R, Gulati T, Jain A. An Intelligent System for Parkinson's Diagnosis Using Hybrid Feature Selection Approach. Int J Softw Innov 2022;10(1):1-13 [doi: 10.4018/ijsi.292027]
46. 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 Mar 10;9(3):e21331 [FREE Full text] [doi: 10.2196/21331] [Medline: 33688838]
47. Tena A, Clarià F, Solsona F, Povedano M. Detecting Bulbar Involvement in Patients with Amyotrophic Lateral Sclerosis Based on Phonatory and Time-Frequency Features. Sensors (Basel) 2022 Feb 02;22(3):1137 [FREE Full text] [doi: 10.3390/s22031137] [Medline: 35161881]
48. Sakar CO, Serbes G, Gunduz A, Tunc HC, Nizam H, Sakar BE, et al. A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform. Applied Soft Computing 2019 Jan;74:255-263 [doi: 10.1016/j.asoc.2018.10.022]
49. Moro-Velazquez L, Gomez-Garcia JA, Godino-Llorente JI, Villalba J, Rusz J, Shattuck-Hufnagel S, et al. A forced gaussians based methodology for the differential evaluation of Parkinson's Disease by means of speech processing. Biomedical Signal Processing and Control 2019 Feb;48:205-220 [doi: 10.1016/j.bspc.2018.10.020]
50. Meghraoui D, Boudraa B, Merazi-Meksen T, Gómez Vilda P. A novel pre-processing technique in pathologic voice detection: Application to Parkinson's disease phonation. Biomedical Signal Processing and Control 2021 Jul;68:102604 [doi: 10.1016/j.bspc.2021.102604]
51. Quan C, Ren K, Luo Z. A Deep Learning Based Method for Parkinson's Disease Detection Using Dynamic Features of Speech. IEEE Access 2021;9:10239-10252 [doi: 10.1109/ACCESS.2021.3051432]
52. Goyal J, Khandnor P, Aseri TC. A Hybrid Approach for Parkinson's Disease diagnosis with Resonance and Time-Frequency based features from Speech signals. Expert Systems with Applications 2021 Nov;182:115283 [doi: 10.1016/j.eswa.2021.115283]
53. Cantürk I, Karabiber F. A Machine Learning System for the Diagnosis of Parkinson's Disease from Speech Signals and Its Application to Multiple Speech Signal Types. Arab J Sci Eng 2016;41:5059 [doi: 10.1007/s13369-016-2206-3]
54. Carrón J, Campos-Roca Y, Madruga M, Pérez CJ. A mobile-assisted voice condition analysis system for Parkinson's disease: assessment of usability conditions. Biomed Eng Online 2021 Nov 21;20(1):114 [FREE Full text] [doi: 10.1186/s12938-021-00951-y] [Medline: 34802448]
55. Yücelbaş C. A new approach: information gain algorithm-based k-nearest neighbors hybrid diagnostic system for Parkinson's disease. Phys Eng Sci Med 2021 Jun;44(2):511-524 [doi: 10.1007/s13246-021-01001-6] [Medline: 33852120]
56. Cai Z, Gu J, Chen H. A New Hybrid Intelligent Framework for Predicting Parkinson's Disease. IEEE Access 2017;5:17188-17200 [doi: 10.1109/ACCESS.2017.2741521]
57. Jahnavi BS, Supraja BS, Lalitha S. A vital neurodegenerative disorder detection using speech cues. IFS 2020 May 29;38(5):6337-6345 [doi: 10.3233/JIFS-179714]
58. Zhang L, Qu Y, Jin B, Jing L, Gao Z, Liang Z. An Intelligent Mobile-Enabled System for Diagnosing Parkinson Disease: Development and Validation of a Speech Impairment Detection System. JMIR Med Inform 2020 Sep 16;8(9):e18689 [FREE Full text] [doi: 10.2196/18689] [Medline: 32936086]
59. Cai Z, Gu J, Wen C, Zhao D, Huang C, Huang H, et al. An Intelligent Parkinson's Disease Diagnostic System Based on a Chaotic Bacterial Foraging Optimization Enhanced Fuzzy KNN Approach. Comput Math Methods Med 2018;2018:2396952 [FREE Full text] [doi: 10.1155/2018/2396952] [Medline: 30034509]
60. Rizvi DR, Nissar I, Masood S, Ahmed M, Ahmad F. An LSTM based deep learning model for voice-based detection of Parkinson's disease. Int J Adv Sci Technol 2020;29(5s):337-343 [FREE Full text]
61. Olivares R, Munoz R, Soto R, Crawford B, Cárdenas D, Ponce A, et al. An Optimized Brain-Based Algorithm for Classifying Parkinson's Disease. Applied Sciences 2020 Mar 06;10(5):1827 [doi: 10.3390/app10051827]

62. Tougui I, Jilbab A, Mhamdi JE. Analysis of Smartphone Recordings in Time, Frequency, and Cepstral Domains to Classify Parkinson's Disease. Healthc Inform Res 2020 Oct;26(4):274-283 [FREE Full text] [doi: 10.4258/hir.2020.26.4.274] [Medline: 33190461]
63. Solana-Lavalle G, Rosas-Romero R. Analysis of voice as an assisting tool for detection of Parkinson's disease and its subsequent clinical interpretation. Biomedical Signal Processing and Control 2021 Apr;66:102415 [doi: 10.1016/j.bspc.2021.102415]
64. Pramanik M, Pradhan R, Nandy P, Qaisar SM, Bhoi AK. Assessment of Acoustic Features and Machine Learning for Parkinson's Detection. J Healthc Eng 2021;2021:9957132 [FREE Full text] [doi: 10.1155/2021/9957132] [Medline: 34471507]
65. Ali L, Zhu C, Zhang Z, Liu Y. Automated Detection of Parkinson's Disease Based on Multiple Types of Sustained Phonations Using Linear Discriminant Analysis and Genetically Optimized Neural Network. IEEE J Transl Eng Health Med 2019;7:2000410 [FREE Full text] [doi: 10.1109/JTEHM.2019.2940900] [Medline: 32166050]
66. Braga D, Madureira AM, Coelho L, Ajith R. Automatic detection of Parkinson's disease based on acoustic analysis of speech. Engineering Applications of Artificial Intelligence 2019 Jan;77:148-158 [doi: 10.1016/j.engappai.2018.09.018]
67. Novotny M, Rusz J, Cmejla R, Ruzicka E. Automatic Evaluation of Articulatory Disorders in Parkinson's Disease. IEEE/ACM Trans. Audio Speech Lang. Process 2014 Sep;22(9):1366-1378 [doi: 10.1109/TASLP.2014.2329734]
68. Solana-Lavalle G, Galán-Hernández JC, Rosas-Romero R. Automatic Parkinson disease detection at early stages as a pre-diagnosis tool by using classifiers and a small set of vocal features. Biocybernetics and Biomedical Engineering 2020 Jan;40(1):505-516 [doi: 10.1016/j.bbe.2020.01.003]
69. Zhang HH, Yang L, Liu Y, Wang P, Yin J, Li Y, et al. Classification of Parkinson's disease utilizing multi-edit nearest-neighbor and ensemble learning algorithms with speech samples. Biomed Eng Online 2016 Nov 16;15(1):122 [FREE Full text] [doi: 10.1186/s12938-016-0242-6] [Medline: 27852279]
70. Khan T, Westin J, Dougherty M. Classification of speech intelligibility in Parkinson's disease. Biocybernetics and Biomedical Engineering 2014;34(1):35-45 [doi: 10.1016/j.bbe.2013.10.003]
71. Berus L, Klancnik S, Brezocnik M, Ficko M. Classifying Parkinson's Disease Based on Acoustic Measures Using Artificial Neural Networks. Sensors (Basel) 2018 Dec 20;19(1):16 [FREE Full text] [doi: 10.3390/s19010016] [Medline: 30577548]
72. García AM, Arias-Vergara TC, C Vasquez-Correa J, Nöth E, Schuster M, Welch AE, et al. Cognitive Determinants of Dysarthria in Parkinson's Disease: An Automated Machine Learning Approach. Mov Disord 2021 Dec;36(12):2862-2873 [doi: 10.1002/mds.28751] [Medline: 34390508]
73. Sakar BE, Isenkul M, Sakar CO, Sertbas A, Gurgen F, Delil S, et al. Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. IEEE J Biomed Health Inform 2013 Jul;17(4):828-834 [doi: 10.1109/JBHI.2013.2245674] [Medline: 25055311]
74. Viswanathan R, Arjunan SP, Bingham A, Jelfs B, Kempster P, Raghav S, et al. Complexity Measures of Voice Recordings as a Discriminative Tool for Parkinson's Disease. Biosensors (Basel) 2019 Dec 20;10(1):10 [FREE Full text] [doi: 10.3390/bios10010001] [Medline: 31861890]
75. Hireš M, Gazda M, Drotár P, Pah ND, Motin MA, Kumar DK. Convolutional neural network ensemble for Parkinson's disease detection from voice recordings. Comput Biol Med 2022 Feb;141:105021 [doi: 10.1016/j.compbiomed.2021.105021] [Medline: 34799077]
76. Majda-Zdancewicz E, Potulska-Chromik A, Jakubowski J, Nojszewska M, Kostera-Pruszczyk A. Deep learning vs feature engineering in the assessment of voice signals for diagnosis in Parkinson's disease. Bull Pol Acad Sci Tech Sci 2021;69(3):e137347 [FREE Full text] [doi: 10.24425/bpasts.2021.137347]
77. Ozkanca Y, Öztürk MG, Ekmekci M, Atkins D, Demiroglu C, Ghomi RH. Depression Screening from Voice Samples of Patients Affected by Parkinson's Disease. Digit Biomark 2019 Jun 12;3(2):72-82 [FREE Full text] [doi: 10.1159/000500354] [Medline: 31872172]
78. Rahman W, Lee S, Islam MS, Antony VN, Ratnu H, Ali MR, et al. Detecting Parkinson Disease Using a Web-Based Speech Task: Observational Study. J Med Internet Res 2021 Oct 19;23(10):e26305 [FREE Full text] [doi: 10.2196/26305] [Medline: 34665148]
79. Almeida JS, Rebouças Filho P, Carneiro T, Wei W, Damaševičius R, Maskeliūnas R, et al. Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques. Pattern Recognition Letters 2019 Jul;125:55-62 [doi: 10.1016/j.patrec.2019.04.005]
80. Benba A, Jilbab A, Hammouch A. Detecting multiple system atrophy, Parkinson and other neurological disorders using voice analysis. Int J Speech Technol 2017 Mar 4;20(2):281-288 [doi: 10.1007/s10772-017-9404-6]
81. Arora S, Baghai-Ravary L, Tsanas A. Developing a large scale population screening tool for the assessment of Parkinson's disease using telephone-quality voice. J Acoust Soc Am 2019 May;145(5):2871 [FREE Full text] [doi: 10.1121/1.5100272] [Medline: 31153319]
82. Yang S, Zheng F, Luo X, Cai S, Wu Y, Liu K, et al. Effective dysphonia detection using feature dimension reduction and kernel density estimation for patients with Parkinson's disease. PLoS One 2014;9(2):e88825 [FREE Full text] [doi: 10.1371/journal.pone.0088825] [Medline: 24586406]

83. Oung QW, Muthusamy H, Basah SN, Lee H, Vijean V. Empirical Wavelet Transform Based Features for Classification of Parkinson's Disease Severity. J Med Syst 2017 Dec 29;42(2):29 [doi: 10.1007/s10916-017-0877-2] [Medline: 29288342]
84. Haq AU, Li JP, Memon MH, khan J, Malik A, Ahmad T, et al. Feature Selection Based on L1-Norm Support Vector Machine and Effective Recognition System for Parkinson's Disease Using Voice Recordings. IEEE Access 2019;7:37718-37734 [doi: 10.1109/ACCESS.2019.2906350]
85. Rusz J, Novotny M, Hlavnicka J, Tykalova T, Ruzicka E. High-accuracy voice-based classification between patients with Parkinson's disease and other neurological diseases may be an easy task with inappropriate experimental design. IEEE Trans Neural Syst Rehabil Eng 2017 Aug;25(8):1319-1321 [doi: 10.1109/TNSRE.2016.2621885] [Medline: 28113773]
86. Klempír O, Krupicka R. Machine learning using speech utterances for Parkinson disease detection. Lekar a Technika 2018 Jan;48(2):66-71 [FREE Full text]
87. Alhussein M. Monitoring Parkinson's Disease in Smart Cities. IEEE Access 2017;5:19835-19841 [doi: 10.1109/ACCESS.2017.2748561]
88. Tsanas A, Little MA, McSharry PE, Spielman J, Ramig LO. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Trans Biomed Eng 2012 May;59(5):1264-1271 [doi: 10.1109/TBME.2012.2183367] [Medline: 22249592]
89. Gómez-Vilda P, Mekyska J, Ferrández JM, Palacios-Alonso D, Gómez-Rodellar A, Rodellar-Biarge V, et al. Parkinson Disease Detection from Speech Articulation Neuromechanics. Front Neuroinform 2017;11:56 [FREE Full text] [doi: 10.3389/fninf.2017.00056] [Medline: 28970792]
90. Zhang T, Zhang Y, Sun H, Shan H. Parkinson disease detection using energy direction features based on EMD from voice signal. Biocybernetics and Biomedical Engineering 2021 Jan;41(1):127-141 [doi: 10.1016/j.bbe.2020.12.009]
91. Karan B, Sahu SS, Mahto K. Parkinson disease prediction using intrinsic mode function based features from speech signal. Biocybernetics and Biomedical Engineering 2020 Jan;40(1):249-264 [doi: 10.1016/j.bbe.2019.05.005]
92. Laganas C, Iakovakis D, Hadjidimitriou SK, Charisis V, Dias SB, Bostantzopoulou S, et al. Parkinson's Disease Detection Based on Running Speech Data From Phone Calls. IEEE Trans Biomed Eng 2022 May;69(5):1573-1584 [doi: 10.1109/TBME.2021.3116935] [Medline: 34596531]
93. Bchir O. Parkinson's Disease Classification using Gaussian Mixture Models with Relevance Feature Weights on Vocal Feature Sets. Int J Adv Comput Sci Appl 2020;11(4):413-419 [FREE Full text] [doi: 10.14569/IJACSA.2020.0110456]
94. Rahman A, Rizvi SS, Khan A, Abbasi AA, Khan SU, Chung TS. Parkinson's Disease Diagnosis in Cepstral Domain Using MFCC and Dimensionality Reduction with SVM Classifier. Mobile Information Systems 2021;2021:1-10 [FREE Full text] [doi: 10.1155/2021/8822069]
95. Fujita T, Luo Z, Quan C, Mori K, Cao S. Performance Evaluation of RNN with Hyperbolic Secant in Gate Structure through Application of Parkinson's Disease Detection. Applied Sciences 2021 May 11;11(10):4361 [doi: 10.3390/app11104361]
96. Vital TPR, Nayak J, Naik B, Jayaram D. Probabilistic Neural Network-based Model for Identification of Parkinson's Disease by using Voice Profile and Personal Data. Arab J Sci Eng 2021 Jan 03;46(4):3383-3407 [doi: 10.1007/s13369-020-05080-7]
97. Azadi H, Akbarzadeh-T M, Kobravi HR, Shoeibi A. Robust Voice Feature Selection Using Interval Type-2 Fuzzy AHP for Automated Diagnosis of Parkinson's Disease. IEEE/ACM Trans. Audio Speech Lang. Process 2021;29:2792-2802 [doi: 10.1109/TASLP.2021.3097215]
98. Amato F, Borzi L, Olmo G, Artusi CA, Imbalzano G, Lopiano L. Speech Impairment in Parkinson's Disease: Acoustic Analysis of Unvoiced Consonants in Italian Native Speakers. IEEE Access 2021;9:166370-166381 [doi: 10.1109/ACCESS.2021.3135626]
99. Hoq M, Uddin MN, Park SB. Vocal Feature Extraction-Based Artificial Intelligent Model for Parkinson's Disease Detection. Diagnostics (Basel) 2021 Jun 11;11(6):11 [FREE Full text] [doi: 10.3390/diagnostics11061076] [Medline: 34208330]
100. Benba A, Jilbab A, Hammouch A. Voice assessments for detecting patients with Parkinson's diseases using PCA and NPCA. Int J Speech Technol 2016 Sep 3;19(4):743-754 [doi: 10.1007/s10772-016-9367-z]
101. Jeancolas L, Petrovska-Delacrétaz D, Mangone G, Benkelfat BE, Corvol JC, Vidailhet M, et al. X-Vectors: New Quantitative Biomarkers for Early Parkinson's Disease Detection From Speech. Front Neuroinform 2021;15:578369 [FREE Full text] [doi: 10.3389/fninf.2021.578369] [Medline: 33679361]
102. Saloni S, Sharma R, Gupta A. Human voice waveform analysis for categorization of healthy and Parkinson subjects. International Journal of Healthcare Information Systems and Informatics 2016;11:21-35 [doi: 10.4018/ijhisi.2016010102]
103. Nasreen S, Rohanian M, Hough J, Purver M. Alzheimer's Dementia Recognition From Spontaneous Speech Using Disfluency and Interactional Features. Front. Comput. Sci 2021 Jun 18;3:3 [doi: 10.3389/fcomp.2021.640669]
104. Gonzalez-Moreira E, Torres-Boza D, Kairuz HA, Ferrer C, Garcia-Zamora M, Espinoza-Cuadros F, et al. Automatic Prosodic Analysis to Identify Mild Dementia. Biomed Res Int 2015;2015:916356 [FREE Full text] [doi: 10.1155/2015/916356] [Medline: 26558287]
105. Shimoda A, Li Y, Hayashi H, Kondo N. Dementia risks identified by vocal features via telephone conversations: A novel machine learning prediction model. PLoS One 2021;16(7):e0253988 [FREE Full text] [doi: 10.1371/journal.pone.0253988] [Medline: 34260593]

106. Tanaka H, Adachi H, Ukita N, Ikeda M, Kazui H, Kudo T, et al. Detecting Dementia Through Interactive Computer Avatars. IEEE J Transl Eng Health Med 2017;5:2200111 [FREE Full text] [doi: 10.1109/JTEHM.2017.2752152] [Medline: 29018636]
107. Khodabakhsh A, Yesil F, Guner E, Demiroglu C. Evaluation of linguistic and prosodic features for detection of Alzheimer's disease in Turkish conversational speech. J AUDIO SPEECH MUSIC PROC 2015 Mar 25;2015(1):9 [doi: 10.1186/s13636-015-0052-y]
108. Nasrolahzadeh M, Mohammadpoory Z, Haddadnia J. Higher-order spectral analysis of spontaneous speech signals in Alzheimer's disease. Cogn Neurodyn 2018 Dec;12(6):583-596 [FREE Full text] [doi: 10.1007/s11571-018-9499-8] [Medline: 30483366]
109. Fraser KC, Meltzer JA, Rudzicz F. Linguistic Features Identify Alzheimer's Disease in Narrative Speech. J Alzheimers Dis 2016;49(2):407-422 [doi: 10.3233/JAD-150520] [Medline: 26484921]
110. Guo Z, Ling Z, Li Y. Detecting Alzheimer's Disease from Continuous Speech Using Language Models. J Alzheimers Dis 2019;70(4):1163-1174 [doi: 10.3233/JAD-190452] [Medline: 31322577]
111. Bourouhou A, Jilbab A, Nacir C, Hammouch A. Classification of cardiovascular disease using dysphonia measurement in speech. Diagnostyka 2021;22(1):31-37 [FREE Full text] [doi: 10.29354/diag/132586]
112. Toth L, Hoffmann I, Gosztolya G, Vincze V, Szatloczki G, Banreti Z, et al. A Speech Recognition-based Solution for the Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech. Curr Alzheimer Res 2018;15(2):130-138 [FREE Full text] [doi: 10.2174/1567205014666171121114930] [Medline: 29165085]
113. Nagumo R, Zhang Y, Ogawa Y, Hosokawa M, Abe K, Ukeda T, et al. Automatic Detection of Cognitive Impairments through Acoustic Analysis of Speech. CAR 2020 Mar 20;17(1):60-68 [FREE Full text] [doi: 10.2174/1567205017666200213094513] [Medline: 32053074]
114. König A, Linz N, Zeghari R, Klinge X, Tröger J, Alexandersson J, et al. Detecting Apathy in Older Adults with Cognitive Disorders Using Automatic Speech Analysis. J Alzheimers Dis 2019;69(4):1183-1193 [doi: 10.3233/JAD-181033] [Medline: 31127764]
115. König A, Linz N, Tröger J, Wolters M, Alexandersson J, Robert P. Fully Automatic Speech-Based Analysis of the Semantic Verbal Fluency Task. Dement Geriatr Cogn Disord 2018;45(3-4):198-209 [doi: 10.1159/000487852] [Medline: 29886493]
116. Wang T, Hong Y, Wang Q, Su R, Ng ML, Xu J, et al. Identification of Mild Cognitive Impairment Among Chinese Based on Multiple Spoken Tasks. JAD 2021 Jun 29;82(1):185-204 [doi: 10.3233/jad-201387]
117. Maskeliūnas R, Damaševičius R, Kulikajevas A, Padervinskis E, Pribuišis K, Uloza V. A Hybrid U-Lossian Deep Learning Network for Screening and Evaluating Parkinson's Disease. Applied Sciences 2022 Nov 15;12(22):11601 [doi: 10.3390/app122211601]
118. Lamba R, Gulati T, Jain A, Rani P. A Speech-Based Hybrid Decision Support System for Early Detection of Parkinson's Disease. Arab J Sci Eng 2022 Sep 12;48(2):2247-2260 [doi: 10.1007/s13369-022-07249-8]
119. Dao SVT, Yu Z, Tran LV, Phan PNK, Huynh TTM, Le TM. An Analysis of Vocal Features for Parkinson's Disease Classification Using Evolutionary Algorithms. Diagnostics (Basel) 2022 Aug 16;12(8):12 [FREE Full text] [doi: 10.3390/diagnostics12081980] [Medline: 36010330]
120. Barukab O, Ahmad A, Khan T, Thayyil Kunhumuhammed MR. Analysis of Parkinson's Disease Using an Imbalanced-Speech Dataset by Employing Decision Tree Ensemble Methods. Diagnostics (Basel) 2022 Nov 30;12(12):12 [FREE Full text] [doi: 10.3390/diagnostics12123000] [Medline: 36553007]
121. Kaya D. Automated gender‐Parkinson's disease detection at the same time via a hybrid deep model using human voice. Concurrency and Computation 2022 Aug 23;34(26):34 [doi: 10.1002/cpe.7289]
122. Hawi S, Alhozami J, AlQahtani R, AlSafran D, Alqarni M, Sahmarany L. Automatic Parkinson's disease detection based on the combination of long-term acoustic features and Mel frequency cepstral coefficients (MFCC). Biomedical Signal Processing and Control 2022 Sep;78:104013 [doi: 10.1016/j.bspc.2022.104013]
123. Quan C, Ren K, Luo Z, Chen Z, Ling Y. End-to-end deep learning approach for Parkinson's disease detection from speech signals. Biocybernetics and Biomedical Engineering 2022 Apr;42(2):556-574 [doi: 10.1016/j.bbe.2022.04.002]
124. El-Habbak O, Abdelalim A, Mohamed N, Abd-Elaty H, Hammouda M, Mohamed Y. Enhancing Parkinson's disease diagnosis accuracy through speech signal algorithm modeling. CMC-Computers Materials & Continua 2022;70:2953-2969 [doi: 10.32604/cmc.2022.020109]
125. Xie JC, Gan Y, Liang P, Lan R, Gao H. Exploring robust computer-aided diagnosis of Parkinson's disease based on various voice signals. Front. Phys 2022 Nov 3;10:10 [doi: 10.3389/fphy.2022.1048833]
126. Abdul Gafoor S, Theagarajan P. Intelligent approach of score-based artificial fish swarm algorithm (SAFSA) for Parkinson's disease diagnosis. IJICC 2022 Jan 17;15(4):540-561 [doi: 10.1108/IJICC-10-2021-0226]
127. Senturk Z. Layer recurrent neural network-based diagnosis of Parkinson's disease using voice features. Biomed Tech (Berl) 2022 Aug 26;67(4):249-266 [doi: 10.1515/bmt-2022-0022] [Medline: 35659859]
128. Tougui I, Jilbab A, Mhamdi JE. Machine Learning Smart System for Parkinson Disease Classification Using the Voice as a Biomarker. Healthc Inform Res 2022 Jul;28(3):210-221 [FREE Full text] [doi: 10.4258/hir.2022.28.3.210] [Medline: 35982595]

129. Almasoud A, Eisa T, Al-Wesabi F, Elsafi A, Al DM, Yaseen I. Parkinson's Detection Using RNN-Graph-LSTM with Optimization Based on Speech Signals. CMC-Computers, Materials & Continua 2022;72(1):871-886 [FREE Full text] [doi: 10.32604/cmc.2022.024596]
130. Motin MA, Pah ND, Raghav S, Kumar DK. Parkinson's Disease Detection Using Smartphone Recorded Phonemes in Real World Conditions. IEEE Access 2022;10:97600-97609 [doi: 10.1109/ACCESS.2022.3203973]
131. Yu Q, Zou X, Quan F, Dong Z, Yin H, Liu J, et al. Parkinson's disease patients with freezing of gait have more severe voice impairment than non-freezers during "ON state". J Neural Transm (Vienna) 2022 Mar;129(3):277-286 [doi: 10.1007/s00702-021-02458-1] [Medline: 34989833]
132. Pah ND, Motin MA, Kumar DK. Phonemes based detection of parkinson's disease for telehealth applications. Sci Rep 2022 Jun 11;12(1):9687 [FREE Full text] [doi: 10.1038/s41598-022-13865-z] [Medline: 35690657]
133. Liu W, Liu J, Peng T, Wang G, Balas VE, Geman O, et al. Prediction of Parkinson's disease based on artificial neural networks using speech datasets. J Ambient Intell Human Comput 2022 Apr 12;1(1):e1 [doi: 10.1007/s12652-022-03825-w]
134. Khaskhoussy R, Ayed Y. Speech processing for early Parkinson's disease diagnosis: machine learning and deep learning-based approach. Soc. Netw. Anal. Min 2022 Jul 04;12(1):12 [doi: 10.1007/s13278-022-00905-9]
135. Pramanik M, Pradhan R, Nandy P, Bhoi AK, Barsocchi P. The ForEx++ based decision tree ensemble approach for robust detection of Parkinson's disease. J Ambient Intell Human Comput 2022 Feb 10;1(1):e1 [doi: 10.1007/s12652-022-03719-x]
136. Bertini F, Allevi D, Lutero G, Calzà L, Montesi D. An automatic Alzheimer's disease classifier based on spontaneous spoken English. Computer Speech & Language 2022 Mar;72:101298 [doi: 10.1016/j.csl.2021.101298]
137. Agbavor F, Liang H. Artificial Intelligence-Enabled End-To-End Detection and Assessment of Alzheimer's Disease Using Voice. Brain Sci 2022 Dec 23;13(1):13 [FREE Full text] [doi: 10.3390/brainsci13010028] [Medline: 36672010]
138. Pérez-Toro PA, Rodríguez-Salas D, Arias-Vergara T, Klumpp P, Schuster M, Nöth E, et al. Interpreting acoustic features for the assessment of Alzheimer's disease using ForestNet. Smart Health 2022 Dec;26:100347 [doi: 10.1016/j.smhl.2022.100347]
139. Hason L, Krishnan S. Spontaneous speech feature analysis for alzheimer's disease screening using a random forest classifier. Front Digit Health 2022;4:901419 [FREE Full text] [doi: 10.3389/fdgth.2022.901419] [Medline: 36465088]
140. Lin Y, Liyanage BN, Sun Y, Lu T, Zhu Z, Liao Y, et al. A deep learning-based model for detecting depression in senior population. Front Psychiatry 2022;13:1016676 [FREE Full text] [doi: 10.3389/fpsyt.2022.1016676] [Medline: 36419976]
141. Othmani A, Zeghina AO, Muzammel M. A Model of Normality Inspired Deep Learning Framework for Depression Relapse Prediction Using Audiovisual Data. Comput Methods Programs Biomed 2022 Nov;226:107132 [doi: 10.1016/j.cmpb.2022.107132] [Medline: 36183638]
142. Hashim NNWN, Basri NA, Ezzi MAEA, Hashim NMHN. Comparison of classifiers using robust features for depression detection on Bahasa Malaysia speech. IAES Int J Artif Intell.. IAES Int J Artif Intell 2022;11:238-253 [doi: 10.11591/ijai.v11.i1.pp238-253]
143. Sharma G, Umapathy K, Krishnan S. Audio texture analysis of COVID-19 cough, breath, and speech sounds. Biomed Signal Process Control 2022 Jul;76:103703 [FREE Full text] [doi: 10.1016/j.bspc.2022.103703] [Medline: 35464186]
144. Dar JA, Srivastava KK, Ahmed Lone S. Design and development of hybrid optimization enabled deep learning model for COVID-19 detection with comparative analysis with DCNN, BIAT-GRU, XGBoost. Comput Biol Med 2022 Oct 03;150:106123 [FREE Full text] [doi: 10.1016/j.compbiomed.2022.106123] [Medline: 36228465]
145. Dang T, Han J, Xia T, Spathis D, Bondareva E, Siegele-Brown C, et al. Exploring Longitudinal Cough, Breath, and Voice Data for COVID-19 Progression Prediction via Sequential Deep Learning: Model Development and Validation. J Med Internet Res 2022 Jun 21;24(6):e37004 [FREE Full text] [doi: 10.2196/37004] [Medline: 35653606]
146. Dash TK, Chakraborty C, Mahapatra S, Panda G. Gradient Boosting Machine and Efficient Combination of Features for Speech-Based Detection of COVID-19. IEEE J. Biomed. Health Inform 2022 Nov;26(11):5364-5371 [FREE Full text] [doi: 10.1109/JBHI.2022.3197910] [Medline: 35947565]
147. Albadr MAA, Tiun S, Ayob M, Al-Dhief FT. Particle Swarm Optimization-Based Extreme Learning Machine for COVID-19 Detection. Cognit Comput 2022 Oct 12:1-16 [FREE Full text] [doi: 10.1007/s12559-022-10063-x] [Medline: 36247809]
148. Ye W, Jiang Z, Li Q, Liu Y, Mou Z. A hybrid model for pathological voice recognition of post-stroke dysarthria by using 1DCNN and double-LSTM networks. Applied Acoustics 2022 Aug;197:108934 [doi: 10.1016/j.apacoust.2022.108934]
149. Svoboda E, Bořil T, Rusz J, Tykalová T, Horáková D, Guttmann CRG, et al. Assessing clinical utility of machine learning and artificial intelligence approaches to analyze speech recordings in multiple sclerosis: A pilot study. Comput Biol Med 2022 Sep;148:105853 [doi: 10.1016/j.compbiomed.2022.105853] [Medline: 35870318]
150. Bertini F, Allevi D, Lutero G, Montesi D, Calzà L. Automatic Speech Classifier for Mild Cognitive Impairment and Early Dementia. ACM Trans Comput Healthcare 2022;3:1-11 [doi: 10.1145/3469089]
151. Chi NA, Washington P, Kline A, Husic A, Hou C, He C, et al. Classifying Autism From Crowdsourced Semistructured Speech Recordings: Machine Learning Model Comparison Study. JMIR Pediatr Parent 2022 Apr 14;5(2):e35406 [FREE Full text] [doi: 10.2196/35406] [Medline: 35436234]
152. Ditthapron A, Lammert AC, Agu EO. Continuous TBI Monitoring From Spontaneous Speech Using Parametrized Sinc Filters and a Cascading GRU. IEEE J Biomed Health Inform 2022 Jul;26(7):3517-3528 [doi: 10.1109/JBHI.2022.3158840] [Medline: 35290191]

153. Cai T, Ni H, Yu M, Huang X, Wong K, Volpi J, et al. DeepStroke: An efficient stroke screening framework for emergency rooms with multimodal adversarial deep learning. Med Image Anal 2022 Aug;80:102522 [doi: 10.1016/j.media.2022.102522] [Medline: 35810587]
154. Farrús M, Codina-Filbà J, Reixach E, Andrés E, Sans M, Garcia N, et al. Speech-Based Support System to Supervise Chronic Obstructive Pulmonary Disease Patient Status. Applied Sciences 2021 Aug 29;11(17):7999 [doi: 10.3390/app11177999]
155. Ben AR, Ben AY. Speech Processing for Early Alzheimer Disease Diagnosis: Machine Learning Based Approach. New York, NY: IEEE; 2018 Presented at: 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA); October 28-November 1, 2018; Aqaba, Jordan p. 1-8 [doi: 10.1109/AICCSA.2018.8612831]
156. Verde L, De Pietro G, Sannino G. Artificial Intelligence Techniques for the Non-invasive Detection of COVID-19 Through the Analysis of Voice Signals. Arab J Sci Eng 2021 Oct 08:1-11 [FREE Full text] [doi: 10.1007/s13369-021-06041-4] [Medline: 34642613]
157. Verde L, De Pietro G, Ghoneim A, Alrashoud M, Al-Mutib KN, Sannino G. Exploring the Use of Artificial Intelligence Techniques to Detect the Presence of Coronavirus Covid-19 Through Speech and Voice Analysis. IEEE Access 2021;9:65750-65757 [FREE Full text] [doi: 10.1109/ACCESS.2021.3075571] [Medline: 35256922]
158. Rong P. A novel hierarchical framework for measuring the complexity and irregularity of multimodal speech signals and its application in the assessment of speech impairment in amyotrophic lateral sclerosis. J Speech Lang Hear Res. 2021;64:2996-3014 [doi: 10.1044/2021/JSLHR/20/00743]
159. 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]
160. Stegmann GM, Hahn S, Duncan CJ, Rutkove SB, Liss J, Shefner JM, et al. Estimation of forced vital capacity using speech acoustics in patients with ALS. Amyotroph Lateral Scler Frontotemporal Degener 2021;22(sup1):14-21 [doi: 10.1080/21678421.2020.1866013] [Medline: 34348537]
161. Rejaibi E, Komaty A, Meriaudeau F, Agrebi S, Othmani A. MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech. Biomedical Signal Processing and Control 2022 Jan;71:103107 [doi: 10.1016/j.bspc.2021.103107]
162. Rao Mv A, Yamini BK, Ketan J, Preetie Shetty A, Pal PK, Shivashankar N, et al. Automatic Classification of Healthy Subjects and Patients With Essential Vocal Tremor Using Probabilistic Source-Filter Model Based Noise Robust Pitch Estimation. J Voice 2023 May;37(3):314-321 [doi: 10.1016/j.jvoice.2021.01.009] [Medline: 33579623]
163. Suppa A, Asci F, Saggio G, Di Leo P, Zarezadeh Z, Ferrazzano G, et al. Voice Analysis with Machine Learning: One Step Closer to an Objective Diagnosis of Essential Tremor. Mov Disord 2021 Jun;36(6):1401-1410 [doi: 10.1002/mds.28508] [Medline: 33528037]
164. Rozenstoks K, Novotny M, Horakova D, Rusz J. Automated Assessment of Oral Diadochokinesis in Multiple Sclerosis Using a Neural Network Approach: Effect of Different Syllable Repetition Paradigms. IEEE Trans Neural Syst Rehabil Eng 2020 Jan;28(1):32-41 [doi: 10.1109/TNSRE.2019.2943064] [Medline: 31545738]
165. Al-Hameed S, Benaissa M, Christensen H, Mirheidari B, Blackburn D, Reuber M. A new diagnostic approach for the identification of patients with neurodegenerative cognitive complaints. PLoS One 2019;14(5):e0217388 [FREE Full text] [doi: 10.1371/journal.pone.0217388] [Medline: 31125389]
166. Roldan-Vasco S, Orozco-Duque A, Suarez-Escudero JC, Orozco-Arroyave JR. Machine learning based analysis of speech dimensions in functional oropharyngeal dysphagia. Comput Methods Programs Biomed 2021 Sep;208:106248 [doi: 10.1016/j.cmpb.2021.106248] [Medline: 34260973]
167. Daqrouq K, Al-Qawasmi AR, Balamesh A, Alghamdi AS, Al-Amoudi MA. The Use of Arabic Vowels to Model the Pathological Effect of Influenza Disease by Wavelets. Comput Math Methods Med 2019;2019 [FREE Full text] [doi: 10.1155/2019/4198462] [Medline: 31915460]
168. Lamba R, Gulati T, Alharbi H, Jain A. A hybrid system for Parkinson's disease diagnosis using machine learning techniques. Int J Speech Technol 2021 Apr 14;25(3):583-593 [doi: 10.1007/s10772-021-09837-9]
169. Benba A, Jilbab A, Hammouch A. Voice assessments for detecting patients with neurological diseases using PCA and NPCA. Int J Speech Technol 2017 Jul 8;20(3):673-683 [doi: 10.1007/s10772-017-9438-9]
170. Arora S, Lo C, Hu M, Tsanas A. Smartphone Speech Testing for Symptom Assessment in Rapid Eye Movement Sleep Behavior Disorder and Parkinson's Disease. IEEE Access 2021;9:44813-44824 [doi: 10.1109/ACCESS.2021.3057715]
171. Jeancolas L, Mangone G, Petrovska-Delacrétaz D, Benali H, Benkelfat BE, Arnulf I, et al. Voice characteristics from isolated rapid eye movement sleep behavior disorder to early Parkinson's disease. Parkinsonism Relat Disord 2022 Feb;95:86-91 [FREE Full text] [doi: 10.1016/j.parkreldis.2022.01.003] [Medline: 35063866]
172. König A, Satt A, Sorin A, Hoory R, Toledo-Ronen O, Derreumaux A, et al. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. Alzheimers Dement (Amst) 2015 Mar;1(1):112-124 [FREE Full text] [doi: 10.1016/j.dadm.2014.11.012] [Medline: 27239498]

173. Gosztolya G, Vincze V, Tóth L, Pákáski M, Kálmán J, Hoffmann I. Identifying Mild Cognitive Impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. Computer Speech & Language 2019 Jan;53:181-197 [doi: 10.1016/j.csl.2018.07.007]
174. Alkenani AH, Li Y, Xu Y, Zhang Q. Predicting Alzheimer's Disease from Spoken and Written Language Using Fusion-Based Stacked Generalization. J Biomed Inform 2021 Jun;118 [FREE Full text] [doi: 10.1016/j.jbi.2021.103803] [Medline: 33965639]
175. Sumali B, Mitsukura Y, Liang K, Yoshimura M, Kitazawa M, Takamiya A, et al. Speech Quality Feature Analysis for Classification of Depression and Dementia Patients. Sensors (Basel) 2020 Jun 26;20(12) [FREE Full text] [doi: 10.3390/s20123599] [Medline: 32604728]
176. Mirzaei S, El Yacoubi M, Garcia-Salicetti S, Boudy J, Kahindo C, Cristancho-Lacroix V, et al. Two-Stage Feature Selection of Voice Parameters for Early Alzheimer's Disease Prediction. IRBM 2018 Dec;39(6):430-435 [doi: 10.1016/j.irbm.2018.10.016]
177. Themistocleous C, Eckerström M, Kokkinakis D. Identification of Mild Cognitive Impairment From Speech in Swedish Using Deep Sequential Neural Networks. Front Neurol 2018;9:975 [FREE Full text] [doi: 10.3389/fneur.2018.00975] [Medline: 30498472]
178. Konig A, Satt A, Sorin A, Hoory R, Derreumaux A, David R, et al. Use of Speech Analyses within a Mobile Application for the Assessment of Cognitive Impairment in Elderly People. Curr Alzheimer Res 2018;15(2):120-129 [doi: 10.2174/1567205014666170829111942] [Medline: 28847279]
179. Ying Y, Yang T, Zhou H. Multimodal fusion for alzheimer's disease recognition. Appl Intell 2022 Dec 01;53(12):16029-16040 [doi: 10.1007/s10489-022-04255-z]
180. Song J, Lee JH, Choi J, Suh MK, Chung MJ, Kim YH, et al. Detection and differentiation of ataxic and hypokinetic dysarthria in cerebellar ataxia and parkinsonian disorders via wave splitting and integrating neural networks. PLoS One 2022;17(6):e0268337 [FREE Full text] [doi: 10.1371/journal.pone.0268337] [Medline: 35658000]
181. Chen F, Yang C, Khishe M. Diagnose Parkinson's disease and cleft lip and palate using deep convolutional neural networks evolved by IP-based chimp optimization algorithm. Biomedical Signal Processing and Control 2022 Aug;77:103688 [doi: 10.1016/j.bspc.2022.103688]
182. Alam MZ, Simonetti A, Brillantino R, Tayler N, Grainge C, Siribaddana P, et al. Predicting Pulmonary Function From the Analysis of Voice: A Machine Learning Approach. Front Digit Health 2022;4:750226 [FREE Full text] [doi: 10.3389/fdgth.2022.750226] [Medline: 35211691]
183. Nilashi M, Ibrahim O, Ahmadi H, Shahmoradi L, Farahmand M. A hybrid intelligent system for the prediction of Parkinson's Disease progression using machine learning techniques. Biocybernetics and Biomedical Engineering 2018;38(1):1-15 [doi: 10.1016/j.bbe.2017.09.002]
184. Tsanas A, Little MA, Fox C, Ramig LO. Objective Automatic Assessment of Rehabilitative Speech Treatment in Parkinson's Disease. IEEE Trans Neural Syst Rehabil Eng 2014 Jan;22(1):181-190 [doi: 10.1109/TNSRE.2013.2293575] [Medline: 26271131]
185. Jain A, Abedinpour K, Polat O, Çalışkan MM, Asaei A, Pfister FMJ, et al. Voice Analysis to Differentiate the Dopaminergic Response in People With Parkinson's Disease. Front Hum Neurosci 2021;15:667997 [FREE Full text] [doi: 10.3389/fnhum.2021.667997] [Medline: 34135742]
186. Pană MA, Busnatu SS, Serbanoiu LI, Vasilescu E, Popescu N, Andrei C. Reducing the heart failure burden in romania by predicting congestive heart failure using artificial intelligence: proof of concept. Appl Sci Switz 2021;11(24):11728 [doi: 10.3390/app112411728]
187. Gao X, Ma K, Yang H, Wang K, Fu B, Zhu Y, et al. A rapid, non-invasive method for fatigue detection based on voice information. Front Cell Dev Biol 2022;10:994001 [FREE Full text] [doi: 10.3389/fcell.2022.994001] [Medline: 36176279]
188. Bárcenas R, Fuentes-García R, Naranjo L. Mixed kernel SVR addressing Parkinson's progression from voice features. PLoS One 2022;17(10):e0275721 [FREE Full text] [doi: 10.1371/journal.pone.0275721] [Medline: 36206238]
189. Lab L. Coswara-Data. GitHub. 2022. URL: https://github.com/iiscleap/Coswara-Data [accessed 2022-10-31]
190. Setia MS. Methodology Series Module 3: Cross-sectional Studies. Indian J Dermatol 2016;61(3):261-264 [FREE Full text] [doi: 10.4103/0019-5154.182410] [Medline: 27293245]
191. Caruana E, Roman M, Hernández-Sánchez J, Solli P. Longitudinal studies. J Thorac Dis 2015 Nov;7(11):E537-E540 [FREE Full text] [doi: 10.3978/j.issn.2072-1439.2015.10.63] [Medline: 26716051]
192. White RT, Arzi HJ. Longitudinal Studies: Designs, Validity, Practicality, and Value. Res Sci Educ 2005 Mar;35(1):137-149 [doi: 10.1007/s11165-004-3437-y]
193. Scherer RW, Saldanha IJ. How should systematic reviewers handle conference abstracts? A view from the trenches. Syst Rev 2019 Nov 07;8(1):264 [FREE Full text] [doi: 10.1186/s13643-019-1188-0] [Medline: 31699124]

Abbreviations

AD: Alzheimer disease
ADBC: Alzheimer Dementia Bank blog corpus

ALS: amyotrophic lateral sclerosis
ANN: artificial neural network
BLA: base line acoustic
CI: cognitive impairment
CFS: collected for study
CHRSD: Corona Hack Respiratory Sound data set
DT: decision tree
GB: gradient boosting
GMM: Gaussian mixture model
HC: healthy control
KNN: K-nearest neighbor
LR: logic regression
MCI: mild cognitive impairment
MeML: mixed effect machine learning
MeSH: Medical Subject Headings
MFCC: Mel-frequency cepstral coefficients
ML: machine learning
NB: naïve Bayes
NCC: neurodegenerative cognitive complaint
NCVS: National Center for Voice and Speech
ND: neurological disease
NG: not given
PA: passive active
PARCZ: Czech Parkinsonian Speech Database
PD: Parkinson disease
PICO: population, intervention, comparison, and outcome
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RF: random forest
RP: recurrence plot
SLR: systematic literature review
SQ: subquestion
SVM: support vector machine
SVR: support vector regression
TBIBank: Traumatic Brain injury bank
TQWT: tunable Q-factor wavelet transform
UCI: University of California, Irvine

Edited by A Mavragani; submitted 30.01.23; peer-reviewed by F Chu, V Martin; comments to author 09.03.23; revised version received 26.04.23; accepted 23.05.23; published 19.07.23

Please cite as:
Idrisoglu A, Dallora AL, Anderberg P, Berglund JS
Applied Machine Learning Techniques to Diagnose Voice-Affecting Conditions and Disorders: Systematic Literature Review
J Med Internet Res 2023;25:e46105
URL: https://www.jmir.org/2023/1/e46105
doi: 10.2196/46105
PMID:

Study II

COPDVD: Automated Classification of Chronic Obstructive Pulmonary Disease on a New Developed and Evaluated Voice Dataset

Published as:

Idrisoglu A, Dallora AL, Cheddad A, Anderberg P, Jakobsson A, Sanmartin Berglund J. COPDVD: Automated Classification of Chronic Obstructive Pulmonary Disease on a New Developed and Evaluated Voice Dataset. Artificial Intelligence in Medicine. 2024;156. doi:10.1016/j.artmed.2024.102953.

Artificial Intelligence In Medicine 156 (2024) 102953

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

COPDVD: Automated classification of chronic obstructive pulmonary disease on a new collected and evaluated voice dataset

Alper Idrisoglu [a,*], Ana Luiza Dallora [a], Abbas Cheddad [a], Peter Anderberg [a], Andreas Jakobsson [b], Johan Sanmartin Berglund [a]

[a] Blekinge Institute of Technology, Valhallavägen 1, 371 41 Karlskrona, Sweden
[b] Lund University, Box 117, SE-221 00 Lund, Sweden

ARTICLE INFO

Keywords:
Acoustic features
Mel Frequency Cepstrum Coefficients
Automated classification
Chronic obstructive pulmonary disease
Machine Learning

ABSTRACT

Background: Chronic obstructive pulmonary disease (COPD) is a severe condition affecting millions worldwide, leading to numerous annual deaths. The absence of significant symptoms in its early stages promotes high underdiagnosis rates for the affected people. Besides pulmonary function failure, another harmful problem of COPD is the systemic effects, e.g., heart failure or voice distortion. However, the systemic effects of COPD might provide valuable information for early detection. In other words, symptoms caused by systemic effects could be helpful to detect the condition in its early stages.
Objective: The proposed study aims to explore whether the voice features extracted from the vowel "a" utterance carry any information that can be predictive of COPD by employing Machine Learning (ML) on a newly collected voice dataset.
Methods: Forty-eight participants were recruited from the pool of research clinic visitors at Blekinge Institute of Technology (BTH) in Sweden between January 2022 and May 2023. A dataset consisting of 1246 recordings from 48 participants was gathered. The collection of voice recordings containing the vowel "a" utterance commenced following an information and consent meeting with each participant using the *VoiceDiagnostic* application. The collected voice data was subjected to silence segment removal, feature extraction of baseline acoustic features, and Mel Frequency Cepstrum Coefficients (MFCC). Sociodemographic data was also collected from the participants. Three ML models were investigated for the binary classification of COPD and healthy controls: Random Forest (RF), Support Vector Machine (SVM), and CatBoost (CB). A nested k-fold cross-validation approach was employed. Additionally, the hyperparameters were optimized using grid-search on each ML model. For best performance assessment, accuracy, F1-score, precision, and recall metrics were computed. Afterward, we further examined the best classifier by utilizing the Area Under the Curve (AUC), Average Precision (AP), and SHapley Additive exPlanations (SHAP) feature-importance measures.
Results: The classifiers RF, SVM, and CB achieved a maximum accuracy of 77 %, 69 %, and 78 % on the test set and 93 %, 78 % and 97 % on the validation set, respectively. The CB classifier outperformed RF and SVM. After further investigation of the best-performing classifier, CB demonstrated the highest performance, producing an AUC of 82 % and AP of 76 %. In addition to age and gender, the mean values of baseline acoustic and MFCC features demonstrate high importance and deterministic characteristics for classification performance in both test and validation sets, though in varied order.
Conclusion: This study concludes that the utterance of vowel "a" recordings contain information that can be captured by the CatBoost classifier with high accuracy for the classification of COPD. Additionally, baseline acoustic and MFCC features, in conjunction with age and gender information, can be employed for classification purposes and benefit healthcare for decision support in COPD diagnosis.
Clinical trial registration number: NCT05897944.

* Corresponding author at: Blekinge Institute of Technology, Valhallavägen 1, Karlskrona 371 41, Sweden.
E-mail address: alper.idrisoglu@bth.se (A. Idrisoglu).

https://doi.org/10.1016/j.artmed.2024.102953
Received 24 January 2024; Received in revised form 26 July 2024; Accepted 13 August 2024
Available online 15 August 2024

1. Introduction

Chronic obstructive pulmonary disease (COPD) is a common lung condition that affects breathing. It is one of the leading causes of illness and death globally, mainly linked to smoking, although many non-smokers also develop COPD [1]. Boers et al. combined the statistics reported in different studies and suggested that in 2019, the prevalence of COPD was around 300 to 400 million people worldwide [2,3]. In the same year, the WHO reported COPD as the third leading cause of death worldwide [4]. COPD reduces lung function, leading to decreased exhalation capacity. Diagnostic tests measure lung function, including the amount of air breathed out in one second, forced expiratory volume in one second (FEV1) and the maximum air exhaled forcefully, forced vital capacity (FVC). Abnormal ratios of FEV1/FVC (below 0.7 L) [5], also known as vital capacity (VC), obtained through spirometry, is considered an essential diagnostic measure to indicate COPD [5]. Reduced ratio of FEV1/FVC can cause great discomfort and poor quality of life for COPD patients [6,7]. Beside its pulmonary effects, the systemic impacts of COPD, such as cardiovascular effects, nervous system effects, and weight loss, are also well-documented [8]. All these data depict the burden of COPD globally and underscore the importance of taking action against it. Despite the promise of voice-based diagnostics, the application of ML techniques for COPD classification remains relatively understudied [9].

In recent years, Machine Learning (ML) techniques have garnered attention in healthcare sectors [10], offering innovative solutions for disease diagnosis and prognosis [11], including COPD. ML, the field that systematically studies algorithms that improve with experience, presents opportunities for enhanced clinical decision-making [12–14]. Notably, ML-based analyses of medical imaging data, such as X-rays and CT scans, have shown promising results in cancer detection [15–21]. Similarly, there has been a growing interest in utilizing voice-based biomarkers for disease diagnosis and monitoring, particularly in respiratory conditions such as COPD. By analyzing vocal features extracted from speech recordings, researchers have explored the potential of ML algorithms to develop automated systems for disease classification and prognosis [9]. These advancements offer a non-invasive and cost-effective approach to healthcare delivery, with the potential to revolutionize diagnostic practices. ML is adept at automatically recognizing these patterns, offering a more efficient and effective decision-support process [22].

This study aims to apply ML techniques to explore whether voice features extracted from recordings of the vowel 'a,' utterance collected via mobile phones, can provide insights into classifying COPD patients versus healthy controls (HC). The investigated machine learning classifiers are: Random Forest (RF), Support Vector Machine (SVM), and CatBoost (CB). Additionally, the study aims to perform the experiment on a newly collected COPD voice dataset based on the performance evaluation results. The study analyzes a dataset consisting of demographics, health information, baseline acoustic features, and Mel Frequency Cepstral Coefficients (MFCC). The latter features are extracted from vowel "a" utterance collected from both COPD and HC participants.

Due to the absence of remarkable symptoms in its early stages, the condition has high underdiagnosis and misdiagnosis rates [23,24]. Detecting COPD in its early stages will not only benefit the treatment course but can also help reduce morbidity and mortality rates, slow down its progression, and increase lifetime expectancy by preventing or delaying its severe consequences [25,26]. The potential implication of voice assessment techniques using ML in the healthcare domain represents a substantial advancement in the widespread automated classification of COPD. The contributions of this work are listed below:

- The evaluation of using ML and voice, specifically with features extracted from vowel "a" utterance, for classifying COPD and HC cases.
- The performance assessment of various ML classifiers, including RF, SVM, and CB, in COPD and HC voice classification.
- The collection of a new Swedish voice dataset for the binary classification of COPD and HC cases.

The remainder of this paper is organized as follows: Section 2 provides an overview of related work in the field, highlighting gaps and motivating our study. Section 3 details the methodology we employed, including data collection, feature extraction, and ML algorithms utilized. Section 4 presents experimental results. Section 5 consists of discussions, followed by future research and limitations. Section 6 provides the conclusions.

2. Related work

The classification manual for voice disorders classifies COPD as a nonlaryngeal aerodigestive voice-affecting disorder [27]. In a related study, voice characteristics were examined in individuals with COPD compared to healthy counterparts [28]. The study focused on baseline acoustic features (BLA) obtained from recordings of vowel "a" utterance. BLA analysis revealed lower fundamental frequency, reduced frequency range, increased pitch and amplitude perturbation measures, and heightened noise measures in the COPD group. Additionally, perceptual assessments indicated increased roughness, breathiness, and strain in the COPD group compared to healthy individuals. Another study delved into the voice alterations in COPD patients [29]. The investigation involved interviews conducted by phoneticians to evaluate the patients. Results unveiled significant positive correlations between the smoking index and parameters like Jitter, Shimmer, and the severity of dysphonia. Additionally, inverse correlations were observed between Jitter, Shimmer, dysphonia severity, and variables such as dry powder inhalers usage, as well as FVC, FEV1, and Maximal Mid-Expiratory Flow of predicted values. Notably, both studies employed statistical methodologies to discern distinctions between COPD and normal voices, without integrating machine learning techniques into the analysis, and, they were only concerned with investigating associations between features and COPD.

There are very few studies using voice analysis techniques in conjunction with ML for COPD assessment [30]. Nallanthighal et al. conducted a study investigating COPD exacerbation based on speech recordings and combining three different approaches: acoustic features identification using a statistical approach, low-level descriptive features with classification with SVM, and speech breathing models based on deep learning architectures to estimate the patients' breathing rate [31]. One study investigates the impact of physical effort and medication on speech patterns in COPD patients, particularly focusing on prosody [32]. By analyzing COPD and control groups, it examines how these factors affect speech and influence the performance of automatic COPD detection systems under various recording conditions employing RF classifier. Findings suggest that speech, notably prosodic features, is influenced by physical exertion and medication, with implications on system design. This research contributes to understanding COPD-related speech changes and aids in the development of automated detection systems for healthcare monitoring. Another study addresses the challenge of continuous monitoring of respiratory diseases like COPD and asthma using mobile devices [33]. It reports RF as the best overall performing algorithm for passive assessment of pulmonary condition: one for detecting obstructive pulmonary disease and the other for estimating pulmonary function based on the FEV1/FVC ratio. All three studies report that they employed speech recordings and k-fold cross-validation in their experiment. In the given context of utilizing voice and machine learning for COPD assessment, the publication frequency is not particularly high, and the studies cited in this section primarily consist of conference proceedings. It is important to note that the provided results in previous studies, except for the one in [32], are preliminary.

Two studies by the same authors utilized recordings of the vowel "a"

utterance and employed RF for distinguishing between COPD and HC voices [34,35]. However, these studies split the dataset without considering individual subjects, resulting in their test set outcomes mirroring the validation set results in this study. Another notable difference lies in the approach to dataset expansion. While they utilized segments from the same recording, our study conducted feature extraction separately for each recording.

In this study, a performance comparison was conducted among three distinct classifiers: RF, SVM, and CB, which deviate from previous approaches. To mitigate the risk of overfitting, nested cross-validation (nCV) was applied, a consideration which is not present in prior studies. Furthermore, the utilization of MFCC features, commonly applied in various applications, particularly in acoustic analysis [36], distinguishes this study from its predecessors.

3. Methods

This section outlines the workflow and components of the study. It focuses on experimenting with raw voice data, specifically, voice recordings of the vowel "a" utterance for the purpose to construct a dataset comprising acoustic voice features, MFCC and sociodemographic data. This dataset is then used for the classification experiment of COPD and HC voices. To achieve this, three machine learning algorithms, RF, SVM, and CB with nCV, were employed, and performance measures of all classifiers are assessed. Fig. 1 shows an overview of the methodology followed in this study.

3.1. Participant recruitment

The recruitment took place at the research clinic of the Blekinge Institute of Technology (BTH), in Sweden. Participants were recruited from the pool of visitors that started in January 2022 and ended in May 2023. A total number of 72 participants (34 COPD and 38 HC) were recruited. Of these, 4 COPD participants did not provide any recordings and were considered dropouts. In total, 1351 recordings were collected. The 30 remaining COPD participants (16 female and 14 male) provided 469 recordings. The 38 HC participants (20 female and 18 male) provided 882 recordings. A total of 33 recordings from the COPD group and 72 from the HC group were deemed improper for feature extraction because the recordings contained background noise (e.g., the voice of surrounding individuals, pets, or machinery) or did not contain a vocal segment. After data scrubbing, the final dataset consisted of 1246 recordings from 68 recruited participants. The average age for the COPD and HC persons were 72.04 (std: 6.86) and 72.13 (std: 6.84), respectively. The average VC for the COPD group was 0.61 L (std: 0.12 L). The inclusion and exclusion criteria employed in participant recruitment are shown below.

Inclusion criteria
COPD group:

- Being 18 years old or older.
- Having a COPD diagnosis given by a physician.
- Having a smartphone.

HC group:

- Being 18 years old or older.
- Not having a voice-affecting disorder diagnosis, i.e., no disorder listed in the categories 'nonlaryngeal aerodigestive disorders affecting voice', 'neurological disorders affecting voice', and 'systemic conditions affecting voice disorder' in the Classification Manual for Voice Disorders [37].
- Having a smartphone.

Exclusion criteria
COPD group:

- Being younger than 18 years old.
- Having a voice-affecting disorder other than COPD.
- Declaring no access or proficiency to use a smartphone.

HC group:

- Being younger than 18 years old.
- Having any voice-affecting disorder.
- Declaring no access or proficiency to use a smartphone.

3.2. Data acquisition

Data collection was done through a mobile application named *VoiceDiagnostic*[1] which is compatible with both Android and iPhones. Participants were matched in terms of age and gender. Age matching was defined within a 5-year range. Recordings from participants who did not match in terms of gender and age within both groups were excluded from the experiments. Research nurses initially briefed the potential participants about the study. Subsequently, they posed two key questions: "*Is this of interest to you*?" and "*Would you like to learn more about the study*?". Those expressing interest proceeded to meet with the first author, who briefed the participant about the study, data collection procedures, and confidentiality of sensitive data. Participants willing to participate signed a written informed consent form. After the registration of age, gender, social security number, phone number, and if an additional diagnosis exists, each participant was assigned a unique ID number to log in to the *VoiceDiagnostic* application. Initial recordings were carried out in a silent room at the BTH research clinic using participants' own mobile phones or an iPad belonging to the institution to ensure the participants comprehended the written instructions for using the application. Participants were instructed to take a deep breath and sustain the utterance of the vowel "a" for as long as possible in a quiet setting, once a week for six months.

The *VoiceDiagnostic* application featured an internal reminder function to prompt participants for their next recording. Before each recording session, participants were asked a set of questions to identify if they had any initial throat issues. Following that, clear instructions were provided on how the recording should take place (see Fig. 2).

3.3. Data safety and ethics

This study was carried out in accordance with the Declaration of Helsinki and was approved by the Swedish ethics review authority in Umeå (DNR: 2020-01045). Written informed consent was collected from all subjects. All data was anonymized and grouped by age and gender.

Recordings were stored in accordance with unique ID groupings, differentiating between COPD and Healthy Control (HC) IDs. Subsequently, the data was stored on a secured server, accessible to authorized administrators for online download. A key list consisting of the personal and contact information of each participant was stored offline locally in a secured location at BTH with limited access to ensure the protection of the integrity of each participant. The signed informed consent forms were also secured in fireproof and locked cabinets.

3.4. Feature extraction

Naturally, the raw voice recordings do not arrive in an optimal form. Hence, audio files were checked by listening to each file for proper recording, and an automatic segmentation procedure was applied to the raw recordings to remove silent segments located at the beginning and ending of the recording. This was done by identifying the vocal segment using a basic moving average filter.

[1] Available from: https://www.voicediagnostic.com, accessed on: 2023-11-15.

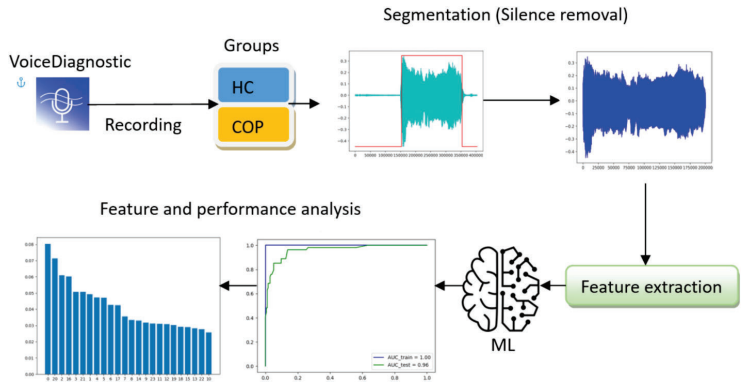


Fig. 1. Overview of the methodology employed in this study.

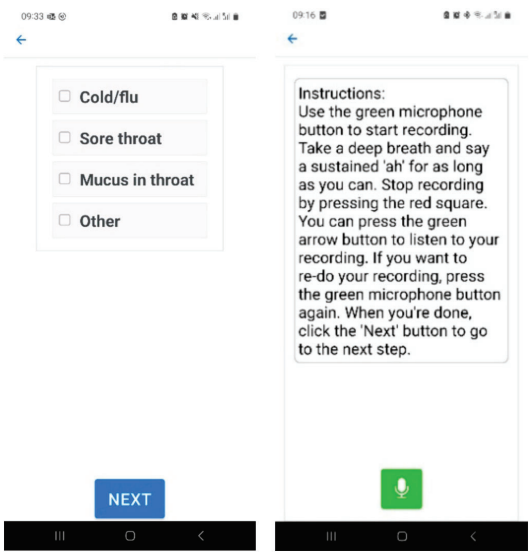


Fig. 2. The initial questionnaire and instructions from the screen of the mobile application for the utterance of the vowel "a" recording.

The threshold was established as the average amplitude, incremented by one, compared against the average power amplitude of each normalized window incremented by one. This comparison was employed to delineate the start and end points of each vocal segment. The longest vocal segment, identified through this method, was then extracted from the original recording for subsequent processing. Please refer to Eqs. (1), (2), and Fig. 3 for further details.

$$Aavg = 1 + \frac{1}{n}\sum_{k=0}^{n-1} x_k \tag{1}$$

where A_{avg} is the incremented arithmetic mean by one, x is the signal value, n is the number of values in the signal, and k is the index.

$$Wpavg = \left(\frac{1}{m}\sum_{k=0}^{m-1}\frac{x_k}{x_{max}} + 1\right)^2 \tag{2}$$

where Wpavg is the window power arithmetic mean of the normalized values +1, x is the data value, m is the number of values in a given window, and k is the index.

The process of reading audio files gives two distinct outputs: the raw data points and the sampling rate associated with each recording. These outputs were subsequently utilized to assess the sampling rate deviation, thereby confirming that all recordings were consistently captured at a frequency of 44,1 kHz.

The BLA and MFCC features were extracted from the vocal segments using the Python code libraries Praat (Parselmouth)[2] and Librosa,[3] which are software packages tailored for speech analysis. Furthermore, an Excel file was generated, encompassing demographic information, labels, BLA, and MFCC features, effectively creating a comprehensive dataset list. MFCC feature extraction method is common in various audio processing applications, such as speech recognition, speaker identification, and audio analysis [37]. The process begins by transforming the signal from the time domain to the frequency domain using the Fourier transform. Then, the power spectrum is mapped onto the Mel scale using a set of triangular bandpass filters. Finally, the Discrete Cosine Transform (DCT) is applied to the logarithm of the Mel scale energies to generate the features [38]. Detailed information about the features in this dataset can be found in Table 1.

3.5. Classification

The experiment was performed on a Dell Precision 7920 MT desktop with 8 GB RAM memory, utilizing Python's Scikit-Learn,[4] Seaborn,[5] and CatBoost[6] libraries for implementing classifiers and evaluation metrics. The extracted features were assigned to the X feature array corresponding to their respective Y labels. To address data imbalance, the X and Y arrays were partitioned into two distinct sets: Balanced X and Y and Unbalanced X and Y. Furthermore, the Balanced X and Y sets were further divided into two subsets, consisting of a training plus validation set and an independent test set. This partitioning was executed on the participant level to ensure that the models were evaluated solely on unseen data. The data within the training and validation set were

[2] Available from: https://parselmouth.readthedocs.io/en/stable/, accessed on: 2024-05-14.
[3] Available from: https://librosa.org/doc/latest/index.html, accessed on: 2024-05-14.
[4] Available from: https://scikit-learn.org/stable/index.html, accessed on: 2024-07-17.
[5] Available from: https://seaborn.pydata.org/index.html, accessed on: 2024-07-17.
[6] Available from: https://catboost.ai/en/docs/, accessed on: 2024-07-17.

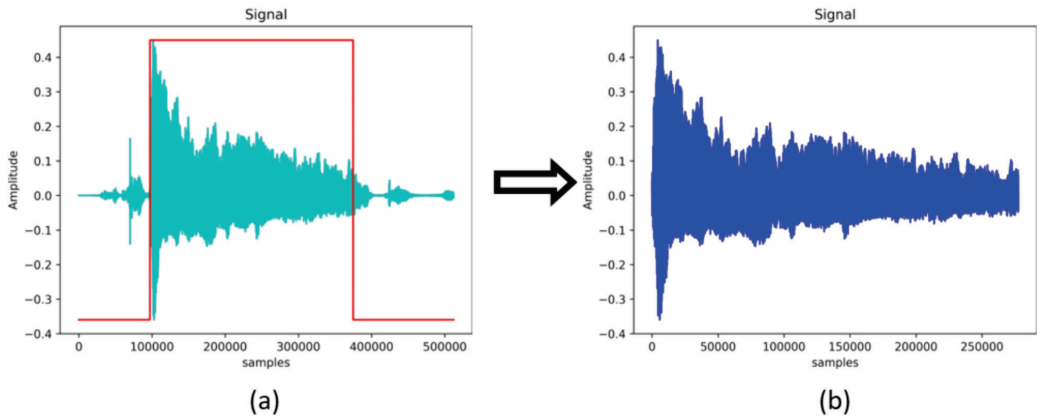


Fig. 3. Silence removal process illustration. Image (a) shows the visual representation of the raw recording and the marked voice region bounded by the red line, and (b) illustrates the voice region after removing the silence segment. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

organized into two groups; however, this arrangement was performed at the data level. Fig. 4 illustrates the data partitioning. To mitigate the risk of unbalanced data effects on ML models, the dataset was relatively balanced regarding gender, age, and number of participants in all groups. However, a marginal deviation in terms of the number of recordings remains.

The training set parameters presented in Table 1 were utilized as input features to train three distinct classifiers: RF, SVM, and CB. Subsequently, each model underwent a series of training, validation, and testing iterations for various KxK-fold combinations, ranging from 2×2 to 10×10, in order to determine the optimal nested nCV configuration, where each combination generates average performance metrics for a specific setup. To fine-tune the hyperparameters for optimal performance, a grid search technique from the Python Sklearn library was employed.

The selection of these ML models was grounded in a previous study [9], which utilized SVM, Artificial Neural Networks (ANN), and RF as the primary models for classifying various disorders using voice and machine learning techniques. It is worth noting that ANNs typically exhibit improved performance with larger datasets, which is not the case of the present study [39,40]. Consequently, SVM and RF were chosen for experimentation alongside CB, which are recommended for their performance in handling heterogeneous medical data [41,42]. The utilization of the nCV method was imperative to mitigate the risk of overfitting for each classifier [43,44].

The grid search method was performed using a parameters grid object for each ML model. The RF model was optimized for the following parameters: number of estimators, max depth, and minimum samples split. The SVM model utilized several C (regularization), kernel, and degree parameters to search for the best parameter combination. The grid search parameters employed for the CB were the number of iterations, learning rate, depth, and l2-leaf regularization. Table 2 depicts the values used in hyperparameter tuning.

3.6. Performance evaluation

The Accuracy, F1-score, Precision, and Recall metrics were used to observe the overall performance of each ML model. The best-performing ML model was further investigated through a closer look at the Area Under the Curve (AUC) and Precision-Recall (PR) curve to evaluate the performance of the classifier. Additionally, the SHAP summary plots were utilized to depict the most important features for the classifiers. SHAP stands for SHapley Additive exPlanations, and it is a method used to explain the output of machine learning models. SHAP values provide an elegant way to understand the contribution of each feature or input variable in a predictive model to the final prediction. They are based on cooperative game theory concepts, specifically Shapley values, which allocate the value of a group to its members in a fair and consistent manner. In simpler terms, when one notes the contribution of each feature in a predictive model, the SHAP values provide a fair and consistent way to measure this contribution. It is like attributing credit to each feature for the model's prediction, considering its collective impact in a manner that is unbiased and equitable. The SHAP summary plots provide a comprehensive visualization of the relative importance of each feature group and elucidate how the actual feature values influence the classifier's prediction [45,46].

Eq. (3) shows the calculation of accuracy, which is a ratio of correct estimates to all estimates. Precision and Recall in Eqs. (4) and (5) refer to the performance of the models for positive prediction and how often the model can predict all positive instances, respectively. On the other hand, the F1_Score in Eq. (6) is an accuracy metric that balances precision and recall metrics on positive instances. The F1_Score becomes more important when there is an imbalance between positive and negative classes [47].

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

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1\_Score = \frac{2(Precision*Recall)}{Precision + Recall} \tag{6}$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative estimated samples, respectively.

4. Results

4.1. Experimental results

Fig. 5 presents the accuracy of the ML models evaluated on different combinations of nCV as well as the average accuracy for each ML model.

Table 1
Features list used for the proposed binary classification of COPD.

Feature group	Feature	Nr of parameters	Description
Demographic	Age	1	Age of participant
	Gender	1	Biological gender
Health	Cold (Cold/Flu)	1	Binary answer to the question "Cold/Flu?"
	Pain (Sore throat)	1	Binary answer to the question: "Sore throat?"
	Slimy (Mucus in throat)	1	Binary answer to the question: "Mucus in throat?"
	Other	1	If the participant is not Cold, has no pain, or is slimy.
BLA	Duration	1	Duration of the voice part
	Mean_F0	1	Mean fundamental frequency
	Std_F0	1	The standard deviation of fundamental frequency
	HNR	1	Harmonic to Noice Ratio
	Local_Jitter	1	The average absolute frequency difference between two consecutive periods, divided by the average period
	Absolute_Jitter	1	Measure of the absolute difference between a clock edge as specified and its observed position
	Rap_Jitter	1	Relative average perturbation
	PPQ5_Jitter	1	Five-point period perturbation quotient
	DDP_Jitter	1	Divided difference between consecutive periods
	Local_Shimmer	1	The average absolute amplitude difference between two consecutive periods, divided by the average period
	LocaldB_Shimmer	1	Decibel representation of Local Shimmer
	APQ3_Shimmer	1	Three-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average
	APQ5_Shimmer	1	Five-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average
	APQ11_Shimmer	1	Eleven-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average
	DDA_Shimmer	1	The average absolute difference between consecutive differences between the amplitudes of consecutive periods
	F1_mean-F4_mean	4	The mean of the four lowest resonant frequencies of the vocal tract.
	F1_median-F4_median	4	The median of the four lowest resonant frequencies of the vocal tract.
MFCC	MFCC1_mean-MFCC13_mean	13	The mean of the 13 Mel Frequency Cepstral Coefficients
	MFCC1_std-MFCC13_std	13	The standard deviation of the 13 Mel Frequency Cepstral Coefficients
	DMFCC1_mean-DMFCC13_mean	13	The mean first derivative of the 13 Mel Frequency Cepstral Coefficients
	DMFCC1_std-DMFCC13_std	13	The standard deviation of the first derivative of the 13 Mel Frequency Cepstral Coefficients
	DDMFCC1_mean-DDMFCC13_mean	13	The mean second derivative of the 13 Mel Frequency Cepstral Coefficients
	DDMFCC1_std-DDMFCC13_std	13	The standard deviation of the second derivative of the 13 Mel Frequency Cepstral Coefficients
	Total parameters	107	

The CB classifier was found to have the highest score with accuracy levels up to 97.0 % on two nCV combination points, namely, on 4 × 7 and 9 × 5, with an average accuracy of 95.0 %. Similarly, the RF classifier, with 93.0 % highest accuracy on the 9 × 5 nCV combination with 90.0 % average accuracy, showed the second-best performance on the validation dataset, whereas the SVM classifier followed with 78.0 % highest and 74.6 % average accuracy scores.

Similarly, Fig. 6 depicts the test accuracy of the ML models evaluated on different combinations of nCV, as well as the average accuracy for each ML model, indicating no change in the performance order of the models. The CB classifier yields the highest accuracy of 78.0 % on the 4 × 9 and 6 × 8 combinations. Additionally, the CB generates the highest average accuracy of 72.8 % overall for 81 nCV combinations. The RF classifier achieved the closest accuracy score to the CB, with the highest score of 77.0 % on the 6 × 5 and 7 × 3 nCV combinations with 70.0 % average accuracy. The SVM achieved the highest and average accuracy of 69.0 % and 65.0 %, respectively. A higher variation in the test accuracy curve is apparent when comparing the accuracy graphs of the validation and test sets, as depicted in Figs. 5 and 6, which indicate the effects of dependent and independent data referring to the validation set and test set, respectively.

To catch the effects of small deviations caused by different number of recordings per classification group, metrics such as F1_Score, Precision, and Recall were also calculated. Since the data were nearly balanced, all other measures followed the accuracy pattern. Table 3 shows the best scores for each ML model and metric for the training, validation, and test set. Graphical results belonging to other metrics other than accuracy can be seen in Appendix 1.

The classification results for CB, RF, and SVM in COPD assessment, including all parameters of the confusion matrix of validation and test sets, are summarized in Table 4. For the validation set, CB correctly classified 101 positive cases and 73 negative cases, with 7 positive cases misclassified as negative and 2 negative cases misclassified as positive. RF achieved 98 true positive classifications and 69 true negative classifications, with 10 positive cases misclassified as negative and 6 negative cases misclassified as positive. SVM achieved 85 true positive classifications and 55 true negative classifications, with 23 positive cases misclassified as negative and 20 negative cases misclassified as positive. For the test set, CB correctly classified 50 positive cases and 61 negative cases, with 7 positive cases misclassified as negative and 25 negative cases misclassified as positive. RF achieved 48 true positive classifications and 62 true negative classifications, with 6 positive cases

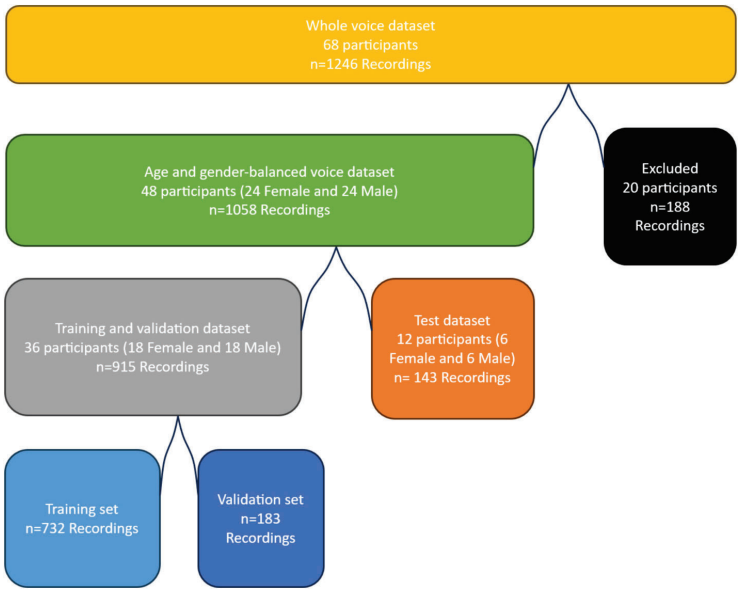


Fig. 4. Data allocation for employing machine learning algorithms in the COPD binary classification tasks. Different colors indicate subsets of the data.

Table 2
Utilized hyperparameters for each ML model in the grid search, where the best-performing values are highlighted in bold.

ML model	Parameter	Values
RF	N_estimators	(50,**100**,200)
	Max_depth	(None, **10**, 20)
	Min_samples_split	(**2**,5,10)
SVM	C (regularization)	(**0.1**,1,10)
	Kernel	(**linear**, rbf, poly)
	Degree	(**2**,3,4)
CB	N_iteration	(100,200,**300**)
	Learning_rate	(0.01,0.1,**0.2**)
	Depth	(**4**,6,8)
	L2_leaf_regularization	(1,**3**,5)

misclassified as negative and 27 negative cases misclassified as positive. SVM achieved 45 true positive classifications and 53 true negative classifications, with 15 positive cases misclassified as negative and 30 negative cases misclassified as positive. These results provide detailed insights into the performance of CB, RF, and SVM in distinguishing between positive and negative instances in the validation and test set.

Apart from methodological variances, Table 5 showcases the highest test accuracy reported in studies within the same context of employing voice analysis and machine learning techniques for assessing COPD and HC individuals. The findings reveal that the proposed CB algorithm achieves the best accuracy of 78 % compared to other studies.

The CB classifier achieved the highest measures in all metrics, with the RF classifier being the closest competitor, whereas the SVM was found to have the worst performance. Furthermore, the performance of the CB classifier was further investigated using the AUC and the PR curve, which gave 99 % and 98 % for the two metrics, respectively (see Fig. 7).

Similar to the validation set, the CB's AUC and AP measurements on the test set also suggest a good performance, reaching 82 % and 76 %, respectively, see Fig. 8.

The SHAP summary plot in Fig. 9 visualizes all the 107 features and their relation to the prediction performance of the validation set arranged in order of importance for the CB classifier from top to bottom. Feature importance order indicates that age is the most important deterministic indicator among all tested features. However, higher ages and high values of *MFCC9_mean* pose a negative impact on the prediction. In contrast, high values of *Absolute_jitter* and *std_F0* are found to have a negative impact on its low values. From a wider perspective, most of the BLA features and MFCC values appear to be important for the classification of COPD and HC voices. Of the examined features, *Age*, *MFCC9_mean*, *Absolute_jitter*, *F1_mean*, *Duration*, *mean_F0*, *std_F0*, *MFCC8_std*, *D2MFCC7_mean*, and *MFCC4_mean,* were all ranked as the top ten most important features.

Nevertheless, the feature importance graph for the test set in Fig. 10 indicates some deviations. Even though the feature-to-feature cooperation shows that the feature values have a similar impact on the prediction performance, the feature importance order deviates from the validation set features in Fig. 9 (e.g., the "duration" feature was ranked as the 5th most important feature for the validation set, and its importance decreased to the 12th most important ranking in the test set). Eight of ten features in the validation set, namely, *Age, MFCC9_mean, Absolute_jitter, F1_mean, mean_F0, std_F0, MFCC8_std, and D2MFCC7_mean,* remained in the top ten most important features' group, and *MFCC8_std* and *MFCC3_mean* moved into this group in the validation set experiment. However, a closer examination of the first and second half of the features' importance order, reveals that the mean values of both BLA and MFCC features dominate the top half of the validation and test set in the SHAP figures. On the other hand, knowing if a participant has a cold, is in pain, or has slime in the throat during the recording does not appear to be a high-ranked feature for the classification of COPD and HC groups since these conditions happen to be the least important; this analogy is valid for both the validation and test sets.

5. Discussion

Introducing ML-based classification of COPD using voice and the collection of a Swedish COPD voice dataset were the aims of this study. Three distinct ML models were experimented on using a dataset

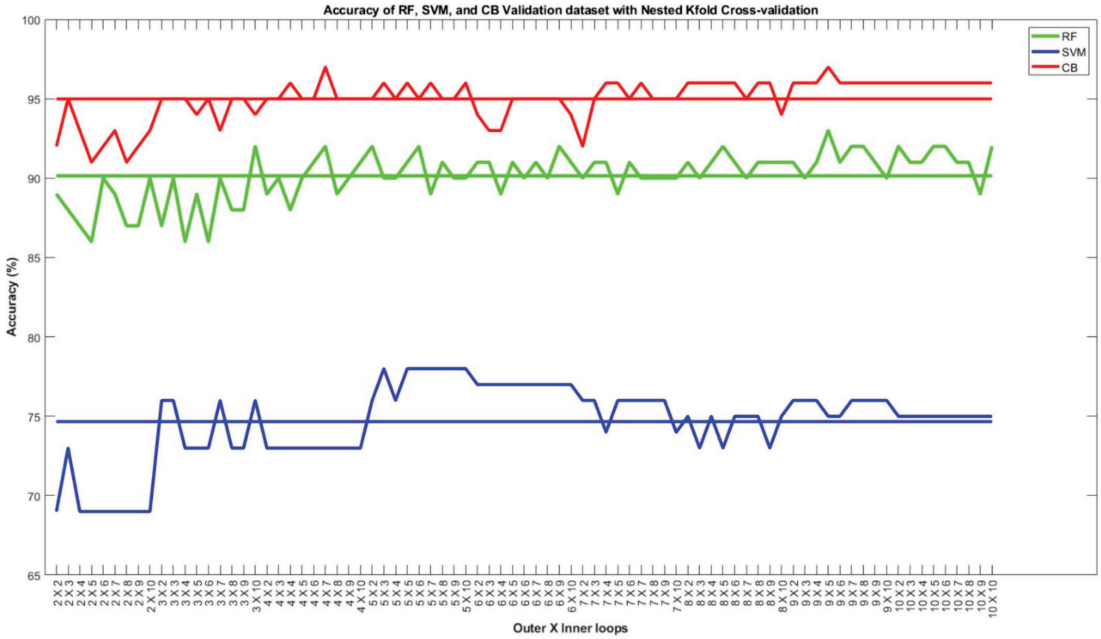


Fig. 5. Average RF, SVM, and CB accuracy over all nCV combinations for the validation set. Different colors refer to different ML models where one color indicates both the achieved accuracy for each nCV combination and the average accuracy for related ML models with flat lines.

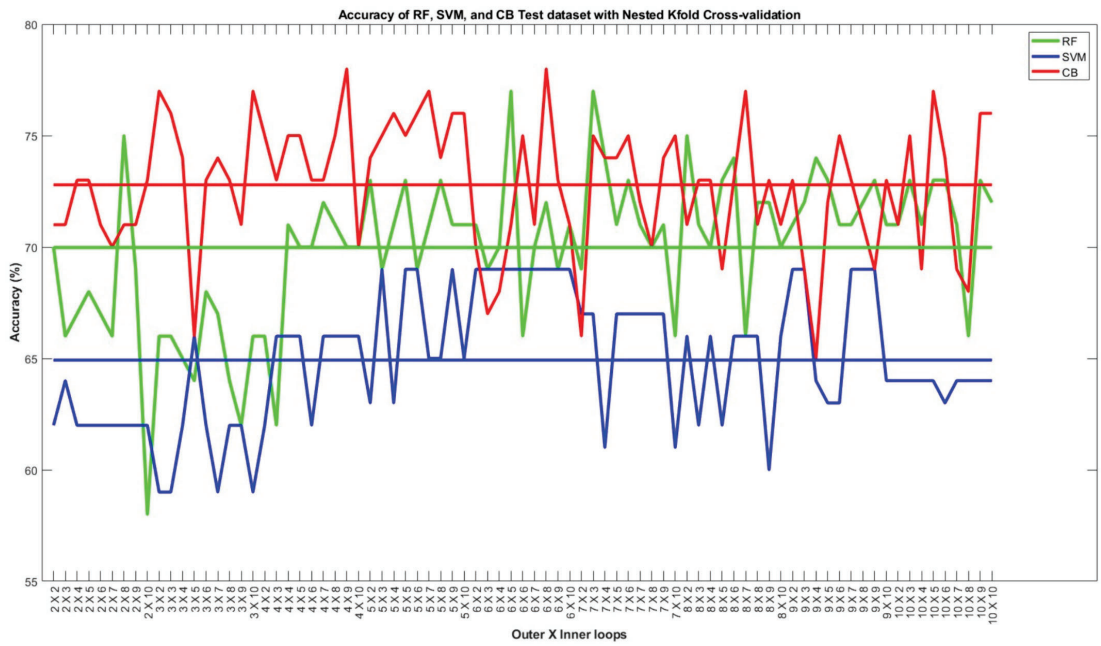


Fig. 6. Average RF, SVM, and CB accuracy over all nCV combinations for the test set. Different colors refer to different ML models where one color indicates both the achieved accuracy for each nCV combination and the average accuracy for related ML models with flat lines.

Table 3
The highest, average and standard deviation scores (STD) scores of each ML model are shown on the training, validation, and test sets, respectively. The best-performing values are highlighted in bold.

Model	Precision (%)		Recall (%)		Accuracy (%)		F1_Score (%)	
	AVG/(STD)	MAX	AVG/(STD)	MAX	AVG/(STD)	MAX	AVG/(STD)	MAX
Training set								
CB	**99/(0.5)**	**100**	**99/(0.5)**	**100**	**99/(0.5)**	**100**	**99/(0.5)**	**100**
RF	98/(0.7)	100	100/(0.2)	100	99/(0.5)	100	99/(0.5)	100
SVM	81/(2.0)	100	84/(1.7)	100	86/(1.8)	100	82/(1.9)	100
Validation set								
CB	**95/(1.3)**	**96**	**95/(1.2)**	**97**	**95/(1.3)**	**97**	**95/(1.4)**	**97**
RF	90/(1.6)	93	90/(1.5)	93	90/(1.5)	93	90/(1.5)	93
SVM	74/(2.5)	77	74/(2.4)	77	75/(2.4)	78	74/(2.5)	77
Test set								
CB	**74/(2.8)**	**79**	73/**(2.7)**	**78**	**73/(2.9)**	**78**	**73/(2.29)**	**78**
RF	72/(3.0)	79	71/(3.3)	78	70/(3.4)	77	70/(3.6)	77
SVM	66/(3.1)	70	65/(3.2)	70	65/(3.0)	69	65/(2.8)	69

Table 4
Confusion Matrix results of each ML classifier for the validation and test sets.

	CB		RF		SVM	
	+	−	+	−	+	−
Validation set						
Positive (+)	101	7	98	10	85	23
Negative (−)	2	73	6	69	20	55
Test set						
Positive (+)	50	25	48	27	45	30
Negative (−)	7	61	6	62	15	53

The table presents the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values for each algorithm.

Table 5
Comparison of CB vs. state-of-the-art studies.

Study	Classifier	Rec. type	Validation	Accuracy (%)
[31]	SVM	Speech	–	75
[32]	Random Forest	Speech	K-Fold CV	75
[33]	Random Forest	Speech	K-Fold CV	74
[34]	Random Forest	Vowel "a"	Leave-One-Subject-Out	68
[35]	Random Forest	Vowel "a"	Leave-One-Subject-Out	77
Proposed	**CatBoost**	**Vowel "a"**	**Nested CV**	**78**

consisting of sociodemographic data, BLA, and MFCC features for the detection of COPD. The features were also analyzed for their importance and impact on the model. All these analyses led to a better understanding of how recordings of vowel "a" utterance can be used as a digital biomarker for the classification of COPD and HC individuals and suggest what features can be exploited.

5.1. Performance

Our results show that among three distinct classifiers, RF, SVM, and CB, the best results were achieved by the CB classifier, with 78 % accuracy for the test set, 97 % accuracy for the validation set, and 100 % accuracy for the training set as the highest. Similarly, the CB classifier generated the highest average accuracy for all nCV combinations for both the test, validation, and training sets, with 73 %, 95 %, and 99 %, respectively. These results indicate that the vowel "a" utterance carries enough information to distinguish people suffering from COPD from healthy individuals. Furthermore, the CB classifier can capture this information with high accuracy and precision. Additionally, this also confirms the rich informative content of a vowel utterance [28,48], which was named in the introduction and is applicable to COPD voice.

Compared to previous studies, the test set results in this study, representing isolated data, are shown to outperform the previous results [31–35]. Additionally, the studies that utilize speech recordings reported lower accuracy compared to this study [31–33]. This might be associated with the highly informative content of the vowel "a" utterance as well as the use of MFCC features, which form the distinguishing aspect from previous studies. This study provides support for researchers analyzing the acoustic relationships with COPD statistically [28,29]. This is demonstrated by the high deterministic importance of jitter and shimmer features, as indicated in the feature importance graphs in this study.

The results do not change for other metrics in Table 3, where the CB classifier outperforms the RF and SVM models. Not having higher variations between accuracy, F1-score, Precision, and Recall, indicates a well-established balance between datasets used for training, validation, and testing, which is an essential aspect not only from an applied health technology point of view but even from a model generalizability perspective. Having the AUC score of 82 % for positive labels in the dataset indicates that the model can predict positive cases with higher precision and 76 % AP, which is a close value to accuracy and F1-score, also indicates a good balance in the datasets. All these results complement the work done previously in the field [34,35]. However, the test set mentioned in earlier studies is what we are calling the validation set in this study because of how the data is divided. This means that the chunks of data mentioned in those previous studies might have come from the same people in both the training and test sets. So, it indicates that the individuals in the test set were not completely separate from those in the training set, which is not the case in our data set-up. Furthermore, utilizing MFCC features, experimenting with several ML algorithms, and focusing on COPD explicitly provides additional and narrower knowledge.

The observed accuracy variation between the validation and the test set provides information on how well the classifiers learn from input data. The validation set refers to recordings isolated from the training set. However, a participant may have recordings from different time stamps in training and validation sets. The validation results indicate that a participant whose data was used for the training can be classified with a high accuracy. This information shows that the individual differences in recordings do not affect the classification accuracy. This finding is also valuable information from the self-recording application's

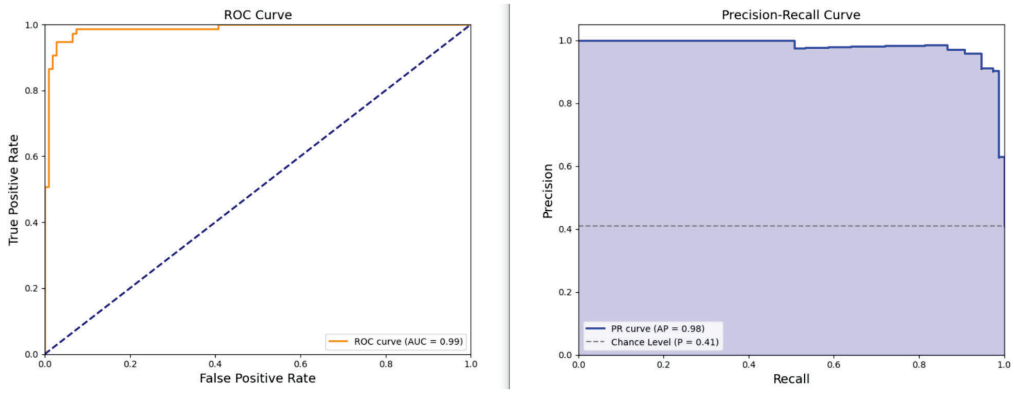


Fig. 7. Receiver Operating Characteristics (ROC) and Precision-Recall (PR) curves for the CB classifier on the validation set.

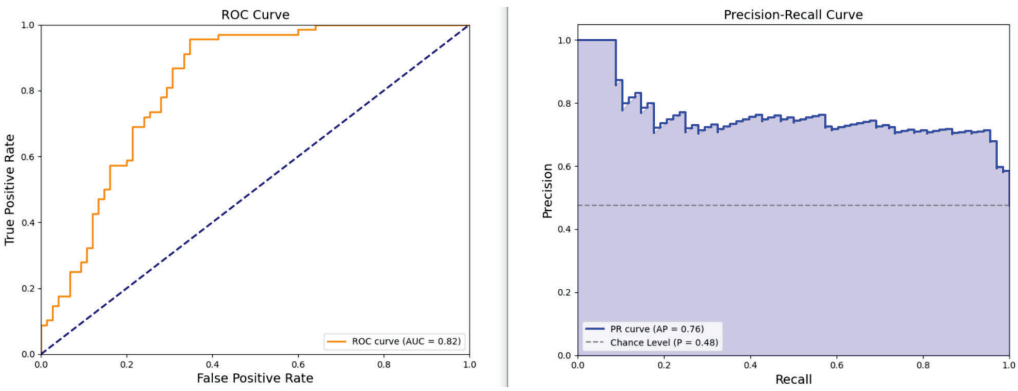


Fig. 8. Receiver Operating Characteristics (ROC) and PR curves for the CB classifier on the test set.

perspective, highlighting that mobile applications may be useful for collecting voice information to support decision-making in clinical settings. One important aspect to remember here is that the participants in this study were informed to perform recordings in quiet settings, which might be crucial for the possible deployment of voice-based decision support systems for voice-based COPD classification. Consequently, the higher variations in the test set accuracy curve for different nCV combinations, may indicate the limited number of participants. Meaning that the data used for the training was not large enough to stabilize the accuracy curve over each nCV combination for unseen data. This phenomenon can be evaluated by increasing the cohort of participants. The data partitioning performed in this experiment is aligned with best practices for ML algorithm development, which provides guidelines for building robust ML models for unbiased results [49].

The performance of the proposed approach provides insights that can be speculated since such a system is useful for continuous monitoring of patients with already known COPD diagnoses whose voices have been used for the training of the ML classifier. Of course, this is a potential venue worth investigating. In the case of monitoring, the severity of COPD is necessary to be taken into consideration to evaluate the progression of the disease, which requires ML models to be trained and tested on the data using severity information.

5.2. Features

Several BLA features seem to play a crucial role in the prediction output. Similarly, the MFCC features also show a deterministic importance together with the Age and Duration characteristics. Nonetheless, mainly having the mean values of features in the top half of the feature importance ranking suggests that the average values of features carry more information than the standard deviation. This means that the dispersity of the values in relation to the mean is less deterministic of the prediction output. Information on whether the participant has a cold, is in pain, or has slime in the throat during the recording does not seem to pose that much importance, which places emphasis on objectively measured acoustic features rather than relying on subjective self-assessment of experienced symptoms like feeling cold or experiencing pain. This information encourages the use of self-recording applications.

Having 107 features utilized for the experiment may give an intuition that the deployment of such a system has the potential to introduce operational difficulties in terms of the number of data to enter. Practically, most of the features and operations can be performed automatically on a remote server. The only required tool for the proposed system is a mobile application enabling voice recording and a connection to the server. The demographic data used in the features set might be a part of the registration process for the system, which automatically composts all the necessary parameters into a feature set for the classification task. The implementation of this system not only constitutes a supplementary

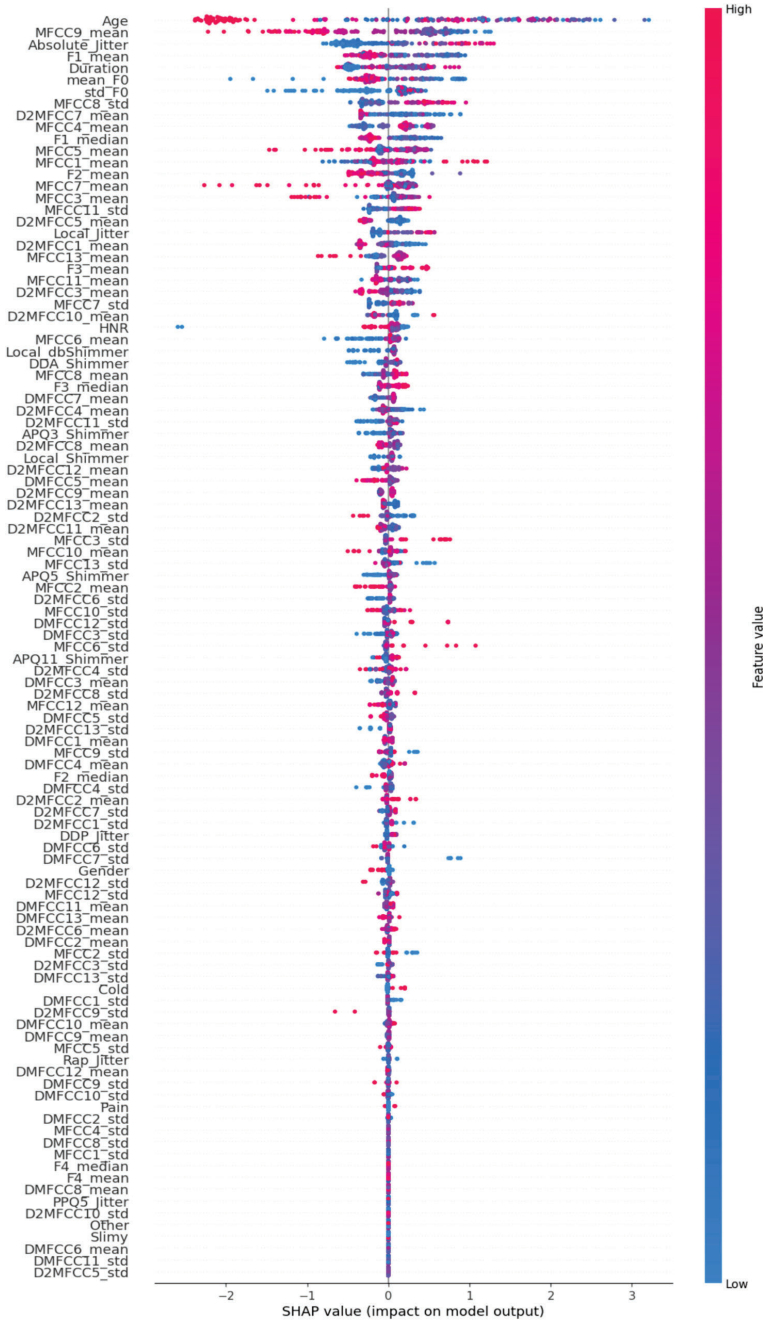


Fig. 9. The features are sorted in order of importance and their impact on the output for the validation set using SHAP feature ranking.

feature to clinical decision support system [50] but also possesses the potential to intricately contribute to the amelioration of the documented underdiagnosis and misdiagnosis rates [23,24].

It is evident that if case feature selection is considered, precautions may be taken to mitigate the risk of performance loss. In other words, if the purpose is an attempt for the classification of COPD, it is critical to ensure the inclusion of the most important features into the feature set. Consequently, since the feature importance's order changes between the

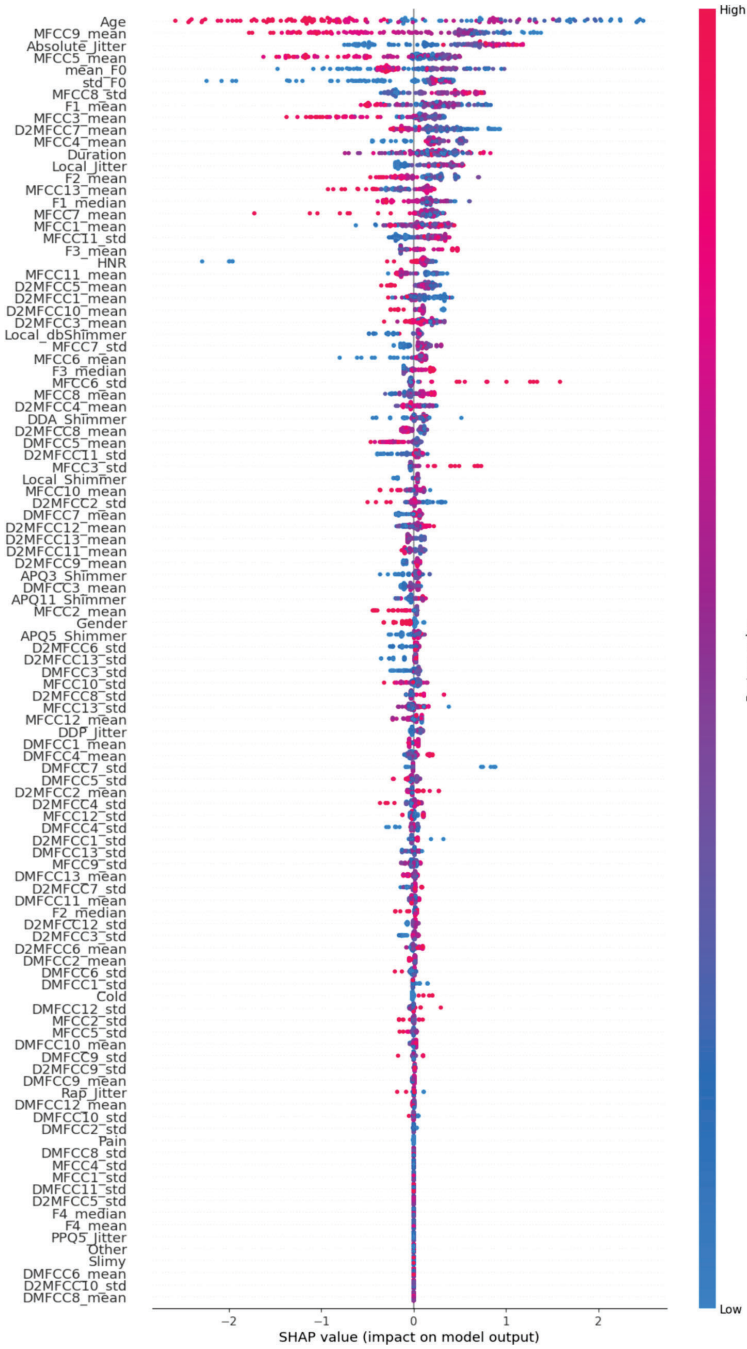


Fig. 10. The features are sorted in order of importance and their impact on the output for the test set using SHAP feature ranking.

validation and test groups, the feature selection approach may introduce a performance loss for the validated users, meaning patients already in the system (e.g., a person being monitored for COPD but not yet diagnosed or has a risk of developing COPD). Therefore, starting from dispatching the least important features of both the test and validation set could be beneficial.

5.3. Limitations

Despite all the satisfactory results of this study, awareness of the limitations regarding the small cohort of participants from the Blekinge region and having only Swedish-speaking participants is important. Furthermore, we recognize the necessity of validating our model with data from multiple centers to ensure its applicability and effectiveness across diverse populations and healthcare settings. From the generalizability point of view, the ML model cannot be used as a universal solution for the classification of COPD in its current stage. However, the method can form a backbone framework for building more generalizable models by including a larger population of participants. Another limitation is that it is computationally expensive to train and develop this kind of systems, as stated in [44] for methods using nCV, which was also observed during the experimentation in this study, where it took almost two weeks to train the models for a single performed experiment.

5.4. Future research

Future research may address the experimentation of more generalizable models and the inclusion of more participants. Additionally, investigating the diversity between COPD and other voice-affecting conditions is an essential step to achieve better performance and precision. Due to the limited number of participants and recordings, this study did not consider using ANN or deep learning, which are data-hungry techniques that may require employing different operations such as data augmentation or extending the data set using segmentation. Moreover, there is a need for further research to investigate the potential for monitoring COPD. To achieve this goal, it is crucial to conduct experiments to assess the classifiers' performance and their ability to predict various stages of COPD progression accurately.

6. Conclusion

The results indicate a potential for vocal features extracted from the utterance of the vowel "a" to be used as a digital biomarker for classifying COPD. Despite the superiority of CB, all three ML models achieved satisfactory performance levels with some degree of variations, demonstrating the binary classification efficiency. An accuracy range between 69 % to 78 % for the test and 78 % to 97 % for the validation sets indicates that the vowel "a" utterance collects enough information to be extracted and utilized as a biomarker for discriminating COPD voice from the normal voice, which supports the results of a previous study [28].

Automated COPD classification systems should carefully select features and tasks as feature importance may vary between the test and validation experiments. Such systems could be useful for healthcare professionals and individuals at risk of COPD. The system could offer a low-cost, user-friendly tool for remote COPD classification via a mobile interface. Moreover, there is a potential for self-detection of COPD, though this deployment requires further research and algorithm training on larger groups.

Overall, the study suggests that an automated decision support system for COPD diagnosis could benefit healthcare systems and patients by providing quicker and better service with fewer resources.

Registration

Clinical trial registration for the study was made both locally at Blekinge Institute of Technology and on clinicaltrials.gov with the **DNR:** BTH-6.1.1-0169-2023 and ID NCT05897944, respectively.

CRediT authorship contribution statement

Alper Idrisoglu: Writing – original draft, Conceptualization. **Ana Luiza Dallora:** Writing – review & editing, Supervision. **Abbas Cheddad:** Writing – review & editing, Supervision. **Peter Anderberg:** Writing – review & editing, Supervision. **Andreas Jakobsson:** Writing – review & editing, Resources. **Johan Sanmartin Berglund:** Writing – review & editing, Supervision, Resources.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Alper Idrisoglu reports financial support provided by the Excellence Center at Linköping and Lund in Information Technology (ELLIIT). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

An anonymized version of the dataset collected for the study will be made available upon a reasonable request from the affiliated department at Blekinge Institute of Technology. Additionally, the source code for the experiment can be reached at AIITPlanet/Code (github.com).

Acknowledgments

The authors would like to thank the Excellence Center at Linköping – Lund in Information Technology (ELLIIT) for funding and supporting this project. The authors also value the participants' keen interest, active involvement, and their significant contributions to the research.

Appendix 1

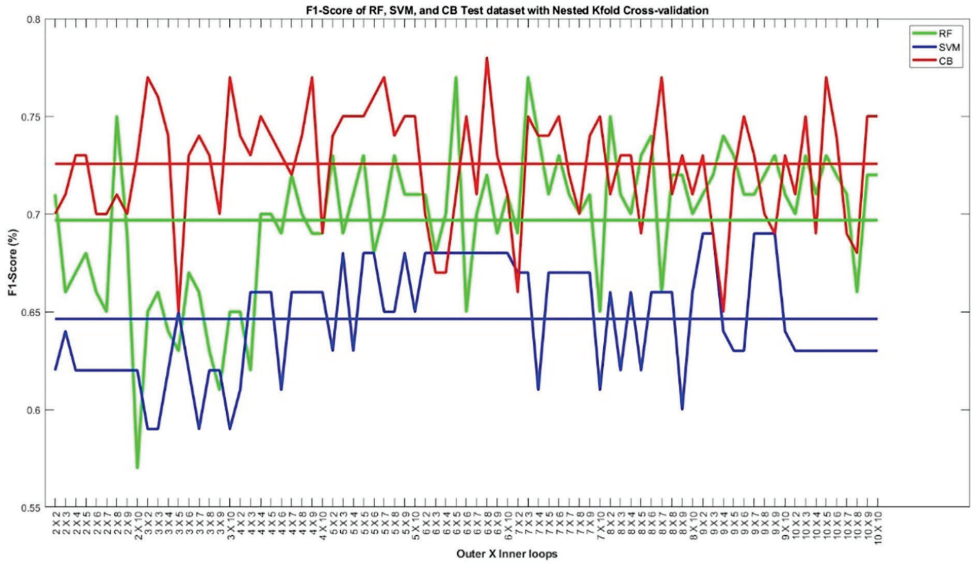


Fig. A1. Achieved average RF, SVM., and CB F1-Score over all nCV combinations for the test set. Different colors refer to different ML models where one color indicates both the achieved accuracy for each nCV combination and the average accuracy for related ML models with flat lines.

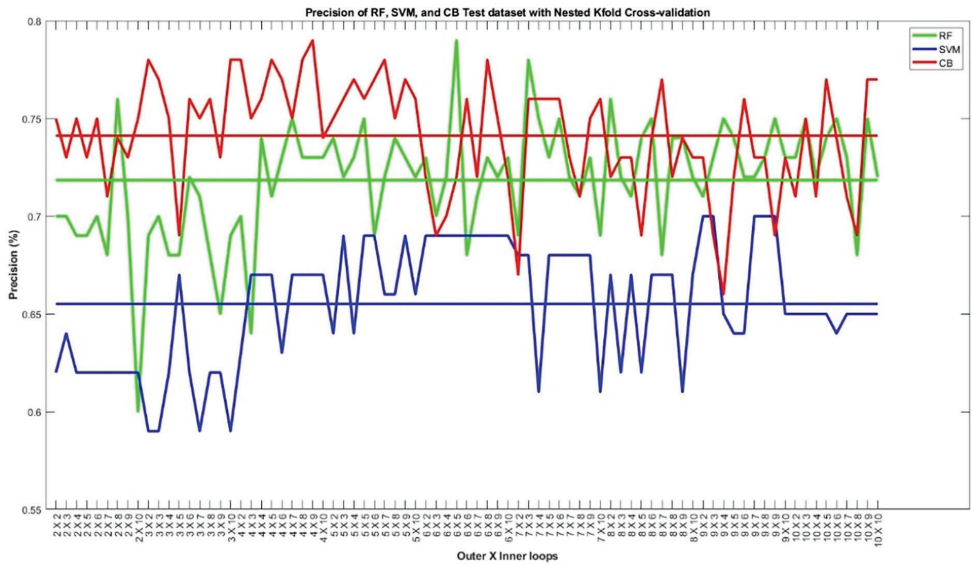


Fig. A2. Achieved average RF, SVM., and CB precision over all nCV combinations for the test set. Different colors refer to different ML models where one color indicates both the achieved accuracy for each nCV combination and the average accuracy for related ML models with flat lines.

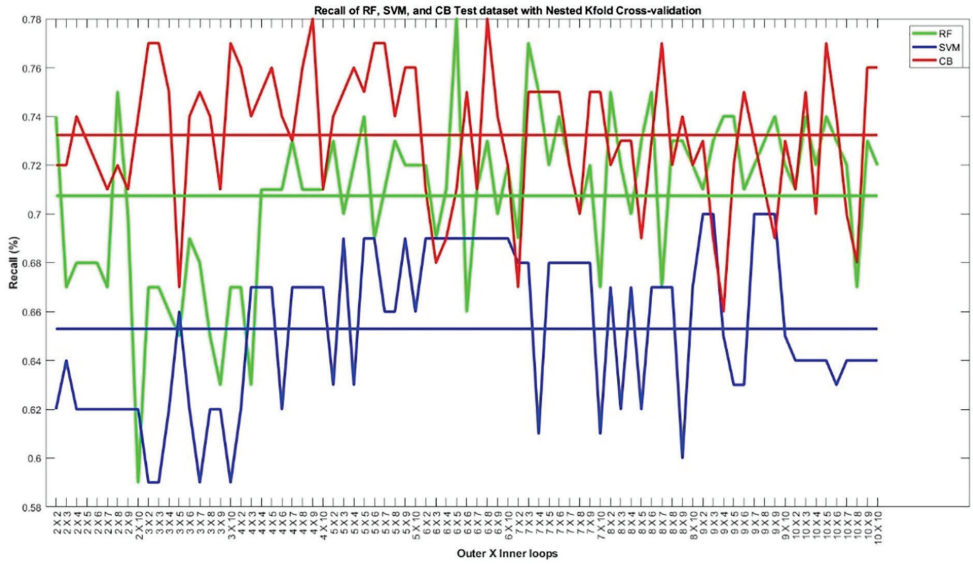


Fig. A3. Achieved average RF, SVM., and CB recall over all nCV combinations for the test set. Different colors refer to different ML models where one color indicates both the achieved accuracy for each nCV combination and the average accuracy for related ML models with flat lines.

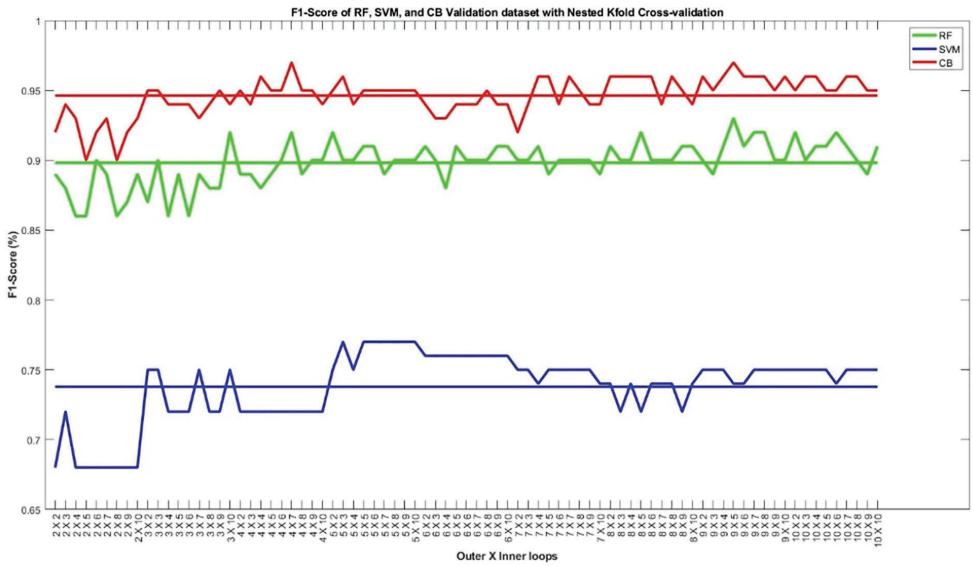


Fig. A4. Achieved average RF, SVM., and CB F1-Score over all nCV combinations for the validation set. Different colors refer to different ML models where one color indicates both the achieved accuracy for each nCV combination and the average accuracy for related ML models with flat lines.

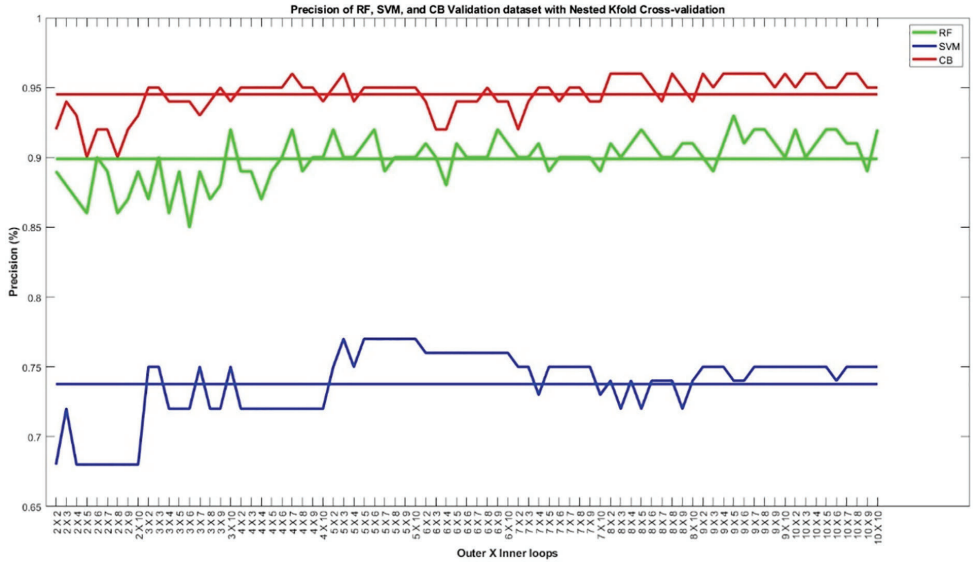


Fig. A5. Achieved average RF, SVM., and CB precision over all nCV combinations for the validation set. Different colors refer to different ML models where one color indicates both the achieved accuracy for each nCV combination and the average accuracy for related ML models with flat lines.

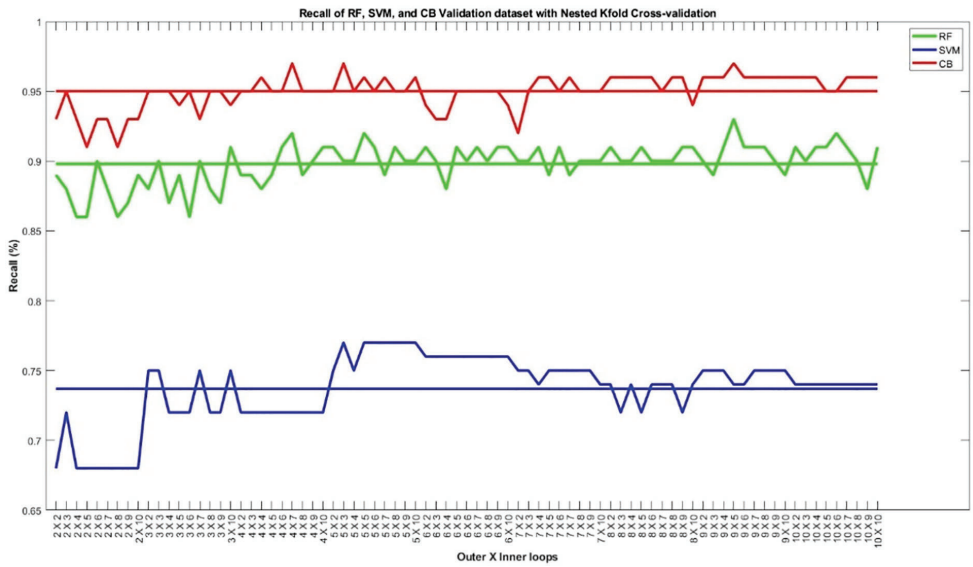


Fig. A6. Achieved average RF, SVM., and CB recall over all nCV combinations for the validation set. Different colors refer to different ML models where one color indicates both the achieved accuracy for each nCV combination and the average accuracy for related ML models with flat lines.

References

[1] Salvi SS, Barnes PJ. Chronic obstructive pulmonary disease in non-smokers. Lancet 2009;374:733–43. https://doi.org/10.1016/S0140-6736(09)61303-9.

[2] Varmaghani M, Dehghani M, Heidari E, Sharifi F, Saeedi Moghaddam S, Farzadfar F. Global prevalence of chronic obstructive pulmonary disease: systematic review and meta-analysis. East Mediterr Health J 2019;25:47–57. https://doi.org/10.26719/emhj.18.014.

[3] Boers E, Barrett M, Su JG, Benjafield AV, Sinha S, Kaye L, et al. Global burden of chronic obstructive pulmonary disease through 2050. JAMA Netw Open 2023;6: e2346598. https://doi.org/10.1001/jamanetworkopen.2023.46598.

[4] Chronic obstructive pulmonary disease (COPD) [cited 13 Nov 2023]. Available: https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd).

[5] Toren K, Olin A-C, Lindberg A, Vikgren J, Brandberg J, Johnsson Å, et al. Vital capacity and COPD: the Swedish CArdioPulmonary bioImage Study (SCAPIS). Int J Chron Obstruct Pulmon Dis 2016:927. https://doi.org/10.2147/COPD.S104644.

[6] Global initiative for chronic obstructive lung disease. In: global initiative for chronic obstructive lung disease - gold [internet] [cited 6 May 2024]. Available: https://goldcopd.org/.
[7] Corlateanu A, Mendez Y, Wang Y, de JA Garnica R, Botnaru V, Siafakas N. Chronic obstructive pulmonary disease and phenotypes: a state-of-the-art. Pulmonology 2020;26:95–100. https://doi.org/10.1016/j.pulmoe.2019.10.006.
[8] Agustí AGN, Noguera A, Sauleda J, Sala E, Pons J, Busquets X. Systemic effects of chronic obstructive pulmonary disease. Eur Respir J 2003;21:347–60. https://doi.org/10.1183/09031936.03.00405703.
[9] Idrisoglu A, Dallora AL, Anderberg P, Sanmartin Berglund J. Applied machine learning techniques to diagnose voice-affecting conditions and disorders: a systematic literature review. JMIR Med Inform 2023:18.
[10] Waring J, Lindvall C, Umeton R. Automated machine learning: review of the state-of-the-art and opportunities for healthcare. Artif Intell Med 2020;104:101822. https://doi.org/10.1016/j.artmed.2020.101822.
[11] Patel H, Shah H, Patel G, Patel A. Hematologic cancer diagnosis and classification using machine and deep learning: state-of-the-art techniques and emerging research directives. Artif Intell Med 2024;152:102883. https://doi.org/10.1016/j.artmed.2024.102883.
[12] Mahesh B. Machine learning algorithms - a review. Int J Sci Res 2018:9. https://doi.org/10.21275/ART20203995.
[13] Flach P. Machine learning: the art and science of algorithms that make sense of data. Cambridge University Press; 2012.
[14] Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. Electron Mark 2021;31:685–95. https://doi.org/10.1007/s12525-021-00475-2.
[15] Liu J, Lei J, Ou Y, Zhao Y, Tuo X, Zhang B, et al. Mammography diagnosis of breast cancer screening through machine learning: a systematic review and meta-analysis. Clin Exp Med 2022. https://doi.org/10.1007/s10238-022-00895-0 [cited 30 Aug 2023].
[16] Singh G, Chaturvedi P, Shrivastava A, Vikram Singh S. Breast cancer screening using machine learning models. In: 2022 3rd international conference on intelligent engineering and management (ICIEM); 2022. p. 961–7. https://doi.org/10.1109/ICIEM54221.2022.9853047.
[17] Cheddad A, Czene K, Eriksson M, Li J, Easton D, Hall P, et al. In: Xiong M, editor. Area and volumetric density estimation in processed full-field digital mammograms for risk assessment of breast cancer. Plos one; 2014. p. 9. e110690, https://doi.org/10.1371/journal.pone.0110690.
[18] Cheddad A. Machine learning in healthcare: breast cancer and diabetes cases. In: Reis T, Bornschlegl MX, Angelini M, Hemmje ML, editors. Advanced visual interfaces supporting artificial intelligence and big data applications. Cham: Springer International Publishing; 2021. p. 125–35. https://doi.org/10.1007/978-3-030-68007-7_8.
[19] Kaplan A, Cao H, FitzGerald JM, Iannotti N, Yang E, Kocks JWH, et al. Artificial intelligence/machine learning in respiratory medicine and potential role in asthma and COPD diagnosis. J Allergy Clin Immunol Pract 2021;9:2255–61. https://doi.org/10.1016/j.jaip.2021.02.014.
[20] Huang P, Lin CT, Li Y, Tammemagi MC, Brock MV, Atkar-Khattra S, et al. Prediction of lung cancer risk at follow-up screening with low-dose CT: a training and validation study of a deep learning method. Lancet Digit Health 2019;1:e353–62. https://doi.org/10.1016/S2589-7500(19)30159-1.
[21] Karimi D, Warfield SK, Gholipour A. Transfer learning in medical image segmentation: new insights from analysis of the dynamics of model parameters and learned representations. Artif Intell Med 2021;116:102078. https://doi.org/10.1016/j.artmed.2021.102078.
[22] Shailaja K, Seetharamulu B, Jabbar MA. Machine learning in healthcare: a review. In: 2018 second international conference on electronics, communication and aerospace technology (ICECA); 2018. p. 910–4. https://doi.org/10.1109/ICECA.2018.8474918.
[23] Ho T, Cusack RP, Chaudhary N, Satia I, Kurmi OP. Under- and over-diagnosis of COPD: a global perspective. Breathe 2019;15:24–35. https://doi.org/10.1183/20734735.0346-2018.
[24] Casas Herrera A, Montes de Oca M, López Varela MV, Aguirre C, Schiavi E, Jardim JR, et al. In: Chotirmall SH, editor. COPD underdiagnosis and misdiagnosis in a high-risk primary care population in four Latin American countries. a key to enhance disease diagnosis: the PUMA study. Plos one; 2016. p. 11. e0152266, https://doi.org/10.1371/journal.pone.0152266.
[25] Tkáč J, Man SFP, Sin DD. Review: systemic consequences of COPD. Ther Adv Respir Dis 2007;1:47–59. https://doi.org/10.1177/1753465807082374.
[26] Sin DD, Anthonisen NR, Soriano JB, Agusti AG. Mortality in COPD: role of comorbidities. Eur Respir J 2006;28:1245–57. https://doi.org/10.1183/09031936.00133805.
[27] Verdolini K, Rosen CA, Branski RC, editors. Classification manual for voice disorders–I. New York: Psychology Press; 2005. https://doi.org/10.4324/9781410617293.
[28] Shastry A, Balasubramanium RK, Acharya PR. Voice analysis in individuals with chronic obstructive pulmonary disease. Int J Phonosurgery Laryngol 2014;4:45–9. https://doi.org/10.5005/jp-journals-10023-1081.
[29] Mohamed EE, El maghraby RA.. Voice changes in patients with chronic obstructive pulmonary disease. Egypt J Chest Dis Tuberc 2014;63:561–7. https://doi.org/10.1016/j.ejcdt.2014.03.006.
[30] Kapetanidis P, Kalioras F, Tsakonas C, Tzamalis P, Kontogiannis G, Karamanidou T, et al. Respiratory diseases diagnosis using audio analysis and artificial intelligence: a systematic review. Sensors 2024;24:1173. https://doi.org/10.3390/s24041173.
[31] Nallanthighal VS, Härmä A, Strik H. Detection of COPD exacerbation from speech: comparison of acoustic features and deep learning based speech breathing models. In: ICASSP 2022–2022 IEEE international conference on acoustics, speech and signal processing (ICASSP); 2022. p. 9097–101. https://doi.org/10.1109/ICASSP43922.2022.9747785.
[32] Farrús M, Codina-Filbà J, Reixach E, Andrés E, Sans M, Garcia N, et al. Speech-based support system to supervise chronic obstructive pulmonary disease patient status. Appl Sci 2021;11:7999. https://doi.org/10.3390/app11177999.
[33] Chun KS, Nathan V, Vatanparvar K, Nemati E, Rahman MM, Blackstock E, et al. Towards passive assessment of pulmonary function from natural speech recorded using a mobile phone. In: 2020 IEEE international conference on pervasive computing and communications (PerCom); 2020. p. 1–10. https://doi.org/10.1109/PerCom45495.2020.9127380.
[34] Nathan V, Vatanparvar K, Rahman MM, Nemati E, Kuang J. Assessment of chronic pulmonary disease patients using biomarkers from natural speech recorded by mobile devices. In: 2019 IEEE 16th international conference on wearable and implantable body sensor networks (BSN); 2019. p. 1–4. https://doi.org/10.1109/BSN.2019.8771043.
[35] Nathan V, Rahman MM, Vatanparvar K, Nemati E, Blackstock E, Kuang J. Extraction of voice parameters from continuous running speech for pulmonary disease monitoring. In: 2019 IEEE international conference on bioinformatics and biomedicine (BIBM); 2019. p. 859–64. https://doi.org/10.1109/BIBM47256.2019.8983115.
[36] Abdul Zkh, Al-Talabani AK. Mel frequency cepstral coefficient and its applications: a review. IEEE Access 2022;10:122136–58. https://doi.org/10.1109/ACCESS.2022.3223444.
[37] Ittichaichareon C, Suksri S, Yingthawornsuk T. Speech recognition using MFCC. Simul Model 2012;9:135–8. https://doi.org/10.13140/RG.2.1.2598.3208.
[38] Gupta S, Jaafar J, Wan Ahmad WF, Bansal A. Feature extraction using Mfcc. Signal Image Process Int J 2013;4:101–8. https://doi.org/10.5121/sipij.2013.4408.
[39] Agatonovic-Kustrin S, Beresford R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. J Pharm Biomed Anal 2000;22:717–27. https://doi.org/10.1016/S0731-7085(99)00272-1.
[40] Cooper AM, Kästner J, Urban A, Artrith N. Efficient training of ANN potentials by including atomic forces via Taylor expansion and application to water and a transition-metal oxide. Npj Comput Mater 2020;6:1–14. https://doi.org/10.1038/s41524-020-0323-8.
[41] Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. J Big Data 2020;7:94. https://doi.org/10.1186/s40537-020-00369-8.
[42] Huang G, Wu L, Ma X, Zhang W, Fan J, Yu X, et al. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. J Hydrol 2019;574:1029–41. https://doi.org/10.1016/j.jhydrol.2019.04.085.
[43] Parvandeh S, Yeh H-W, Paulus MP, McKinney BA. Consensus features nested cross-validation. In: Valencia A, editor. Bioinformatics. 36; 2020. p. 3093–8. https://doi.org/10.1093/bioinformatics/btaa046.
[44] Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res 2010;11:2079–107.
[45] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Advances in neural information processing systems. Curran Associates, Inc.; 2017. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.
[46] Marcílio WE, Eler DM. From explanations to feature selection: assessing SHAP values as feature selection mechanism. In: 2020 33rd SIBGRAPI conference on graphics, patterns and images (SIBGRAPI); 2020. p. 340–7. https://doi.org/10.1109/SIBGRAPI51738.2020.00053.
[47] Foody GM. Challenges in the real world use of classification accuracy metrics: from recall and precision to the Matthews correlation coefficient. PloS One 2023;18:e0291908. https://doi.org/10.1371/journal.pone.0291908.
[48] Boyanov B, Hadjitodorov S. Acoustic analysis of pathological voices. A voice analysis system for the screening of laryngeal diseases. IEEE Eng Med Biol Mag 1997;16:74–82. https://doi.org/10.1109/51.603651.
[49] Wujek B, Hall P, Güneş F. Best practices for machine learning applications. SAS Inst Inc 2016:1–23.
[50] Alanazi A. Using machine learning for healthcare challenges and opportunities. Inform Med Unlocked 2022;30:100924. https://doi.org/10.1016/j.imu.2022.100924.

Abbreviations

ANN: artificial neural network
AUC: Area Under the Curve
AP: Average Precision
BLA: base line acoustic
BTH: Blekinge Institute of Technology
CB: CatBoost
CDSS: clinical decision support system

CT: Computerized Tomography
COPD: chronic obstructive pulmonary disease
FEV1: forced expiratory volume in one second
FVC: forced vital capacity
Ml: Machine Learning
MFCC: Mel-Frequency Cepstral Coefficient
nCV: nested cross validation
PR: Precision Recall
RF: Random Forest
SHAP: SHapley Additive exPlanations
SVM: Support Vector Machine
VC: vital capacity
WHO: World Health Organization

Study III

Vowel Segmentation Impact on Machine Learning Classification for Chronic Obstructive Pulmonary Disease

Published as:

Idrisoglu, A., Moraes, A. L. D., Cheddad, A., Anderberg, P., Jakobsson, A., & Berglund, J. S. (2025). Vowel segmentation impact on machine learning classification for chronic obstructive pulmonary disease. *Scientific Reports*, *15*(1), 9930. https://doi.org/10.1038/s41598-025-95320-3.

scientific reports

Check for updates

OPEN

Vowel segmentation impact on machine learning classification for chronic obstructive pulmonary disease

Alper Idrisoglu[1]✉, Ana Luiza Dallora Moraes[1,4], Abbas Cheddad[1,2,4], Peter Anderberg[1,4], Andreas Jakobsson[3,4] & Johan Sanmartin Berglund[1,4]

Vowel-based voice analysis is gaining attention as a potential non-invasive tool for COPD classification, offering insights into phonatory function. The growing need for voice data has necessitated the adoption of various techniques, including segmentation, to augment existing datasets for training comprehensive Machine Learning (ML) modelsThis study aims to investigate the possible effects of segmentation of the utterance of vowel "a" on the performance of ML classifiers CatBoost (CB), Random Forest (RF), and Support Vector Machine (SVM). This research involves training individual ML models using three distinct dataset constructions: full-sequence, segment-wise, and group-wise, derived from the utterance of the vowel "a" which consists of 1058 recordings belonging to 48 participants. This approach comprehensively analyzes how each data categorization impacts the model's performance and results. A nested cross-validation (nCV) approach was implemented with grid search for hyperparameter optimization. This rigorous methodology was employed to minimize overfitting risks and maximize model performance. Compared to the full-sequence dataset, the findings indicate that the second segment yielded higher results within the four-segment category. Specifically, the CB model achieved superior accuracy, attaining 97.8% and 84.6% on the validation and test sets, respectively. The same category for the CB model also demonstrated the best balance regarding true positive rate (TPR) and true negative rate (TNR), making it the most clinically effective choice. These findings suggest that time-sensitive properties in vowel production are important for COPD classification and that segmentation can aid in capturing these properties. Despite these promising results, the dataset size and demographic homogeneity limit generalizability, highlighting areas for future research.

Trial registration **The study is registered on clinicaltrials.gov with ID: NCT06160674.**

Keywords Classification, Chronic obstructive pulmonary disease (COPD), Machine learning, Vowel segmentation

Chronic Obstructive Pulmonary Disease (COPD) is a progressive respiratory disorder characterized by a gradual diminution of airflow and lung tissue deterioration. It has emerged as a significant global health concern, ranking as the third leading cause of mortality and morbidity worldwide[1,2]. In 2015, approximately 174 million individuals were diagnosed with COPD, with an estimated 3.2 million deaths, likely underestimated due to high underdiagnosis rates[3,4]. COPD is not only related to pulmonary problems; it is known to have systemic effects[5], meaning that having COPD may lead to the malfunction in other organs. Even though the main evaluation is based on spirometry and Computerized Tomography (CT)[6], recent research has investigated the possibility of using systemic effects to support decision-making processes[7,8]. COPD is known to affect voice production[9,10], which has increased interest in investigating the potential utilization of various vocal parameters as decision-support cues for COPD diagnosis[11,12]. The underlying premise is to use machine learning (ML) algorithms to

[1]Department of Health, Blekinge Institute of Technology, 371 41 Karlskrona, Sweden. [2]Institute of Computer Science, University of Tartu, Narva mnt 18, 51009 Tartu, Estonia. [3]Mathematical Statistic, Lund University, 221 00 Lund, SE, Sweden. [4]Ana Luiza Dallora Moraes, Abbas Cheddad, Peter Anderberg, Andreas Jakobsson and Johan Sanmartin Berglund contributed equally to this work. ✉email: alper.idrisoglu@bth.se

extract latent information embedded within vocal characteristics to support the decision-making process for early diagnosis.

Speech processing encompasses various techniques, including noise reduction methods to enhance signal clarity[13], feature extraction approaches to analyze vocal characteristics[14], and strategies like additive white noise to improve model robustness under different SNR levels[15]. The process of extracting information from voice entails the mathematical computation of attributes associated with individual voice samples, commonly referred to as voice, vocal features, and vocal biomarkers[16]. These features can be derived from time, frequency, and spectral representations of the raw voice recordings, such as baseline acoustic features (BLA), Jitter and Shimmer, and Mel Frequency Cepstral Coefficients (MFCC), which are techniques that emanate from voice recognition and provide a foundation for research in the field of voice-based decision support systems[16,17]. The characteristics of voices differ widely. Different voice-affecting disorders influence different voice characteristics. For example, Parkinson's disease tends to affect the vowel "a" phonation; on the other hand, Alzheimer's disease influences free speech[17]. Even individual differences in voice types are highlighted in the literature referring to the unliterary and dynamic characteristic of voice production[18].

The evaluation of Artificial Intelligence (AI) for performing medical tasks is underway across various fields of practice. Using voice recordings, vocal features, and ML to diagnose disorders that affect the voice is a growing area of interest among researchers[19–21]. The idea is to extract information from voice recordings and let ML assess patterns that can be used for clinical purposes, such as detection, classification, and monitoring, to support decision-making processes[22–24]. ML is an active research area involving the systematic development of algorithms for better performance to mimic humankind's abilities based on the collected data[25,26]. Additionally, the performance of ML in complex data analyses is another factor for the increased usage in clinical research[27–31]. However, the common denominator for ML-based experiments is the demand for data, which in some cases might be challenging to work with and require additional techniques to train ML models on small datasets[32]. There are several techniques employed to expand the voice datasets to make it possible to train ML algorithms on more data, such as the collection of several recordings at the same time[33–36], using windowing with some degree of overlap to create several feature vectors from one single recording[37–39], or dividing the recording into time frames and treating each frame as a new recording[40–42]. However, these methods are applied mostly on long speech recordings, with very few studies investigating their efficacy for vowel recordings. Since vowel production exhibits dynamic characteristics[43,44], segmentation techniques tailored for vowels may have a different impact on ML performance compared to their application in continuous speech. Furthermore, shorter time frames may capture more stationary characteristics of voice signals[45], which could influence classifier performance in ways not yet fully explored. While these methods are widely employed, it is essential to explore their potential impact on the performance of ML classifier, particularly in the context of vowel-based analysis.

This article investigates whether time frame-wise differences in the utterance of the vowel 'a' collected from Swedish-speaking individuals affect the binary classification performance of ML algorithms CB, RF, and SVM in distinguishing between COPD and non-COPD voices. The aim is to apply segmentation to the utterance of the vowel 'a' to assess performance differences across individual and grouped datasets compared to the full sequence of recordings and analyze the classification results from a clinical perspective, exploring whether segmentation enables a more refined analysis of disease-related vocal characteristics and enhances the diagnostic relevance of voice-based features. The potential contributions of this study include:

- Introducing or refining a segmentation method could provide insights into how analyzing smaller segments of vocal data, rather than the entire recordings, impacts classification performance.
- A comparison of the CB, RF, and SVM on the performance effects of segmented vs. full sequence data offers valuable knowledge on which algorithms are best suited for segmentation-based extended datasets.
- By focusing on time frame-wise differences, the study may uncover whether there is a time sensitivity in recordings critical for COPD classification, which could help in developing more advanced and precise speech analysis models and signal processing frameworks.
- By analyzing time frame-wise differences in vowel utterances, the study may enhance the accuracy of COPD and non-COPD voice classification, leading to more effective early diagnosis tools using ML algorithms.

Results

This section provides an analysis of the effects of segmentation of the vowel "a" utterance for binary classification performance (COPD vs. No COPD) from an experiment involving three machine learning classifiers: CB, RF, and SVM. Confusion matrix results are provided to compare the overall accuracy metrics for segment-based and group-wise results, a clinically relevant perspective on performance. Additionally, the Receiver Operating Characteristic (ROC) curves provides a comparation of the ML results between using full-sequence dataset and segmented dataset that achieved the best performance.

Experimental results

The experiment that forms the basis of Fig. 1 yielded 15 distinct results for each machine learning classifier. The different combinations of nested cross-validation (nCV), starting from 2X2 to 5X5, have generated 16 results for each performance metric: precision, recall, accuracy, and F1 score, associated with each segment in five different categories, starting from the full sequence and ending with five equally divided segments of the same recording.

Figure 1 illustrates the distribution of the accuracy results and the highest accuracy achieved for the training validation and test datasets, displayed at the top of each boxplot and for each dataset. The classifiers CB, RF, and SVM ranked from highest to lowest based on their performance: training set accuracies of 99.9%, 99.2%, 87.8%; validation set accuracies of 97.8%, 93.4%, and 80.3%; and the test set accuracies of 84.6%, 77.6%, and 72.7%, respectively. The highest accuracies were measured mostly in segment-based results, with the exception of the

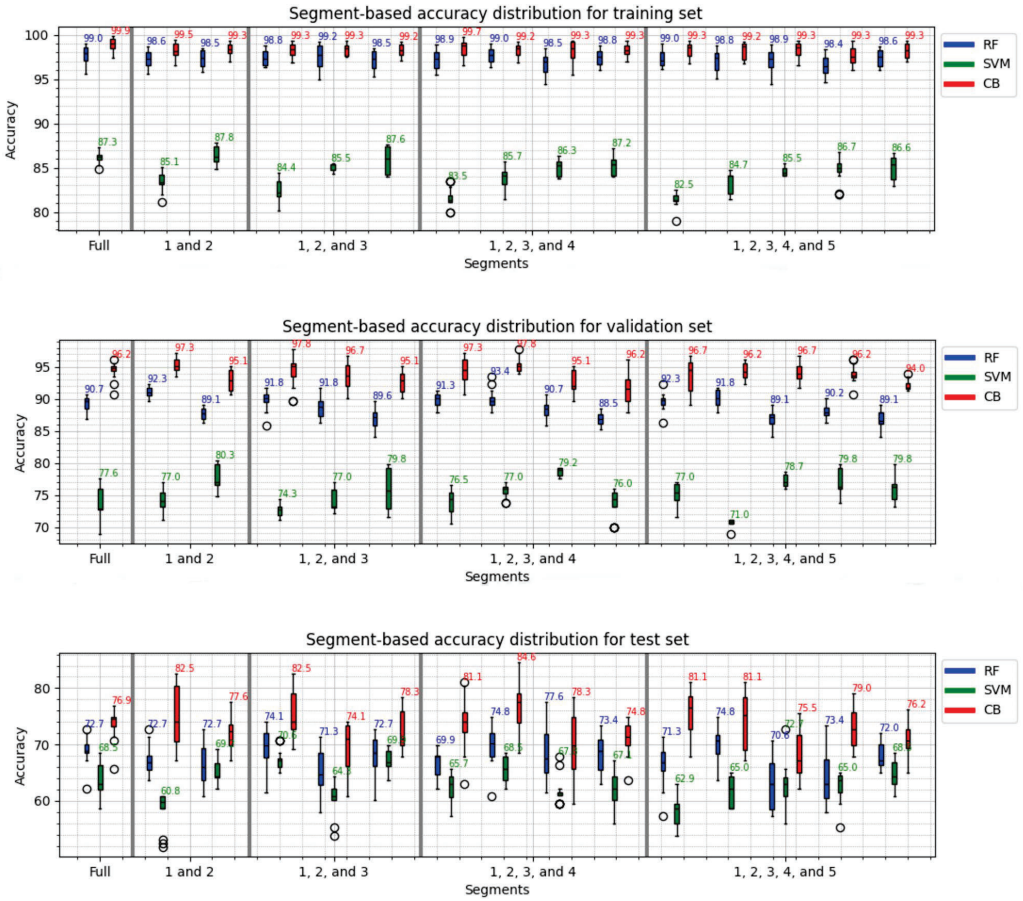


Fig. 1. Segment-wise accuracy results show the maximum accuracy achieved for each segment with three different ML classifiers for the training, validation, and test sets, where the whole sequence is divided into several segment combinations, starting from full sequence to 5 different segments.

CB classifier in the training set, where the highest accuracy of 99.9% was achieved in the full segment. Compared with the full-sequence results, an overall improvement in performance metrics was noted for segment-based results. However, the performance improvement in the test set seems to come with a cost of increased variance, which is not just across different segments but also when comparing the segmented results to the full sequences for both the validation and test sets, while the training set results look stable. Moreover, when examining the results of the validation and the test sets on a pairwise and segment-wise basis, each classifier exhibits a unique trend slope. For the validation set, the trend lines for the CB and RF classifiers exhibit similar patterns, whereas the SVM displays a divergent trend. Conversely, the test set results demonstrate a greater variation among the trend lines of all three classifiers compared to the training and validation sets. Another interesting observation is that the ensemble learning-based classifiers CB and RF show better performance and achieve higher average accuracy mostly in the first half of the recordings when divided into two halves, with performance metrics summed and averaged separately for each half. However, SVM follows the opposite trend by having the highest accuracies in the second half of the recordings for the validation set and test set.

Table 1 presents all the validation and test set results for all segments and all performance metrics associated with each ML classifier. The analysis of the table reveals distinct performance metrics for CB, RF, and SVM models across various all categories and metrics within both validation and test sets. Specifically, the CB model performed consistently better than the other models across most metrics, achieving its peak precision of 97.6%, the highest accuracy of 97.8%, and F1 scores of 97.8%, in the 3-segment category and a recall of 98.2%, in 4-segment category during validation. The RF model shows strong performance, especially in the 4-segment category for a precision of 93.0% and recall of 93.8%, and it reaches its best performance with an accuracy of

Validation set metrics															
		2 Seg		3 Seg			4 Seg				5 Seg				
Model	Full	1	2	1	2	3	1	2	3	4	1	2	3	4	5
Precision (%)															
CB	95.8	97.0	94.7	**97.6**	96.4	94.8	97.0	97.5	94.7	96.0	96.5	96.0	96.8	95.8	93.6
RF	90.5	91.9	88.6	91.3	91.3	89.1	90.8	**93.0**	90.2	88.0	91.9	91.5	88.6	89.7	88.5
SVM	76.9	76.5	**79.7**	73.8	76.4	79.1	75.8	76.4	78.8	75.2	76.5	70.3	78.1	79.1	79.1
Recall (%)															
CB	96.6	97.5	95.4	97.9	97.0	95.0	97.5	**98.2**	95.4	96.2	96.8	96.4	96.4	96.6	94.1
RF	90.7	92.7	88.9	92.0	92.0	89.8	91.6	**93.8**	90.9	88.4	92.7	92.0	89.1	90.7	89.3
SVM	76.7	75.5	**79.6**	72.3	76.3	**79.6**	76.2	75.9	79.4	74.9	75.5	70.8	77.5	79.4	79.4
Accuracy (%)															
CB	96.2	97.3	95.1	**97.8**	96.7	95.1	97.3	**97.8**	95.1	96.2	96.7	96.2	96.7	96.2	94.0
RF	90.7	92.4	89.1	91.8	91.8	89.6	91.3	**93.4**	90.7	88.5	92.4	91.8	89.1	90.2	89.1
SVM	77.6	77.1	**80.3**	74.3	77.1	79.8	76.5	77.1	79.2	76.0	77.1	71.0	78.7	79.8	79.8
F1_Score (%)															
CB	96.1	97.2	95.0	**97.8**	96.6	94.9	97.2	**97.8**	95.0	96.1	96.6	96.1	96.6	96.1	93.8
RF	90.5	92.2	88.8	91.6	91.6	89.4	91.1	**93.3**	90.5	88.2	92.2	91.6	88.8	90.0	88.8
SVM	76.8	75.8	**79.6**	72.7	76.3	79.3	75.9	76.1	78.8	75.0	75.8	70.4	77.7	79.2	79.2
Test set metrics															
Precision (%)															
CB	78.4	82.6	78.1	82.7	75.8	78.9	81.2	**84.7**	78.3	77.3	81.2	82.0	75.9	79.3	78.3
RF	75.8	72.8	73.5	75.2	71.3	74.1	70.2	75.1	**77.7**	74.8	71.7	75.5	70.6	74.3	72.6
SVM	69.2	61.6	71.3	70.7	66.2	72.3	65.7	69.0	69.4	68.9	62.8	65.3	**73.7**	66.5	70.6
Recall (%)															
CB	77.5	82.7	77.9	82.3	74.6	78.7	81.3	**84.8**	78.4	75.4	80.1	81.5	75.8	79.3	76.9
RF	73.5	72.8	73.1	74.6	71.3	73.3	70.2	75.0	**77.8**	73.7	71.6	75.2	70.6	73.8	72.4
SVM	68.9	61.3	69.9	70.8	65.0	70.7	65.7	68.8	68.4	67.8	62.8	65.2	**73.2**	65.6	69.0
Accuracy (%)															
CB	76.9	82.5	77.6	82.5	74.1	78.3	81.1	**84.6**	78.3	74.8	81.1	81.1	75.5	79.2	76.2
RF	72.7	72.7	72.7	74.1	71.3	72.7	69.9	74.8	**77.6**	73.4	71.3	74.8	70.6	73.4	72.0
SVM	68.5	60.8	69.2	70.6	64.3	69.9	65.7	68.5	67.8	67.1	62.9	65.0	**72.7**	65.0	68.5
F1_Score (%)															
CB	76.8	82.5	77.6	82.4	74.1	78.3	81.1	**84.6**	78.3	74.7	81.0	81.1	75.5	79.0	76.0
RF	72.3	72.7	72.7	74.0	71.3	72.7	69.9	74.8	**77.6**	73.4	71.3	74.8	70.6	73.4	72.0
SVM	68.5	60.8	68.9	70.6	64.3	69.6	65.7	68.5	67.6	66.8	62.8	65.0	**72.6**	64.7	68.4

Table 1. Validation and test set scores for three different ML classifiers using four different performance metrics for full sequence and each segment. The highest performance is shown in bold.

93.4% and F1 score of 93.3% within the 4-segment category. On the other hand, SVM lags behind with its best precision of 79.7%, recall of 79.6%, accuracy of 80.3%, and F1 score of 79.6%, all occurring in the 2-segment category, underlining its comparative underperformance. When shifting focus to the performance on the test set, CB's supremacy persists with the highest precision of 84.7%, a top recall rate of 84.8%, and a leading accuracy and F1 scores of 84.6% and 84.6%, respectively, all in the 4-segment category. RF maintained its highest test performance with precision, recall, accuracy, and F1-score of 77.2%, 77.8%, 77.6%, and 77.6%, respectively, in the 4-segment category. Conversely, SVM falls behind results in the test set similar to the validation set, with its highest precision of 73.7%, recall of 73.2%, accuracy of 73.7%, and F1 score of 72.6%, respectively, in the 5-segment categories. However, when comparing the performance drop between the highest validation and test set results for each classifier, the SVM classifier demonstrates a higher degree of generalizability. It shows the smallest decrease across all metrics, with drops of 6.0% in precision, 6.4% in recall, 7.6% in accuracy, and 7.0% in F1-score. The CB classifier takes second place with precision, recall, accuracy, and F1-score of 12.9%, 13.1%, 13.2%, and 13.2%, respectively. The RF classifier suffers the highest performance drop with precision, recall, accuracy, and F1-score of 15.3%, 16.0%, 15.8%, and 15.7%, respectively. The result indicates an increased level of possible overfitting within a performance drop comparison between the full sequence results of validation end test sets, where performance drop for full segment occurs for SVM, 7.7%, 7.7%, 8.1%, 8.3%, and CB, 17.4%, 19.1%, 19.3%, 19.3%, and lastly for RF, 14.7%, 17.2%, 18.0%, 18.2%, on the performance metrics precision, recall, accuracy, and F1-score, respectively.

CB outperforms RF and SVM across all categories, achieving the highest accuracy of 97.8%, precision of 97.6%, recall of 98.2%, and F1-score of 97.8% in the validation set. In the test set, CB maintains its lead with an

accuracy of 84.6%. RF follows but shows the highest performance drop, indicating potential overfitting. SVM, while the lowest-performing model, exhibits the smallest drop in performance, suggesting better generalizability.

Figure 2 presents the results when the feature vectors in each segment-based data merged into one dataset to create an expanded dataset with a factor of the number of segments in the specific group. Additionally, all-segment group is created by merging all feature vectors from all segment-based groups and the full sequence together, creating a dataset with a factor of 15 from one single recording if the recordings are divided into five segments as highest. The CB classifier performs with the highest accuracy of 100% on the training set and validation set overall groups. RF classifier follows CB with a small decrease in training and validation set accuracy of 99.5% and around 97.0% on average, respectively. However, SVM falls behind with an accuracy of 84.0% and 82.0% average for training and validation set results, respectively. Test set results do not change the ranking where CB, RF, and SVM classifiers are placed in chronological order from highest to lowest performance with accuracies of 79.3% in 5 segment group, 74.3% 5 segment group, and 69.6% in 2 segment group, respectively. In a comparison between Figs. 1 and 2, it is clear that variation in performance decreases when the grouped dataset is used. On the other hand, the higher accuracies achieved in Table 1 results are not presented in the group-wise results. Regardless of which dataset is employed, the results indicate a higher performance than using features extracted from a full sequence of vowel "a" utterance regarding the performance accuracy.

Group-wise results of all three classifiers for validation and test set are presented in Table 2. The results suggest that the CB performs the best in all metrics with a score of 100% in two segment groups, 4 and all-segment groups in validation sets. The RF classifier gives the second-best performance with the highest score of 99.6% in all metrics in the all-segment group for the validation set. The SVM classifiers take the last position in

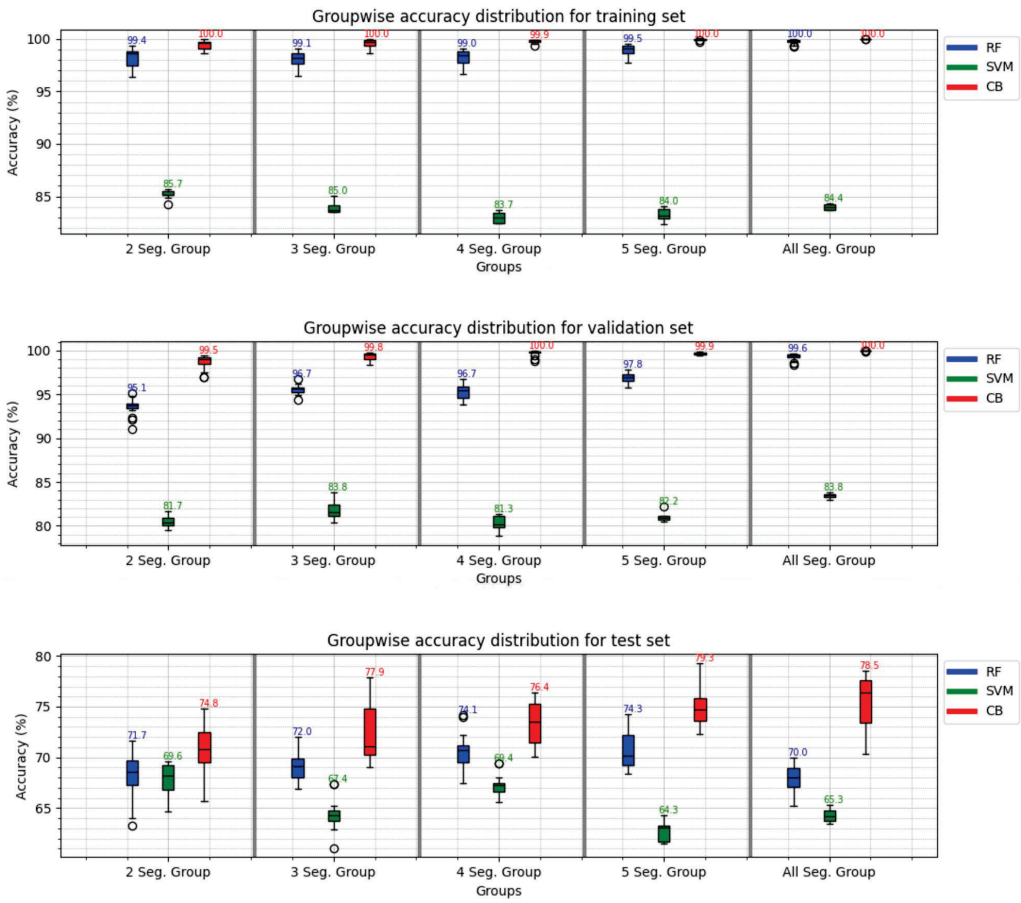


Fig. 2. Group-wise accuracy results show the max accuracy achieved in each group with three different ML classifiers for training validation and test sets, where segments in Fig. 1 are merged into a dataset to create respective groups, first from two to five and then all segments together.

Validation set metrics					
Model	**2 Seg. group**	**3 Seg. group**	**4 Seg. group**	**5 Seg. group**	**All Seg. group**
Precision (%)					
CB	99.3	99.8	**100.0**	99.9	**100.0**
RF	94.6	96.2	96.6	97.8	**99.6**
SVM	80.8	82.8	81.1	81.8	**83.3**
Recall (%)					
CB	99.6	99.9	**100.0**	99.9	**100.0**
RF	95.2	97.2	96.7	97.7	**99.6**
SVM	80.5	83.7	80.8	81.7	**83.5**
Accuracy (%)					
CB	99.5	99.8	**100.0**	99.9	**100.0**
RF	95.1	96.7	96.7	97.8	**99.6**
SVM	81.7	83.8	81.3	82.2	**83.8**
F1_Score (%)					
CB	99.4	99.8	**100.0**	99.9	**100.0**
RF	94.9	96.6	96.6	97.8	**99.6**
SVM	80.7	83.2	80.8	81.8	**83.3**
Test set metrics					
Precision (%)					
CB	75.5	78.0	76.6	80.0	**80.6**
RF	72.2	72.1	74.1	**74.2**	69.9
SVM	**70.0**	68.5	70.0	65.5	66.6
Recall (%)					
CB	75.2	78.0	76.2	**78.9**	77.9
RF	72.0	71.8	74.0	**74.3**	69.9
SVM	**69.9**	67.9	69.7	64.7	65.9
Accuracy (%)					
CB	74.8	77.9	76.4	**79.3**	78.5
RF	71.7	72.0	74.1	**74.3**	70.0
SVM	**69.6**	67.4	69.4	64.3	65.3
F1_Score (%)					
CB	74.8	77.9	76.2	**79.0**	78.0
RF	71.7	71.8	74.0	**74.2**	69.9
SVM	**69.6**	67.2	69.4	64.3	65.1

Table 2. Validation and test set scores for three different ML classifiers using four different performance metrics for each group. The highest performance is shown in bold.

performance for the validation set with the highest scores of 83.3%, 83.5%, 83.8%, and 83.3%, Precision, Recall, Accuracy, and F1-score, respectively, in the all-segment group. The group-wise results for the test set suggest that the CB classifier is the best-performing one with scores of 80.6%, 78.9%, 79.3%, and 79.0%, in precision, recall, accuracy, and F1 score, respectively, in the 5-segment group with the exception of precision that reaches the highest score in the all-segment group. RF reaches its highest precision, recall, accuracy, and F1-score of 74.2%, 74.3%, 74.3%, and 74.2% in the 5-segment group for the test set, respectively. SVM falls behind with precision, recall, accuracy, and F1 scores of 70.0%, 69.9%, 69.6%, and 69.6%, in the 2-segment group for the test set, respectively. From the generalizability point of view, regarding performance drops between unseen data and unseen participants, presented by validation and test set, respectively. SVM, CB, and RF follow the chronological order regarding the lowest performance drop in precision, recall, accuracy, and F1 score by 13.3%, 15.8%, 13.9%, 13.7%, and 19.4%, 21.1%, 20.7%, 21.0%, and lastly 25.4%, 25.3%, 25.3%, 25.4%, respectively. That indicates that the models trained on the group-wise expanded dataset generate a higher performance drop than those results presented in Table 1.

CB remains the strongest model, achieving perfect scores of 100% in multiple validation set groups. RF follows with a peak accuracy of 99.6%, while SVM lags behind, with its best accuracy at 83.8%. In the test set, CB maintains the highest accuracy of 79.3%, followed by RF (74.3%) and SVM (69.6%). Performance drops are more pronounced in group-wise datasets, with SVM showing the least decline, further supporting its generalizability.

The performance results underscored the robustness of CB and superior performance across all dataset configurations, with RF following closely behind. Although SVM showed notable generalizability, it generally trailed behind the other models. In addition to the performance metrics, Table 3 presents the confusion matrix results, which offer a clinical perspective of the ML model's performance. Here, the recall metric, also known

Segment-wise											
Validation set											
		Full		**2 Seg**		**3 Seg**		**4 Seg**		**5 Seg**	
		Predicted (+)	**Predicted (−)**	**Predicted (+)**	**Predicted (−)**	**Predicted (+)**	**Predicted (−)**	**Predicted (+)**	**Predicted (−)**	**Predicted (+)**	**Predicted (−)**
CB	Actual(+)	74	1	74	1	74	1	**75**	**0**	73	2
	Actual(−)	6	102	4	104	3	105	**4**	**104**	4	104
RF	Actual(+)	68	7	71	4	70	5	**72**	**3**	71	4
	Actual(−)	10	98	10	98	10	98	**9**	**99**	10	98
SVM	Actual(+)	54	21	56	19	**59**	**16**	56	19	58	17
	Actual(−)	20	80	17	91	**21**	**87**	19	89	20	80
Test set											
CB	Actual(+)	60	8	58	10	53	15	**60**	**8**	61	7
	Actual(−)	25	50	15	60	10	65	**14**	**61**	20	55
RF	Actual(+)	**61**	**7**	55	13	57	11	55	13	56	12
	Actual(−)	**32**	**43**	26	49	26	49	19	56	24	51
SVM	Actual(+)	52	16	**57**	**11**	50	18	54	14	56	12
	Actual(−)	29	46	**33**	**42**	24	51	25	50	27	48
Group-wise											
Validation set											
		2 Seg. group		**3 Seg. group**		**4 Seg. group**		**5 Seg. group**		**All Seg. group**	
CB	Actual(+)	142	0	213	0	**309**	**0**	387	0	**1124**	**0**
	Actual(−)	2	222	1	335	**0**	**357**	1	527	**0**	**1626**
RF	Actual(+)	136	6	211	2	298	11	376	11	**1119**	**5**
	Actual(−)	12	212	16	320	13	410	9	519	**5**	**1616**
SVM	Actual(+)	107	35	**177**	**36**	229	80	305	82	910	214
	Actual(−)	32	192	**53**	**283**	57	366	81	447	230	1391
Test set											
CB	Actual(+)	**112**	**24**	165	93	208	64	240	100	663	357
	Actual(−)	**48**	**102**	56	169	111	189	48	327	104	1021
RF	Actual(+)	**107**	**29**	136	68	195	77	252	88	702	318
	Actual(−)	**52**	**98**	52	173	71	229	96	279	326	799
SVM	Actual(+)	103	33	**160**	**44**	208	64	243	97	787	233
	Actual(−)	54	96	**96**	**129**	11	189	158	217	511	614

Table 3. Confusion matrix results for validation and test sets gained from the ML models showed the highest average accuracy in each dataset configuration, where the best result is bolded for each ML classifier. The highest performance is shown in bold.

as the true positive rate (TPR), is highlighted as it indicates the model's ability to correctly identify positive cases, a clinically relevant factor in clinical applications. The confusion matrix results show that while the best performance order is preserved as CB, RF, and SVM, the highest TPR is aligned with the highest accuracies in the performance metrics results in Tables 1 and 2. The best TPR of 75/75 and 72/75 for CB and RF, respectively, were achieved in the 4 Seg. category, and for SVM, the highest TPR of 59/75 has been achieved in the 3 Seg. category for the validation set. The confusion matrix test set results also show a deviation compared to performance metrics results in Tables 1 and 2, where the best TPR occurs in 4 Seg, Full, and 2 Seg. categories are 60/68, 61/68, and 67/68 for CB, RF, and SVM, respectively. Group-wise confusion matrix results for the validation set are aligned for CB and RF with the performance metrics results in Tables 1 and 2, which are TPR of 1124/1124 and 1119/1124, respectively. However, SVM diverges and reaches its highest TPR of 177/213 in 3 Seg. group. For the test set results, CB and RF reached their highest TPR of 112/136 and 107/136 in the 2 Seg. Groups and SVM have a TPR of 160/204 in 3 Seg. Group. Almost all high TPRs are achieved using segment-wise datasets or group-wise datasets with only one exception, where the RF classifier performed the best TPR in the full segment category.

CB and RF achieve the highest recall (TPR) in the 4-segment validation category, while SVM's best TPR is in the 3-segment group. In the test set, CB and RF peak in the 4-segment and full-segment groups, whereas SVM's best TPR is in the 2-segment group. The results confirm CB's superior classification performance, RF's strong but slightly overfitting nature, and SVM's generalizability despite lower overall accuracy.

Examining the performance improvement between the models trained on the full voice sequence and the best-performing segmented version, using the ROC curves presented in Fig. 3, reveal notable differences across classifiers. The CB classifier, which initially achieved an AUC of 0.81 when trained on the full sequence, improved to 0.90 when using the segmented data, demonstrating the most substantial performance gain. Similarly, the RF classifier showed a slight improvement from 0.80 to 0.82, while the SVM model increased from 0.73 to 0.78.

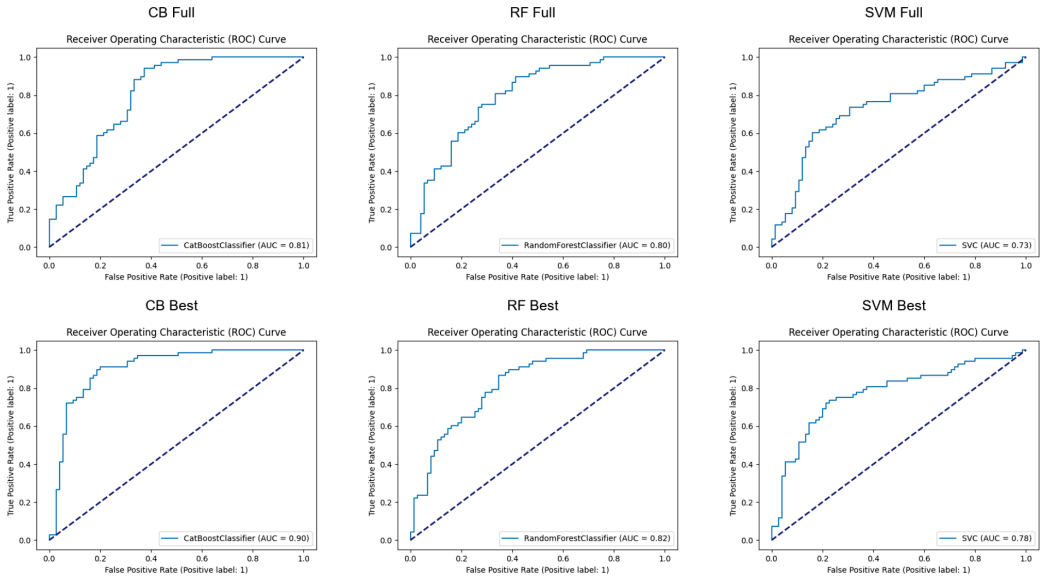


Fig. 3. ROC curves for ML models trained on the full voice sequence and segmented voice data, where the best performance was achieved.

Similar to the previous metrics presented in Tables 1,2, and 3, the ROC curves also indicate that CB outperforms the other models across all decision thresholds. The steeper initial rise of the ROC curve for CB Best suggests a stronger separation of classes compared to the full-sequence approach.

Discussion

This study employed a segmentation method to observe how different segments and groups of segments of the utterance of the vowel "a" affect the binary classification performance of three classifiers, RF, SVM, and CB, for the classification of COPD and non-COPD voices. This study uniquely contributes to the field by specifically examining the impact of segmentation on ML performance, a topic not fully explored in prior research. It provides new insights into how different uses of time frames affect model outcomes, offering a deeper understanding of their potential to enhance performance consequently generating academic engagement with this research direction. The key findings in this study are as follows:

- Using segments individually or in groups mostly generated higher performance metric scores as compared to using the features extracted from the full sequence.
- Segmenting data increases classification accuracy. However, this increase comes with a cost of higher variation in classification performance.
- Expanding the dataset by merging the segments in groups reduces the variance observed in segment-wise classification performance but never reaches the highest performance achieved in the segments.
- The features extracted from the first half of the recordings show a higher classification performance on average than the second half of the recordings for tree-based algorithms CB and RF. However, the distance-based SVM shows fluctuation.

In general, segmentation proves to be an effective method for creating datasets composed of BLA and MFCC features extracted from time frames of the vowel "a" recordings. This might be connected to the voice signals depicting stationary characteristics in shorter time frames[45]. However, this approach appears to benefit tree-based algorithms, such as RF and CB, more than distance-based algorithms, like SVM. The SVM classifier did not show as much improvement as RF and CB. The fluctuations in SVM performance across the datasets, compared to CB and RF, may be due to its sensitivity to feature distribution. Unlike CB and RF, which adaptively handle non-linear patterns, SVM's hyperparameter sensitivity may cause inconsistencies across the datasets, even with grid search optimization. The ability of CB and RF to weigh features differently may contribute to their more consistent performance. Considering the increased variation, a key challenge with this method is identifying the correct segment within the recording. In this study, the second segment out of a four-segment category achieved the highest performance, with validation and test set accuracies of 97.8% and 84.7%, respectively, using the CB classifier. This result is further supported from a clinical standpoint by the confusion matrix, which shows a TPR of 100% (75/75) on the validation set and 88.2% (60/68) on the test set. When

combined with other performance metrics for the same segment, the high recall suggests that this approach may help ensure that most moderate-level COPD cases are identified[46,47]. This is particularly relevant given that the COPD cohort in this study primarily consists of participants in moderate stages, with an average ratio of Forced Expiratory Volume in 1 s (FEV1) and Forced Vital Capacity (FVC) around 0.61%. From a clinical perspective, another critical consideration is maintaining a balance between TPR and False Positive Rate (FPR) to avoid unnecessary overdiagnosis[6,47]. In this regard, CB seems to achieve a reasonable balance in both the validation and test sets within the four-segment category, providing reliable performance without compromising clinical relevance. Group-wise results in Table 2 and Fig. 2 show a smoothing effect on the results with decreased variation. However, the greater performance drop between the validation and test sets indicates a higher degree of possible overfitting to the validation data[48,49]. This problem might be mitigated by increasing participant variability, including factors like balanced age and health status, increased sample size, using regularization techniques, and data augmentation in future studies. Additionally, from a clinical point of view, expanding a dataset based on grouping the small time frames does not seem to support real-world scenarios, as represented by the test set results in Table 3. The best TPR results in the group-wise analysis do not reflect better performance than segment-wise or full-sequence results.

The results suggest that certain time frames within the recordings exhibit more deterministic properties, which could enhance classification performance. This situation may be due to the more pronounced differences in certain voice feature characteristics, as reported in previous studies[50,51]. Specifically, the second segment in the four-segment group appears to have performed the best, which could be attributed to the unique temporal or acoustic properties that distinguish it from other segments. It would be valuable to investigate why the second segment yielded superior results, whether it's due to a specific change in vocalization patterns or a shift in the underlying physiological state of the subjects. This also reflects the dynamic characteristics of the vowel production[43,44]. However, the findings indicate that different ML models, such as CB, RF, and SVM, excel in different segments of the recordings. This implies that the signal processing steps should be optimized to match the characteristics of the chosen model. Consequently, employing a fixed recording duration, as seen in other data collection methods[52], or data extracted from a fixed frame of a signal[53] may not fully leverage the performance potential of models for COPD classification. By integrating segmentation strategies with ML models, future diagnostic tools could achieve higher accuracy and robustness, enabling earlier and more precise identification of COPD. These findings could inform the development of clinical workflows that leverage vocal biomarkers for screening, monitoring, and early intervention.

The computational demands of this study were substantial due to 21 dataset configurations, 3 ML models, 3 hyperparameter options, and a 5×5 nCV framework, resulting in $625+$ training runs per model and a total training period exceeding 3 months. While the primary objective was to assess the impact of segmentation on model performance, an important observation was that grouping segments into new datasets to expand them significantly increased training time and memory usage across all models, in proportion to the number of segments grouped. However, segmentation itself did not increase memory usage or the time required to train the models compared to using the full sequence of the recording and remained 220 MB–33.8 s, 204 Mb–40.3 s, and194 MB–384.0 s for the best performing CB, RF, and SVM, respectively, as the total number of data points remained unchanged. This suggests that while segmentation enhances model performance, dataset expansion strategies may introduce computational trade-offs that should be considered in practical ML applications for voice analysis in COPD classification.

With respect to all the strengths in this study, such as employing well-known ML methods and voice features, applying regularization and nested cross-validation to minimize overfitting and increase generalizability, it is essential to acknowledge the limitations of the present study. This study is based on a dataset collected from a small cohort of individuals who primarily speak the dialect of southern Sweden. This limited sample size and linguistic homogeneity may restrict the generalizability of the findings. Applying the findings to other populations or dialects may present challenges due to differences in speech patterns, vocal characteristics, and demographic factors. Variations in accent, pronunciation, and language use across regions or dialects could impact the model's ability to generalize. Additionally, limited representation of certain age and gender groups could introduce biases and affect the generalizability of the findings. While the study provides valuable insights, future work should focus on increasing the dataset size and ensuring a more balanced demographic distribution to improve robustness and applicability across diverse populations. However, this limitation may provide a more concentrated analysis that could be more challenging to achieve with a widespread dataset. Another limitation is that the analysis is constrained to CB, RF, and SVM classifiers in the context of COPD, may not be applicable to other models, such as artificial neural networks, or to other voice-affecting disorders, such as Parkinson's disease[54]. Additionally, the study focuses solely on sustained vowel "a", which may not fully capture articulatory and phonatory variations present in other vowels, consonant–vowel pairs, or connected speech. Expanding the analysis to include these elements, along with deep learning approaches, could offer a more comprehensive understanding of respiratory-phonatory coordination in real-world conditions. Future studies should analyze statistical differences more thoroughly to better understand the statistical significance of feature variations. Furthermore, the cross-sectional design of the study does not provide longitudinal analysis for understanding changes and trends over time, which might be another area worth investigating alongside the investigation of the computational efficiency of different segmentation strategies aspect in more detail, particularly in the context of real-time applications or resource-constrained environments in future research.

The variability in segment-wise results suggests that alternative techniques, such as wavelet transformation[55], may capture the temporal characteristics of voice without requiring segmentation. This approach could eliminate the need to identify the most suitable segment for classification. Future research should explore this potential and investigate further optimization of voice assessment systems for COPD detection. Additionally, addressing this

study's limitations by testing larger datasets, exploring multilingual applications, and incorporating longitudinal data will help to enhance the reliability of voice-based diagnostic tools.

Conclusion

These findings suggest that time-sensitive properties in vowel production are important for COPD classification, and that segmentation can help capture these properties. However, expanding the dataset by grouping segments does not necessarily improve performance, especially in real-world scenarios. Potential future applications of a voice assessment system for COPD include aiding early diagnosis through vocal feature analysis, supporting disease monitoring, and facilitating personalized management strategies. However, the clinical utility of such a system would depend on further validation in larger, real-world datasets. Additionally, if demonstrated to be effective, such a system could serve as a decision-support tool for clinicians, potentially improving diagnostic accuracy and optimizing resource allocation in healthcare settings.

Methods

General description

This study conducts experiments on the utterance of entire vowel "a" recordings by dividing the entire recording into several subsegments and comparing the binary classification performance of three ML models: CB, RF, and SVM segment-wise and group-wise. Segment-wise datasets involve training models on smaller time frames of the full recording, while group-wise datasets involve merging multiple segments into a single dataset, increasing the number of samples available for training. The models were trained using nCV on different combinations in the number of inner and outer folds. Figure 4 illustrates the workflow and segmentation proceeds of this study.

Data acquisition

The dataset employed in this study was created from Swedish utterances of the vowel "a" recordings collected through a mobile application *VoiceDiagnostic*, from a pool of research clinic participants who were recruited at the Blekinge Institute of Technology (BTH) in Sweden. Sixty-eight participants provided 1246 recordings in total. 30 COPD (16 female and 14 male) and 38 Healthy Control (HC) group (20 female and 18 male) made 436 and 810, respectively. Participants with COPD had an average FEV1/FVC of 0.61%, with a standard deviation of 0.12%. The voices recorded by VoiceDiagnostic, which is an application compatible with Android and iPhone, allows participants to make two types of recordings: one single utterance of the vowel "a" with the maximum possible duration and a scripted speech provided in the application. However, this study analyses only utterances of the vowel "a" recordings because the sustained vowel "a" is widely recognized as providing a controlled, reliable measure of vocal fold function and acoustic stability[56–58]. Participants were enrolled after a brief introduction to the study was given by a nurse with experience in research and after meeting the first author, who provided deeper information about the study. Participants were instructed to record in a quiet environment, free from background noise. Each recording was manually checked by the author to confirm it was free from any unwanted sounds. The participant's integrity and safety against the risk of data leakage were ensured by anonymizing the data and securing both physical and digital data in secure cabinets and safe databases. Each participant was assigned a unique ID to ensure that no personal information was used, and all data were anonymized to protect privacy, especially given the sensitivity of voice data. The study was approved by the Swedish ethics review authority in Umeå (DNR: 2020-01045) and followed the principles of the Declaration of Helsinki. All participants signed a written informed consent form that allows the collection of voice samples, health data, and sociodemographic information during a six-month period. The recruitment was based on the inclusion and exclusion criteria given below:

Inclusion criteria

- COPD group
  Participants starting from 18 years old or older who have a COPD diagnosis, with access and proficiency to use a smartphone.
- HC group
  Participants starting from 18 years old and older who do not declaring that they have a voice-affecting disorder diagnosis, i.e., no disorder listed in the categories' nonlaryngeal aerodigestive disorders affecting voice', 'neurological disorders affecting voice', and 'systematic conditions affecting voice' in the Classification Manual for Voice Disorders[54], with access and proficiency to use a smartphone.

Exclusion criteria

- COPD group
  Participants younger than 18 years old with a voice-affecting disorder other than COPD, or declare no access or proficiency to smartphone use.
- HC group
  Participants younger than 18 years old with a voice-affecting disorder or declaring no access or proficiency to use a smartphone.

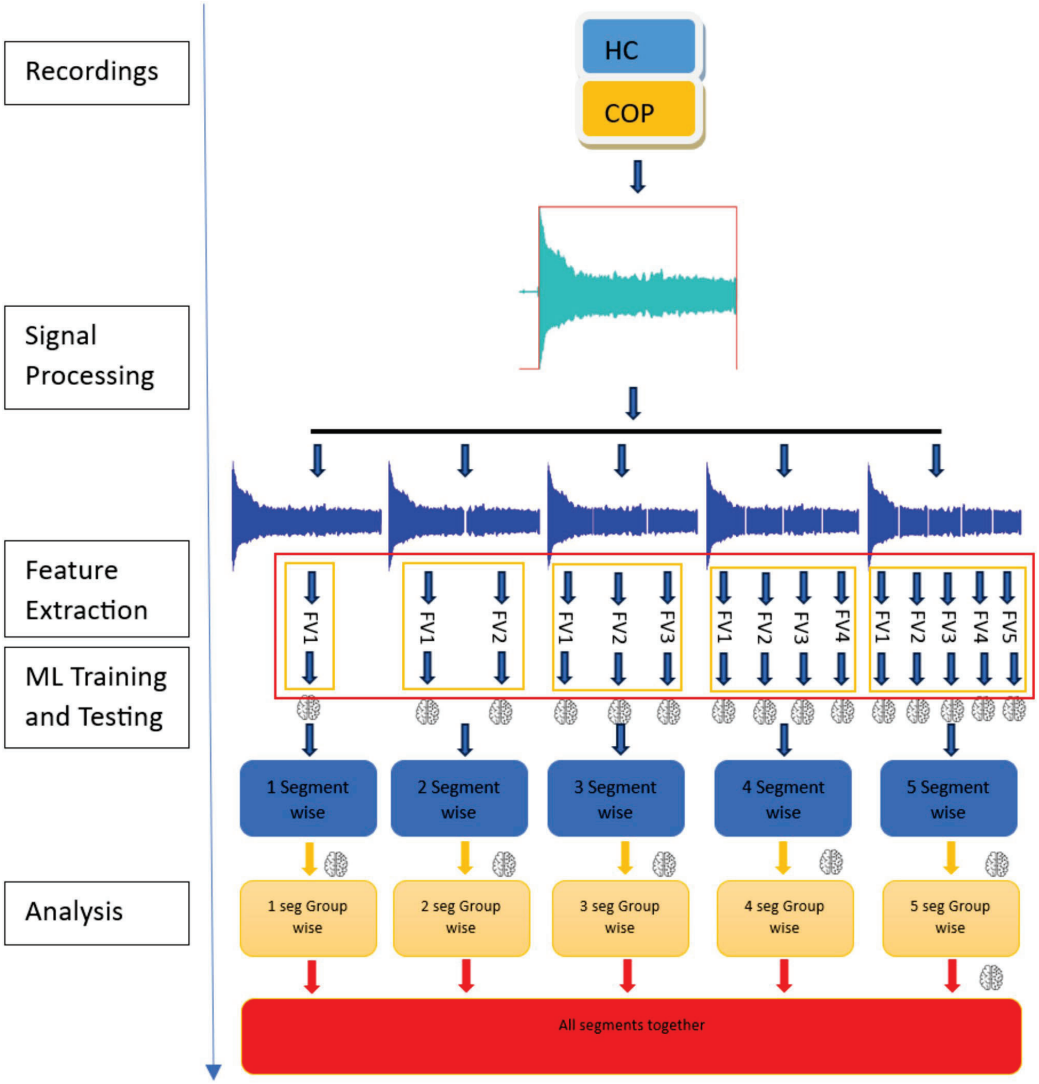


Fig. 4. Chronological overview of the workflow. FV stands for feature vectors extracted from each sample.

Data preparation and feature extraction

The voice recordings were checked and standardized on parameters regarding sampling frequency and silence-free sequences. 44.1kHZ frequency was common for all voice recordings, and the silence part, which is usually common at the beginning and end of the vowel recordings, was dispatched from the recordings using a basic moving average filter and adaptive threshold that standardized the process for each recording[50]. The silence-free voice signals were divided equally into several one-dimensional segments from 2 to 5, separately based on the total duration of the voice part of the recording as visualized in Fig. 4. These choices were made to comply with the constrains of potential mobile implementations such as limited processing power, memory, and battery life. By including the original full recording, 15 different feature vectors were calculated for each recording for five distinct experiments, as it is shown in Fig. 1.

The feature vectors contained 107 parameters related to some demographics (Age, and Gender), baseline acoustics (BLA) (Duration, 2 Fundamental Frequency measures, Harmonic to noise ratio, 5 Jitter measures, 5 Shimmer measures, 4 mean formant measures, and 4 median formant measures) and Mel-Frequency Cepstral

Coefficients (MFCC) (Standard deviation and mean of first 13 MFCCs and first and second derivative of them), where a detailed description of the features extracted from the vowel "a" are described in a previous study performed on the same dataset[50]. These features were chosen for their promising performance in earlier studies on different types of COPD voices[50,51]. Jitter and Shimmer measure frequency and amplitude perturbations, respectively, reflecting vocal fold stability. Increased values are linked to respiratory and laryngeal impairments, including COPD. MFCCs capture spectral properties of speech and are widely used in pathological voice classification due to their effectiveness in modeling vocal tract characteristics[56,59,60]. The features were extracted using Praat (Parselmouth) and Librosa libraries using Python.

Experimentation with machine learning

In order to mitigate the imbalance problems that affect the performance, the data set was balanced by matching the feature vectors based on gender and age. That resulted in a dataset containing 1058 recordings belonging to 48 participants (24 females and 24 males), which was used in ML experiments. A subset of 25% (12/48 participants and 143/1058 recordings) of participants (12 participants, 6 females, and 6 males, 143 recordings) based on matched age with ± 6.7 years old standard deviation with an average age of 73.2 years old were isolated for the test dataset for the evaluation of ML models on unseen data for the training set. Further, the remaining 75% (36/48 participants and 917/1058 recordings) of participants (36 participants, 18 females, and 18 males) with ± 6.5 years old standard deviation with an average age of 75, 1 years old were further divided into two sub-datasets, 80% (732/917 recordings) training dataset and 20% (183/917 recordings) validation dataset corresponding to 732 and 183 recordings, respectively. This data distribution was done to observe ML models' learning and classification performance on data collected from the same participant from different time stamps.

Nested cross-validation nCV, also known as double k-fold cross-validation, is a method utilized to mitigate the overfitting[61,62]. This technique combines traditional k-fold cross-validation in two stages: an outer loop and an inner loop. In this method, the outer loop divides the data into k folds, and for each fold, the inner loop performs cross-validation on the training data to tune the hyperparameters. This process helps ensure that the model is not overfitting to the training data and provides a more reliable estimate of model performance. It was employed for the training of CB, RF, and SVM classifiers[63–65], which were suggested for their ability to handle tabular data with relatively small sample sizes while effectively capturing non-linear patterns. These classifiers have demonstrated strong performance in similar datasets across multiple studies[19,66,67], including a previous investigation focused on COPD classification[50]. Additionally, the limited sample size in this study constrained the feasibility of testing larger models, as proven by an initial LSTM test. The LSTM model exhibited high overfitting due to insufficient data, ultimately leading to the decision to abandon the pursuit of larger models. Hyperparameter optimization was performed using a grid search within the inner loop of nested cross-validation to choose the best performing model. The optimized hyperparameters that provided the best performance with specific nCV combinations on the models are as follows:

- CB: 5X2nCV, ('depth': 4, 'iterations': 300, 'l2_leaf_reg': 5, 'learning rate': 0.1).
- RF: 4X4nCV, ('max depth: None, 'min samples split': 5, 'n estimators': 200).
- SVM: 5X3nCV, ('C': 1, 'degree': 2, 'kernel': 'linear').

Analysis

Alongside the performance measures of accuracy, F1-score, precision, recall metrics and ROC curves, the confusion matrix was used to elucidate the results from a clinical perspective. The observed results were presented in the form of graphs and tables.

Data availability

The raw recordings cannot be made available due to ethical and general data protection regulations. However, an anonymized version of the dataset after the pre-processing of voice, generated during the present study will be made available from the corresponding author's institution upon reasonable request. The code for repeating the experiments can be found on GitHub: https://github.com/AIITPlanet/Code/blob/main/Analysis_RF_SVM_CB_Nested_ForPartitionsToExcel.py.

Received: 25 November 2024; Accepted: 20 March 2025
Published online: 22 March 2025

References

1. Jarhyan, P., Hutchinson, A., Khaw, D., Prabhakaran, D. & Mohan, S. Prevalence of chronic obstructive pulmonary disease and chronic bronchitis in eight countries: A systematic review and meta-analysis. *Bull World Health Organ.* **100**, 216–230 (2022).
2. Chronic obstructive pulmonary disease (COPD). https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd).
3. Agarwal, A. K., Raja, A. & Brown, B. D. Chronic Obstructive Pulmonary Disease. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2024).
4. Diab, N. et al. Underdiagnosis and overdiagnosis of chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* **198**, 1130–1139 (2018).
5. Agustí, A. G. N. et al. Systemic effects of chronic obstructive pulmonary disease. *Eur Respir J* **21**, 347–360 (2003).
6. Singh, D. et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease: The GOLD science committee report 2019. *Eur. Respir. J.* **53**, 1900164 (2019).
7. Peng, J. et al. A machine-learning approach to forecast aggravation risk in patients with acute exacerbation of chronic obstructive pulmonary disease with clinical indicators. *Sci. Rep.* **10**, 3118 (2020).
8. Badnjevic, A., Gurbeta, L. & Custovic, E. An expert diagnostic system to automatically identify asthma and chronic obstructive pulmonary disease in clinical settings. *Sci. Rep.* **8**, 11645 (2018).

9. Shastry, A., Balasubramanium, R. K. & Acharya, P. R. Voice analysis in individuals with chronic obstructive pulmonary disease. *Int. J. Phonosurg. Laryngol.* **4**, 45–49 (2014).
10. Mohamed, E. E. & El Maghraby, R. A. Voice changes in patients with chronic obstructive pulmonary disease. *Egypt. J. Chest Dis. Tuberculosis* **63**, 561–567 (2014).
11. Naqvi, S. Z. H. & Choudhry, M. A. An automated system for classification of chronic obstructive pulmonary disease and pneumonia patients using lung sound analysis. *Sensors* **20**, 6512 (2020).
12. Haider, N. S., Singh, B. K., Periyasamy, R. & Behera, A. K. Respiratory sound based classification of chronic obstructive pulmonary disease: A risk stratification approach in machine learning paradigm. *J. Med. Syst.* **43**, 255 (2019).
13. Korkmaz, Y. SS-ESC: A spectral subtraction denoising based deep network model on environmental sound classification. *SIViP* **19**, 50 (2024).
14. Korkmaz, Y. & Boyacı, A. Classification of Turkish Vowels Based on Formant Frequencies. In *2018 International conference on artificial intelligence and data processing (IDAP)* 1–4 (2018). https://doi.org/10.1109/IDAP.2018.8620877.
15. Korkmaz, Y. & Boyacı, A. milVAD: A bag-level MNIST modelling of voice activity detection using deep multiple instance learning. *Biomed. Signal Process. Control* **74**, 103520 (2022).
16. Tirumala, S. S., Shahamiri, S. R., Garhwal, A. S. & Wang, R. Speaker identification features extraction methods: A systematic review. *Expert Syst. Appl.* **90**, 250–271 (2017).
17. Fagherazzi, G., Fischer, A., Ismael, M. & Despotovic, V. Voice for health: The use of vocal biomarkers from research to clinical practice. *Digit. Biomark* **5**, 78–88 (2021).
18. Little, M., McSharry, P., Roberts, S., Costello, D. & Moroz, I. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *Nat. Prec.* 1–1 (2007) https://doi.org/10.1038/npre.2007.326.1.
19. Idrisoglu, A., Dallora, A. L., Anderberg, P. & Sanmartin Berglund, J. Applied machine learning techniques to diagnose voice-affecting conditions and disorders: A systematic literature review. *JMIR Med. Inform.* **18**, 46105 (2023).
20. Verma, V. et al. A novel hybrid model integrating MFCC and acoustic parameters for voice disorder detection. *Sci. Rep.* **13**, 22719 (2023).
21. Peng, X., Xu, H., Liu, J., Wang, J. & He, C. Voice disorder classification using convolutional neural network based on deep transfer learning. *Sci. Rep.* **13**, 7264 (2023).
22. Quan, C., Ren, K. & Luo, Z. A deep learning based method for parkinson's disease detection using dynamic features of speech. *IEEE Access* **9**, 10239–10252 (2021).
23. Fujimura, S. et al. Classification of voice disorders using a one-dimensional convolutional neural network. *J. Voice* **36**, 15–20 (2022).
24. Luz, S. Longitudinal monitoring and detection of alzheimer's type dementia from spontaneous speech data. In vols 2017-June 45–46 (2017).
25. Mahesh, B. Machine learning algorithms - A review. *Int. J. Sci. Res.* **9**, 381–386 (2018).
26. Flach, P. *Machine learning: The art and science of algorithms that make sense of data* (Cambridge University Press, 2012).
27. Yoo, C., Ramirez, L. & Liuzzi, J. Big data analysis using modern statistical and machine learning methods in medicine. *Int. Neurourol. J.* **18**, 50–57 (2014).
28. Xu, C. & Jackson, S. A. Machine learning and complex biological data. *Genome Biol.* **20**, 76 (2019).
29. König, I. R. et al. Machine learning and data mining in complex genomic data—A review on the lessons learned in Genetic Analysis Workshop 19. *BMC Genet.* **17**, S1 (2016).
30. Bayram, B., Kunduracioglu, I., Ince, S. & Pacal, I. A systematic review of deep learning in MRI-based cerebral vascular occlusion-based brain diseases. *Neuroscience* **568**, 76–94 (2025).
31. Elbedwehy, S., Hassan, E., Saber, A. & Elmonier, R. Integrating neural networks with advanced optimization techniques for accurate kidney disease diagnosis. *Sci. Rep.* **14**, 21740 (2024).
32. Shaikhina, T. et al. Machine learning for predictive modelling based on small data in biomedical engineering. *IFAC-PapersOnLine* **48**, 469–474 (2015).
33. Braga, D., Madureira, A. M., Coelho, L. & Ajith, R. Automatic detection of Parkinson's disease based on acoustic analysis of speech. *Eng. Appl. Artif. Intell.* **77**, 148–158 (2019).
34. Ali, L., Zhu, C., Zhang, Z. & Liu, Y. Automated detection of parkinson's disease based on multiple types of sustained phonations using linear discriminant analysis and genetically optimized neural network. *IEEE J. Transl. Eng. Health Med.* **7**, 1–10 (2019).
35. Pramanik, M., Pradhan, R., Nandy, P., Qaisar, S. M. & Bhoi, A. K. Assessment of acoustic features and machine learning for parkinson's detection. *J. Healthcare Eng.* **2021**, 1–13 (2021).
36. Tsanas, A., Little, M. A., McSharry, P. E. & Ramig, L. O. Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests. *IEEE Trans. Biomed. Eng.* **57**, 884–893 (2010).
37. Zhang, L. et al. An intelligent mobile-enabled system for diagnosing parkinson disease: Development and validation of a speech impairment detection system. *JMIR Med. Inform.* **8**, e18689 (2020).
38. Chun, K. S. et al. Towards passive assessment of pulmonary function from natural speech recorded using a mobile phone. in *2020 IEEE International conference on pervasive computing and communications (PerCom)* 1–10 (2020). https://doi.org/10.1109/PerCom45495.2020.9127380.
39. Farrús, M. et al. Speech-based support system to supervise chronic obstructive pulmonary disease patient status. *Appl. Sci.* **11**, 7999 (2021).
40. Soumaya, Z., Taoufiq, B. D., Benayad, N., Achraf, B. & Ammoumou, A. A Hybrid method for the diagnosis and classifying parkinson's patients based on time–frequency domain properties and K-nearest neighbor. *J. Med. Signals Sens.* **10**, 60–66 (2020).
41. Nathan, V. et al. Extraction of voice parameters from continuous running speech for pulmonary disease monitoring. In *2019 IEEE international conference on bioinformatics and biomedicine (BIBM)* 859–864 (2019). https://doi.org/10.1109/BIBM47256.2019.8983115.
42. Nathan, V., Vatanparvar, K., Rahman, M. M., Nemati, E. & Kuang, J. Assessment of chronic pulmonary disease patients using biomarkers from natural speech recorded by mobile devices. In *2019 IEEE 16th international conference on wearable and implantable body sensor networks (BSN)* 1–4 (2019). https://doi.org/10.1109/BSN.2019.8771043.
43. Kent, R. D. & Rountrey, C. What acoustic studies tell us about vowels in developing and disordered speech. *Am. J. Speech Lang Pathol.* **29**, 1749–1778 (2020).
44. Fujisaki, H. Dynamic characteristics of voice fundamental frequency in speech and singing. In *The Production of speech* (ed. MacNeilage, P. F.) 39–55 (Springer, 1983).
45. Vieira, V. J. D., Costa, S. C. & Correia, S. E. N. Non-stationarity-based adaptive segmentation applied to voice disorder discrimination. *IEEE Access* **11**, 54750–54759 (2023).
46. Curtis, J. R. & Patrick, D. L. The assessment of health status among patients with COPD. *Eur. Respir. J.* **21**, 36s–45s (2003).
47. Shen, X. & Liu, H. Using machine learning for early detection of chronic obstructive pulmonary disease: A narrative review. *Respir. Res.* **25**, 336 (2024).
48. Peng, Y. & Nagata, M. H. An empirical overview of nonlinearity and overfitting in machine learning using COVID-19 data. *Chaos, Solit. Fract.* **139**, 110055 (2020).
49. Kolluri, J., Kotte, V. K., Phridviraj, M. S. B. & Razia, S. Reducing overfitting problem in machine learning using novel L1/4 regularization method. In *2020 4th international conference on trends in electronics and informatics (ICOEI)(48184)* 934–938 (2020). https://doi.org/10.1109/ICOEI48184.2020.9142992.

50. Idrisoglu, A. et al. COPDVD: Automated classification of chronic obstructive pulmonary disease on a new collected and evaluated voice dataset. *Artif. Intell. Med.* **156**, 102953 (2024).
51. Pramono, R. X. A., Imtiaz, S. A. & Rodriguez-Villegas, E. Evaluation of features for classification of wheezes and normal respiratory sounds. *PLoS ONE* **14**, e0213659 (2019).
52. Bot, B. M. et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. *Sci. Data* **3**, 160011 (2016).
53. Tena, A., Clarià, F., Solsona, F. & Povedano, M. Voiceprint and machine learning models for early detection of bulbar dysfunction in ALS. *Comput. Methods Programs Biomed.* **229**, 107309 (2023).
54. *Classification Manual for Voice Disorders–I.* (Psychology Press, New York, 2005). https://doi.org/10.4324/9781410617293.
55. Moufidi, A., Rousseau, D. & Rasti, P. wavelet scattering transform depth benefit, an application for speaker identification. In *Artificial Neural Networks in Pattern Recognition* (eds El Gayar, N. et al.) 97–106 (Springer International Publishing, 2023).
56. Godino-Llorente, J. I., Gomez-Vilda, P. & Blanco-Velasco, M. Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short-term cepstral parameters. *IEEE Trans. Biomed. Eng.* **53**, 1943–1953 (2006).
57. Baken, R. J. & Orlikoff, R. F. *Clinical measurement of speech and voice* (Singular Thomson Learning, 2000).
58. Titze, I. R. *Principles of voice production* (Prentice Hall, 1994).
59. Ma, E.P.-M. & Yiu, E.M.-L. Multiparametric evaluation of dysphonic severity. *J. Voice* **20**, 380–390 (2006).
60. Mekyska, J. et al. Robust and complex approach of pathological speech signal analysis. *Neurocomputing* **167**, 94–111 (2015).
61. Parvandeh, S., Yeh, H.-W., Paulus, M. P. & McKinney, B. A. Consensus features nested cross-validation. *Bioinformatics* **36**, 3093–3098 (2020).
62. Cawley, G. C. & Talbot, N. L. C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **11**, 2079–2107 (2010).
63. Schölkopf, B. *Support vector learning* (Oldenbourg München, 1997).
64. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
65. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. In *Advances in neural information processing systems* vol. 31 (Curran Associates, Inc., 2018).
66. Hancock, J. T. & Khoshgoftaar, T. M. CatBoost for big data: an interdisciplinary review. *J. Big Data* **7**, 94 (2020).
67. Huang, G. et al. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. *J. Hydrol.* **574**, 1029–1041 (2019).

Author contributions

AI conducted the experiment and conceptualized and wrote the original draft of the manuscript. ALD, AC, PA, AJ, and JSB analyzed the results, revised the manuscript, and supervised and provided resources. All authors reviewed the manuscript.

Funding

Open access funding provided by Blekinge Institute of Technology.
Excellence Center at Linköping –Lund in Information Technology (ELLIIT).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.I.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Study IV

Feature Analysis of the Vowel [a :] In individuals with Chronic Obstructive Pulmonary Disease

Submitted as:

Idrisoglu A, Dallora AL, Cheddad A, Anderberg P, Jakobsson A, Sanmartin Berglund J. Feature Analysis of the Vowel [a:] in Individuals with Chronic Obstructive Pulmonary Disease and Healthy Controls. (Submitted to Journal).

Original Paper

Feature Analysis of the Vowel [a:] in Individuals with Chronic Obstructive Pulmonary Disease and Healthy Controls

Alper Idrisoglu[1], Ana Luiza Dallora Moraes[1], Abbas Cheddad[1,2], Peter Anderberg[1], Susanna Whitling[4], Andreas Jakobsson[3], Johan Sanmartin Berglund[1]

[1]Blekinge Institute of Technology, Department of Health, Karlskrona, 371 41, Sweden.
[2]University of Tartu, Institute of Computer Science, Narva mnt 18, 51009 Tartu, Estonia.
[3]Lund University, Mathematical Statistic, SE 221 00, Lund, Sweden.
[4]Lund University, Department of Logopedics, Phoniatrics, and Audiology, SE 221 00, Lund, Sweden.

Corresponding Author:
Alper Idrisoglu
Blekinge Institute of Technology
Valhallavägen 1,
Karlskrona, 371 41
Phone: 46 701462619
Email: alper.idrisoglu@bth.se

Abstract

Background: In addition to impairing the lung function, Chronic Obstructive Pulmonary Disease (COPD) also affects phonatory characteristics. Recent research highlights the potential of voice as a digital biomarker to support clinical decision-making. While machine learning (ML) can detect disease patterns from acoustic features, clinical relevance requires understanding the relationship between the disorder and acoustic features.

Objective: This study investigates both statistical and clinical significance using Baseline Acoustic (BLA) and Mel-Frequency Cepstral Coefficient (MFCC) features with focusing on individuals with COPD and healthy controls (HC).

Method: Acoustic features derived from Swedish utterances of the vowel [a:], recorded via mobile phones from 48 age-matched participants (24 COPD, 24 HC; equal gender distribution), was analyzed. To reduce bias from varying recording counts, features were aggregated by averaging 10 randomly selected recordings per participant over 100 iterations. Vowel articulation was visualized in the vowel quadrilateral space using F1 (tongue height) and F2 (tongue advancement). Group differences were assessed using the Shapiro-Wilk test, Mann-Whitney U test ($\alpha = 0.05$), Benjamini-Hochberg (BH) and Bonferroni corrections, multivariate Permanova Test, and Cliff's Delta (δ).

Results: Of 101 features, 29 remained significant after BH correction and one after Bonferroni. Multivariate testing ($p = 0.019$) showed group separation. Additionally, 34 features demonstrated large effect sizes, suggesting potential as digital biomarkers.

Conclusion: Voice data recorded via mobile phones captures meaningful acoustic differences associated with COPD. These findings support the integration of voice-based assessments into eHealth platforms for non-invasive COPD screening and monitoring, which is pending further Validation on larger populations.

Keywords: Chronic obstructive pulmonary disease; Effect size; Mel-frequency cepstral coefficient; Mobile phone-recorded voice data; Statistical analysis; Voice features; Vowel quadrilateral space.

Clinical Trial: NCT06705647

Introduction

Chronic Obstructive Pulmonary Disease (COPD) is a progressive condition characterized by reduced exhalation capacity and lung function, driven by tissue degradation and an inflammatory response in the lungs [1,2]. In addition to its respiratory impact, COPD is classified as a non-laryngeal aerodigestive voice disorder, according to the Classification Manual for Voice Disorders [3]. High rates of underdiagnosis and misdiagnosis reported in various studies highlight the challenges in accurately identifying individuals with COPD, emphasizing the need for improved diagnostic approaches [4]. Timely diagnosis and treatment lead to less use of healthcare resources as compared to those who receive care based on common clinical practices [5].

Many researchers realize this need for timely intervention as they focus on leveraging the power of machine learning (ML) to analyze different and large datasets for the early diagnosis of COPD [6–9]. Earlier research has indicated that ML could be effectively utilized to differentiate between healthy voices and those of individuals diagnosed with COPD [10–12]. The idea is to utilize the data-based learning ability of ML and applying it to health data to analyze the patterns for supporting early decision-making [13]. In this context, a growing area of research focuses on using voice as a biomarker, utilizing the vocal features for supporting decision-making in clinical practices [14].

ML-aided voice analysis is based on extracting vocal properties such as acoustic, prosodic, and linguistic features and using them for training different ML models to support a clinical outcome [10,15]. These studies usually focus on the predictive performance of the ML models, and some of them pay attention to the importance of features in the ML model's output. However, this approach often does not elucidate the pathological relationship between the disease and the vocal features that contribute to it. Sustained vowel phonation contains enough acoustic information to assess vocal quality, with the advantage of stimuli being resistant to language differences [15,16]. As a result, sustained vowels are frequently used in ML-based voice research for clinical applications, with the vowel [a:] being one of the most commonly employed [17]. Vocal features that may be extracted from the acoustic voice signal are viewed as biomarkers, that may support cues on individual health, just as any other traditional biomarker [18–20]. Given this conceptual similarity, a better understanding of digital vocal biomarkers and their association with specific diseases may increase scientific evidence and reliability for applying voice as a biomarker for any clinical outcome.

The present study investigates the statistical and practical significance of vocal differences between COPD patients and healthy controls (HC) based on Swedish utterances of the vowel [a:] collected via mobile phones, approaching a real-world scenario. The work is based on using two types of features that are commonly applied in voice recognition and voice-based clinical research [15,17,21,22]: Type

1) baseline acoustic (BLA) features such as jitter, shimmer, harmonic-to-noise ratio (HNR), and Type 2) Mel frequency cepstral coefficients (MFCC).

The main contributions of this study are as follows:

- Analysis of decentralized voice recordings collected via mobile phones to analyze the impact of COPD on vocal characteristics, presenting a cost-effective and scalable approach to disease assessment.
- Detailed statistical analysis of acoustic features between COPD individuals and HC groups across multiple confidence intervals, identifying significant BLA and MFCC features.
- By incorporating effect size analysis, the study measures the magnitude and direction of feature differences between the groups, providing robust evidence for potential diagnostic applications and enhancing clinical relevance.

Related work

In recent years, there has been a growing interest in utilizing voice as a digital biomarker to support clinical decision-making across a variety of health conditions [17]. According to a recent literature review, voice analysis offers a non-invasive, cost-effective approach for health monitoring and disease detection [23]. The review highlighted the extensive use of BLA features and MFCC in classifying pathological voice conditions. However, while MFCCs are widely used in general voice pathology detection, their application in COPD research has remained relatively limited, with most studies focusing on traditional perturbation measures [14]. Furthermore, the majority of these investigations have employed ML models to distinguish between healthy and pathological voices, emphasizing predictive performance over analyses of statistical significance or clinical relevance.

Several studies have examined statistical differences between COPD patients and HC groups voice [1,22,24–26]. These studies performed acoustic and perceptual analyses on vowel phonations and running speech recordings, all collected in controlled environments using standardized recording protocols and equipment. While most of the studies report increased perturbation measures and reduced HNR in COPD patients, conflicting results have also been reported.

For example, Shastry et al. [1] demonstrated that individuals with COPD exhibit lower fundamental frequency, reduced frequency range, increased jitter and shimmer, and elevated noise measures as compared to HC. This aligns with findings from Mohamed and El Maghraby [25], who reported significant correlations between jitter, shimmer, and dysphonia severity in COPD patients. They highlighted additional factors influencing voice quality, such as smoking history and inhaled corticosteroid (ICS) usage, showing that both disease pathology and treatments contribute to vocal disturbances. Similarly, Hassan et al. [24] confirmed that dysphonia is prevalent in COPD patients and directly correlates with pulmonary function decline. They specifically identified reductions in maximum phonation time (MPT) and phonatory efficiency, along with increased phonatory resistance, as the lung function worsens. This reinforces the physiological interplay between respiratory health and vocal function. Saeed et al. [26] similarly reported impaired vocal quality in both COPD and bronchial asthma patients, with acoustic analyses revealing heightened jitter and shimmer and diminished HNR. They noted that these effects were more pronounced in patients using metered-dose inhalers (MDI) as compared to those using dry powder inhalers (DPI), suggesting that medication delivery methods may influence vocal health. Expanding on the relationship between respiratory function and voice, Hassan et al. [24] linked dysphonia severity to pulmonary function metrics, such as forced expiratory volume (FEV1) and MPT. Their study demonstrated that as the lung function declined, phonatory efficiency also decreased, reinforcing the physiological link between respiratory impairment and vocal output.

A common finding in these studies is an increase in vocal perturbation, showed as higher measures of jitter and shimmer in COPD groups, indicating measurable vocal differences between healthy individuals and those diagnosed with COPD. However, Węglarz et al. [22] presented a more complex picture. Their analysis of jitter, shimmer, HNR, and MFCCs revealed that COPD patients exhibited lower jitter and shimmer values as compared to healthy controls, which stands in contrast to earlier studies. Additionally, no significant differences in HNR were found between COPD patients and HC groups. However, MFCC values were significantly higher in COPD patients, positioning spectral features as potentially more sensitive markers of vocal changes than traditional perturbation measures.

Unlike previous studies, which primarily relied on centralized recording environments, this study focuses on Swedish speakers producing the vowel [a:] in decentralized, participant-driven settings via mobile phones, thereby enhancing ecological validity and capturing real-world variability in vocal production. This approach diverges from the strictly controlled laboratory settings of earlier studies and aligns with the increasing focus on scalable, accessible health monitoring solutions. In addition to analyzing BLA features, this study incorporates MFCCs, further extending the analytical scope to spectral features. While the integration of MFCCs aligns with the work of Węglarz et al. [22], the use of mobile recordings in decentralized environments remains a key contribution of this research.

Material and methods

Data characteristics

This study utilized a gender and age-balanced dataset of vocal features extracted from the Swedish utterance of the vowel [a:] recordings, where the details regarding the data acquisition are shared in a previous study [27]. The dataset consists of acoustic and MFCC features from 48 participants (24 males and 24 females with COPD and their corresponding HC counterparts), who provided 1058 recordings (COPD = 429, HC = 629) over time. The recordings were gathered at different intervals, leading to a varying number of recordings for each participant. The total cohort consisted of individuals with a mean age of 72.1 and a ±6.8 standard deviation. Similarly, HC and COPD groups have an age distribution of 72.1 years old with ±6.8 standard deviation and 72.0 years old with ±6.9 standard deviation, respectively. The average rate of forced expiratory volume in one second (FEV1)/ Forced vital capacity (FVC) was 0.61% with a standard deviation of ± 0.11% for the COPD cohort in this study. The study received approval from the Swedish Ethical Review Authority in Umeå (DNR: 2020-01045) and was conducted in accordance with the principles of the Declaration of Helsinki. All participants provided written informed consent permitting the collection of voice samples. The study was conducted at the Blekinge Institute of Technology (BTH) in Sweden.

Inclusion and Exclusion criteria

The inclusion criteria for the COPD group required participants to be 18 years or older with a confirmed diagnosis of COPD. Additionally, they were required to have access to a smartphone and be proficient in using it. For the HC group, participants had to be 18 years or older and have no documented history of voice disorders. Specifically, they must not have had any diagnoses categorized as "non-laryngeal aerodigestive disorders affecting voice," "neurological disorders affecting voice," or "systematic conditions affecting voice" as defined in the Classification Manual for Voice Disorders [28]. Similar to the COPD group, HC participants were also required to have access to and be proficient in using a smartphone.

The exclusion criteria for the COPD group include participants younger than 18 years old, those diagnosed with a voice-affecting disorder unrelated to COPD, and those without access to or proficiency in using a smartphone. For the HC group, the exclusion applies to individuals under 18

years old, those with any form of voice-affecting disorder, or those without access to or proficiency in smartphone use.

Feature extraction

To ensure consistency, the voice recordings were reviewed and preprocessed. It was observed that all recordings shared a sampling frequency of 44.1 kHz, requiring no adjustments. Furthermore, the silent segments typically present at the beginning and end of the vowel recordings were identified and automatically removed, leaving only the relevant vocal portions for analysis, described in detail in a previous study [27]. This preprocessing step ensured that the data contained only the voice activity section and was ready for further evaluation. Analysis in the present study relies on 101 vocal features extracted from vowel [a:] recordings using the Parselmouth and Librosa modules from Python repositories. The Parselmouth module was used to extract BLA features, and the Librosa module was employed to extract MFCC features. The use of BLA features alongside MFCCs is well-established in voice biomarker research, as these features capture both physiological voice stability and spectral characteristics. Their combined use enhances the sensitivity of ML models in detecting subtle vocal changes associated with respiratory, neurological, and affective disorders [17,29,30], which is the ground for their choice. The feature list, associated module, and the description of the individual features are given in Table 1.

Table 1. The list of BLA and MFCC features used in the analysis.

Feature Group	Feature	Length	Module	Description
BLA	Duration	1	Parselmouth	Duration of the voice part
	Mean_F0	1	Parselmouth	Mean fundamental frequency
	Std_F0	1	Parselmouth	The standard deviation of the fundamental frequency
	HNR	1	Parselmouth	Harmonic to Noice Ratio
	Local_jitter	1	Parselmouth	The average absolute frequency difference between two consecutive periods, divided by the average period
	Absolute_jitter	1	Parselmouth	Measure of the absolute difference between a clock edge as specified and its observed position
	Rap_jitter	1	Parselmouth	Relative average perturbation
	PPQ5_jitter	1	Parselmouth	Five-point period perturbation quotient
	DDP_jitter	1	Parselmouth	Divided difference between consecutive periods
	Local_Shimmer	1	Parselmouth	The average absolute amplitude difference between two consecutive periods, divided by the average period
	LocaldB_Shimmer	1	Parselmouth	Decibel representation of Local Shimmer
	APQ3_Shimmer	1	Parselmouth	Three-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average
	APQ5_Shimmer	1	Parselmouth	Five-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average
	APQ11_Shimmer	1	Parselmouth	Eleven-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average
	DDA_Shimmer	1	Parselmouth	Captures the average difference in amplitude variation across consecutive cycles.
	F1_mean-F4_mean	4	Parselmouth	The mean of the four lowest resonant frequencies of the vocal tract.
	F1_median-F4_median	4	Parselmouth	The median of the four lowest resonant frequencies of the vocal tract.
MFCC	MFCC1_mean-MFCC13_mean	13	Librosa	The mean of the 13 Mel Frequency Cepstral Coefficients
	MFCC1_std-MFCC13_std	13	Librosa	The standard deviation of the 13 Mel Frequency Cepstral Coefficients
	DMFCC1_mean-DMFCC13_mean	13	Librosa	The mean first derivative of the 13 Mel Frequency Cepstral Coefficients
	DMFCC1_std-DMFCC13_std	13	Librosa	The standard deviation of the first derivative of the 13 Mel Frequency Cepstral Coefficients
	DDMFCC1_mean-DDMFCC13_mean	13	Librosa	The mean second derivative of the 13 Mel Frequency Cepstral Coefficients

	DDMFCC1_std-DDMFCC13_std	13	Librosa	The standard deviation of the second derivative of the 13 Mel Frequency Cepstral Coefficients

Statistical experiment and analysis

The data, which consisted of subgroups representing COPD patients and HC group, were analyzed to capture vocal differences between these groups. To minimize potential biases due to accent and language characteristics, which could affect phonation and introduce confounding variables, the first and second formant frequencies (F1 (Tongue Height) and F2 (Tongue Advancement)), were plotted on a vowel quadrilateral space plot using the Matplotlib module in Python. This step aimed to visually assess the similarity of vowel utterances across participants. The vowel quadrilateral plot was used to visualize the clustering of similar utterances and divergence in different utterances, indicating group differences in vocal production [31–33].

This method was applied to comparisons between COPD and HC groups, as well as to gender-based subgroups within each group. The analysis takes into account demographic factors such as gender to ensure that these do not confound the results. This approach aims to provide insights into the vocal characteristics associated with COPD and non-COPD individuals while controlling for potential confounding factors related to language and accent.

Based on previous studies indicating that individuals with COPD tend to exhibit higher acoustic values (as discussed in the related work section), the data were statistically analyzed using the Mann-Whitney U test, a non-parametric test appropriate for comparing two independent groups when the data do not follow a normal distribution. The null hypothesis (H_0) stated that there is no difference in acoustic values between COPD patients and healthy controls, while the alternative hypothesis (H_1) posited that COPD patients have higher acoustic values than healthy controls. The Mann-Whitney U test given in Equation 1 was chosen for its suitability in handling heterogeneous data distributions and its robustness when assumptions of normality are not met [34]. See Figure A1 in the appendix for the distribution of the features. The Shapiro-Wilk (SW) test, a powerful tool for checking the normality of data, was also used to assess the distribution of features, which is an essential step in selecting the appropriate statistical test [35–37]. However, having multiple recordings per participant over time posed a risk of overrepresentation, potentially introducing bias by disproportionately influencing the statistical outcomes. To mitigate this risk and to maintain comparable variance between participants, given the varying number of recordings per participant, the data were aggregated by averaging all measures over 10 randomly selected recordings across 100 iterations at the participant level. Selecting 10 recordings per participant provided a balance between retaining a sufficient number of participants and capturing representative within-subject variability. Conducting 100 iterations reduced the impact of random selection bias, ensuring stable and reproducible aggregate measures across participants. This ensured that each participant contributed equally to the analysis, regardless of the total number of recordings. Consequently, the final statistical analysis included only participants with more than 10 recordings, comprising 27 individuals: 13 healthy controls (5 males, 8 females) and 14 individuals with COPD (6 males, 8 females). This approach aligns with the study's goal of comparing groups (COPD vs. HC) rather than individual recordings. Furthermore, the Mann-Whitney U test assumes independence between observations, and aggregation ensures that each participant contributes a single, independent data point. To further evaluate whether the set of significant features collectively distinguished the groups, a multivariate Permanova (Permutational analysis of variance) test was conducted. Permanova offers a robust statistical approach for comparing multivariate voice feature profiles between unbalanced clinical groups, such as individuals with COPD and HC, with varying numbers of participants [38,39]. Unlike traditional parametric tests, it accommodates the non-normal distribution and high dimensionality typical of acoustic biomarkers like MFCCs and BLA features [38,39]. This makes it particularly valuable in identifying clinically meaningful voice differences. To control multiple comparisons and balance the false positives/false negatives rate and p-hacking, the

Benjamini-Hochberg (BH) correction was applied. In parallel, the Bonferroni correction was also used to identify features with the strongest evidence, emphasizing results with the lowest likelihood of false negatives. The analysis was conducted using the Statsmodels and SciPy libraries, with results visualized using Seaborn and Matplotlib. An alpha threshold of 0.05 was set for the test. Equation 2 shows the calculation of *p*-values for each feature in the dataset. For practical significance, enhancing clinical relevance, and strengthening the analysis, effect size calculations using Cliff's Delta were included in the analysis. This metric, suitable for non-parametric tests, was accompanied by a 95% confidence interval for effect sizes to strengthen the analysis [40–46]. Cliff's delta, as given in Equation 3, was chosen because it quantifies the magnitude of differences between two independent groups without assuming normality or equal variance, making it ideal for non-normally distributed acoustic features. The effect size categorization follows standard thresholds, where values indicate negligible, small, medium, or large effects, providing a more meaningful interpretation of statistical differences [40–44]. The effect size was categorized as below [43]:

$$
\begin{aligned}
|\,\delta\,| &> 0.47 \rightarrow \text{Large effect} \\
0.47 \leq |\,\delta\,| &> 0.33 \rightarrow \text{Medium effect} \\
0.33 \leq |\,\delta\,| &> 0.15 \rightarrow \text{Small effect} \\
|\,\delta\,| &\leq 0.15 \rightarrow \text{Negligible effect}
\end{aligned}
$$

$$
\begin{aligned}
U_1 &= R_1 - \frac{n_1(n_1+1)}{2} \\
U_2 &= R_2 - \frac{n_2(n_2+1)}{2} \\
\mathrm{U} &= \min(n_1, n_2)
\end{aligned}
\tag{1}
$$

where

- R_1 = rank sum for COPD
- R_2 = rank sum for HC
- n_1, n_2= number of samples in each group
- U = Mann–Whitney U statistic (In the implementation, the *alternative* parameter was set to ‘greater' in the Python code.)

$$
\begin{aligned}
\mu\mathrm{U} &= \frac{n_1 n_2}{2} \\
\sigma_U &= \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \\
Z &= \frac{U - \mu\mathrm{U}}{\sigma_U} \\
\mathrm{p} &= 1 - \mathrm{P}(Z)
\end{aligned}
\tag{2}
$$

where

- μU = expected U
- σ_U = standard error for U
- n_1, n_2= number of samples in each group
- Z = z-score
- P(Z)= cumulative distribution function of normal distribution.
- p = p-value

$$\delta = \frac{1}{n_1, n_2} \sum_{x \in COPD} \sum_{y \in HC} sign(x - y)$$

$$sign(x - y) \begin{cases} 1, if\ x > y\ (COPD\ value\ > HC\ value) \\ 0, if\ x = y\ (COPD\ value = HC\ value) \\ -1, if\ x < y\ (COPD\ value\ < HC\ value) \end{cases} \quad (3)$$

where

- x = feature value for COPD
- y = feature value for HC
- n_1, n_2 = number of samples in each group
- δ = Cliff's delta

Results

The vowel quadrilateral plot for the Swedish utterance of the vowel [a:] is presented in Figure 1 using the formant frequencies F1 and F2. The figure includes six different plots, each comparing HC groups, COPD patients, and gender-based subgroups. Each plot contains ellipses that consist of 95% of the data associated with each group, represented by the same color of data points. Figure 1(a), located at the top-left, represents all recordings, where the non-normalized F1 and F2 frequencies range from approximately 400–1200 Hz and 600–1300 Hz, respectively. The group-wise decomposition reveals a decrease in both F1 and F2 frequencies by around 100 Hz for the COPD group as compared to HC participants, while the F2 range remains consistent (Figure 1(b)). Gender-wise decomposition indicates that male voices exhibit lower F1 and F2 frequencies, whereas female voices show higher frequencies, with some overlap between genders (Figure 1(c)). When combining all groups, visual distinctions between subgroups become less apparent (Figure 1(d)). However, when examining the subgroups individually, COPD, HC, female, and male from right to left in the second row (Figure 1(e)), a higher degree of similarity in the articulation of the vowel [a:] is observed. The most distinct subgroup difference appears in the gender decomposition within the HC group, where the ellipses for male and female participants show minimal overlap (Figure 1(f)). The plots in Figure 1(g) and Figure 1(h) represent the COPD and HC groups by gender, respectively, and do not show a strong distinction. Overall, Figure 1 demonstrates that mobile phones are capable of capturing articulatory and dialectal characteristics.

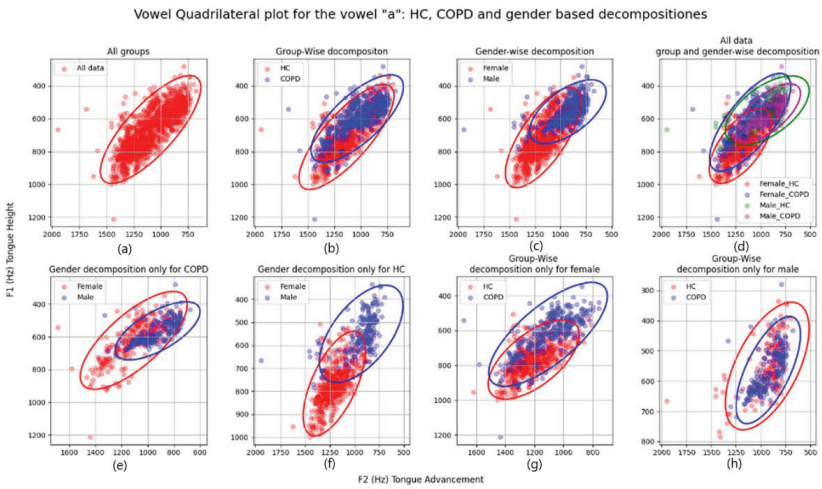


Figure 1. Vowel quadrilateral space shows the vowel [a:] articulation and its acoustic properties.

The Man-Whitney U-test result in Figure 2 shows notable differences between COPD ptients and HCs regarding several acoustic and MFCC features. Out of 101 features, 38 features presented in Table 1 indicate statistical significance ($p<0.05$), leading to the rejection of the null hypothesis. The features represented by red bars suggest higher U-values, signifying greater distinction between the two groups. The remaining features represented by blue bars did not meet the significance threshold ($p \geq 0.05$), suggesting no significant differences between groups for those features.

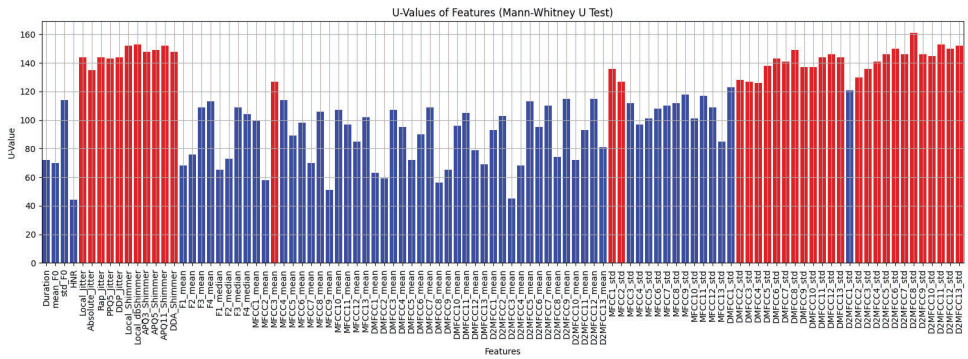


Figure 2. U-statistic ranks for all features, where the red bars represent the statistically significant ones (p<0.05).

The most significant features are highlighted in Figure 3 in ascending order. The 38 features in the figure depict varying levels of significance. It is observable that the features with the highest U-values exhibit a higher level of significance, where 26 features remained under the p-value of 0.01. Six features end up between p-values of 0.01 and 0.02. The remaining 2/6 features occur between the p-values of 0.03 and 0.04. The last four features find a place between the p-value of 0.04 and the threshold alpha of 0.05. Half of the significant features, 26/38, are found to be under the significance level of 0.01 p-value, indicating a high potential for a distinction between COPD patients and HC group. Considering the significant features that remain under the p-value of 0.02, which corresponds to almost 84% (32/38) of the features in Figure 3, high distinctive probabilities regarding the two groups are shown.

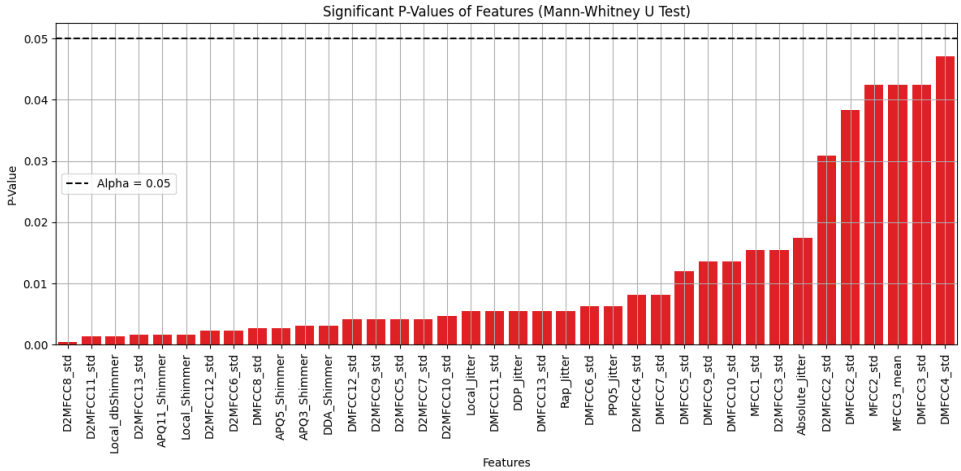


Figure 3. The features where the p-value remains under the significance threshold alpha 0.05, starting from the lowest p-value from the left.

After applying the BH and Bonferroni corrections to account for multiple comparisons, the number of statistically significant features was reduced to 29 after the BH correction and one after the Bonferroni correction, represented by red and blue bars in Figure 4, respectively. Despite this adjustment, 29 out of 38 features remained below the p-value threshold of 0.05, showing their potential to distinguish COPD patients and HC voices. Notably, the feature *D2MFCC8_std* remained statistically significant after both the BH and Bonferroni corrections. This distinction is further supported by the multivariate PERMANOVA test between HC and COPD patient voices using the 29 significant features, which yielded a significant result ($p = 0.019$). Multiple BLA features related to jitter and shimmer measures showed significant differences, with higher values observed in the COPD group as compared to the HC group. Among these, shimmer-based measures demonstrated the highest U-values. Many MFCC features, including both first- and second-order derivatives of standard deviations, also showed statistically significant differences between the groups.

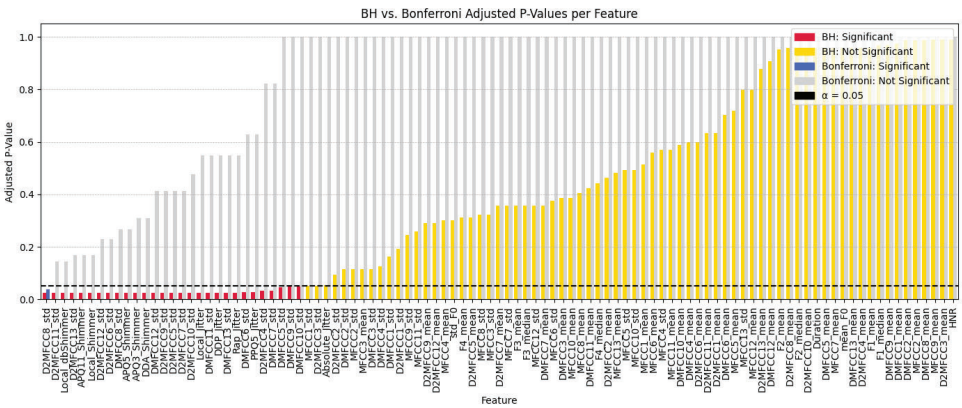


Figure 4. the significant features after the Benjamini-Hochberg and Bonferroni corrections.

The effect size for all extracted features was calculated using Cliff's Delta (δ) along with their respective 95% confidence intervals, as shown in Figure 5. The features are ordered from left to right based on their effect size, with the most discriminative features appearing on the left. Features

represented by red boxplots (n=29) were identified as statistically significant ($p < 0.05$) after the p correction and 34 features (marked by red dots) exhibited large effect sizes ($\delta \geq 0.47$). Additionally, 11 features, indicated by magenta-colored dots, fall within the medium effect size range. Among the remaining features, 36 out of 56 are classified as having a small effect size. The blue dotted 20 features represent the features with negligible effect size.

Additionally, 5 non-statistically significant features depict a large effect size. All of these 34 features have a 95% confidence interval spanning between small and large effect sizes, indicating important discriminative information. Features such as D2MFCC3_mean and HNR exhibit a large effect size in the negative range, signifying systematic differences between COPD patient and HC groups and higher values for HC group regarding these two features. It is also evident that most of the features that were not statistically significant have a wide effect size distribution, suggesting unreliability in their discriminative power. In contrast, features with large effect sizes tend to have a narrower effect size distribution, indicating higher reliability in their discriminative power. As the features progress towards the right side of the plot, the effect sizes decrease, with some confidence intervals crossing the zero threshold. This suggests that these features provide little to no practical discriminative power between the groups.

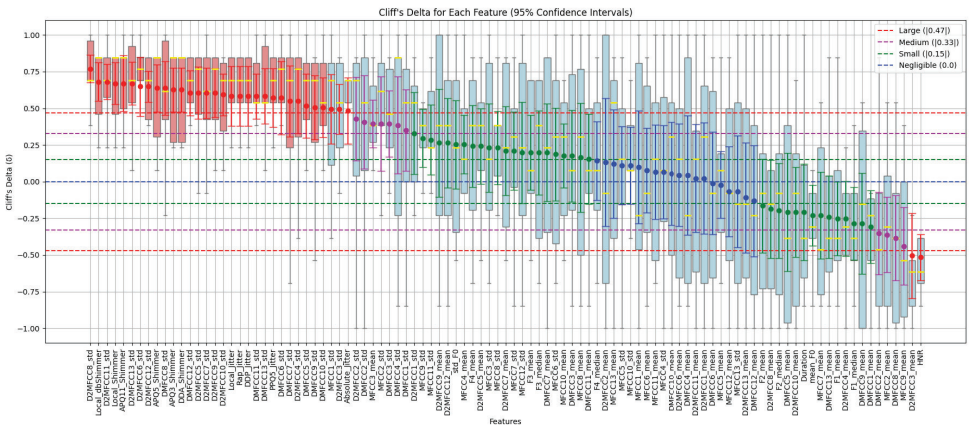


Figure 5. Clif's Delta effect size results for non-parametric test starting from the highest δ on the left side. The boxplots depict the distribution of δ among participants and each feature. The red colored boxplots indicate the significant features (p<0.05). The yellow-colored lines within the boxplots represent the median, and the dots with error bars represent the mean with 95% confidence intervals for each delta (δ).

Discussion

This study set out to investigate whether vocal characteristics, specifically BLA and MFCC features, differ significantly between individuals with COPD and HC when collected via mobile phones in participant-driven settings. The results of this study underline the differences between BLA and MFCC features extracted from COPD patient and HC voice recordings, offering insights into potential vocal biomarkers for COPD diagnosis. Through the integration of statistical significance testing and effect size analyses, the study not only confirms the presence of vocal differences but also evaluates their clinical relevance, an aspect often overlooked in prior machine learning-driven research. The analysis of the results has led to the following findings:

1. Voice recordings collected from different mobile phones retain key articulatory and dialectic characteristics, confirming that mobile-based voice analysis is a viable approach for assessing vocal features.

2. Statistically significant differences were identified between COPD patient and HC groups, suggesting a strong potential for differentiation based on vocal features.
3. A high magnitude of differences was observed between COPD patients and HC vocal features, demonstrating potential integration into clinical assessment protocols for diagnosing and monitoring COPD.

The vowel quadrilaterals presented in the results section (Figure 1) help to understand how the utterance of the vowel [a:] differs among the voice recordings. Having 95% of the recordings within the ellipse shows that the vast majority of the participants articulate the vowel [a:] similarly. Compared to many studies that usually collect the recording in a controlled environment using professional microphones [11,12,17,47], this study shows that even mobile phones can be effective in capturing acoustic information in user-specific uncontrolled settings regarding the distance to the mic and background noise. These findings are even supported by prior research results highlighting the potential of using mobile phones in clinical settings [48]. This understanding does not only apply to specific brands or types of microphones [49], which supports the methodology of using different types and brands in the present study.

The statistical analysis of the features reveals the 29 significant features in Figure 3 exhibit a high probability of distinction between the groups, and more than half of them are below the ($p<0.01$) in regards to associated U-values, indicating strong evidence that those features are higher for the COPD patient group, that supports the initial assumption for this study, which was based on the previous study results [1,24–26]. Alongside BLA features (e.g., jitter and shimmer measures), MFCC and MFCC derivatives also show differences. These findings emphasize that both traditional acoustic metrics and advanced spectral features can provide complementary insights. The features with the largest U-values, demonstrate the potential of these variables as robust discriminators between COPD patient and HC groups. The statistical results show alignment with majority of previous studies suggesting higher jitter and shimmer measures for individuals with COPD diagnosis [1,24–26].

The 34/101 (34%) of the features in the large effect size zone imply great practical significance ($\delta \geq 0.47$) for the features in this zone. However, some features, such as MFCC1_std, D2MFCC3_std, Absolute_jitter, D2MFCC3_mean and HNR, with a large effect size, do not remain among the statistically significant features. This finding highlights the importance of effect size measures in addition to p-values, as effect sizes provide essential insights into the magnitude and clinical relevance of the observed differences [50]. Recent literature emphasizes the growing recognition that reliance solely on p-values can be misleading, as statistical significance does not necessarily equate to practical or clinical significance. While p-values only indicate the probability of observing an effect under the null hypothesis, effect sizes offer a direct measure of the strength and relevance of the relationship between variables [51]. As suggested in previous research, incorporating effect size estimates, especially in fields like medical science, psychology, and health technology, promotes a more nuanced understanding of study outcomes [50,52]. This study adheres to these principles by reporting both effect sizes with associated confidence intervals, p-values, and corrected p-values, thereby enhancing the interpretability of the findings and ensuring that features with large effect sizes, even if not statistically significant, are not overlooked in clinical assessments.

From a health technology perspective, the identification of 34 jitter and MFCC-based spectral features as potential biomarkers for COPD patients presents opportunities for non-invasive diagnostics and digital health monitoring. Traditional COPD assessment relies on spirometry and imaging techniques, which, while effective, can be costly, invasive, and dependent on clinical visits [53]. On the other hand, the high underdiagnosis rates and the importance of early diagnosis to slow down the progression and better treatment are highlighted in many studies [4,5,47,53–55]. The integration of voice-based biomarkers into health technology platforms, such as smartphone applications, telemedicine services, and Artificial Intelligence (AI)-driven decision support tools, offers a scalable, accessible, and cost-efficient alternative. These features may provide early warning indicators of

phonatory instability, helping clinicians track disease progression and response to treatment remotely, considering the moderate lung capacity of 0,61% of the participants in this study. Additionally, the application of machine learning models to analyze these spectral variations could enhance precision medicine approaches, enabling personalized risk assessment and continuous COPD monitoring without the need for frequent hospital visits. As digital health solutions continue to evolve, the integration of these biomarkers into wearable technology, remote monitoring platforms, and AI-assisted diagnostics could revolutionize COPD care, making early intervention and disease management more proactive, patient-centric, and data-driven. All aforementioned potential digital solutions may help reduce healthcare resource use and offer cost-effective alternative tools in COPD assessment.

While positive effect sizes indicate that individuals with a COPD diagnosis tend to have higher values for these features compared to the HC group, negative effect sizes suggest that the COPD patient group tends to have lower values relative to the HC group. In that regard, the presence of 34% of features in the large effect size zone ($\delta \geq 0.47$) underscores the practical significance of these voice parameters in distinguishing COPD patients from healthy controls. The strong effects observed in features such as jitter and shimmer measures, alongside mostly the second derivative standard deviations of MFCC components, highlight the profound impact of COPD on phonatory stability and spectral consistency [1,14]. According to many studies, the increased jitter values reflect heightened pitch instability [56–61], likely caused by weakened vocal fold control and airflow limitations, which contribute to the characteristic breathy and rough voice quality observed in COPD [1]. Additionally, the elevated D2MFCC values indicate greater spectral variability, suggesting that respiratory muscle fatigue and airflow resistance may have the potential to disrupt speech-breathing coordination, leading to unstable articulatory movements [62,63]. These findings emphasize the clinical importance of 34 jitter and MFCC-based spectral features as potential biomarkers for COPD assessment, particularly in regard to the participants' 0,61% moderate lung capacity, which offers promising applications in early detection, disease progression monitoring, and non-invasive screening approaches.

Combining the vowel quadrilateral plot to visualize the acoustic characteristics of the groups, not relying on only a statistical p-value threshold by including different levels of confidence intervals into the analysis, applying p-value correction to reflect different clinical implications such as balanced precision or minimizing false negatives, and providing information for practical significance by adding effect size measures for better clinical relevance are the strengths of this research. Consequently, the small sample size of this research and the geographical homogeneity of the participant pool pose a limitation from the perspective of generalizability. However, even though these limitations exist, other research results that investigate the acoustic characteristics of Swedish vowels give an intuition that the results might be expandible in other regions, too, since the acoustic characteristics collected in other regions in Sweden show high-level similarities regarding the utterance of the vowel [a:] [64].

While this study identifies significant acoustic and statistical differences in the vowel [a:] between COPD patients and HC, it does not account for other conditions that may similarly affect voice production. Disorders such as neuromuscular diseases, laryngeal pathologies, and other respiratory conditions (e.g., asthma, vocal fold paralysis, or Parkinson's disease) can also lead to changes in jitter, shimmer, and MFCC-based spectral features [26,58,65]. This raises a potential limitation in diagnostic specificity, as the observed differences may not be exclusive to COPD patients but rather indicative of broader phonatory and respiratory impairments. Given this potential overlap, rather than suggesting these acoustic markers for standalone diagnostic purposes, they may be more suited for monitoring voice-related changes over time in COPD patients or as part of a multi-modal assessment approach alongside clinical evaluation.

Future research should consider expanding the analysis to include a wider range of clinical populations, investigating how voice characteristics differ not only between COPD patients and HC

individuals but also among other conditions that impact respiratory and phonatory functions. By incorporating additional disease groups, applying machine learning classification models, and utilizing larger, more diverse datasets, future studies can enhance the diagnostic accuracy and specificity of voice-based digital biomarkers for COPD and related disorders. Nonetheless, the conflicting results mentioned in the related work section might be caused by the effects of differences in language or type of voice recordings, which underscore a need for further investigation in future research.

Additionally, the present study involved a cross-sectional analysis of the vocal characteristics using BLA and MFCC features of COPS and HC groups. For the brother application, future research could consider multi-linguistic analysis and longitudinal studies to track changes over time in larger populations with greater demographic variety, providing insights into the disorder's progression and potential causal relationships. Furthermore, including different voice and speech characteristics might help to identify more features being affected by COPD.

Conclusion

This study demonstrates the feasibility and potential of mobile phone-based voice analysis for COPD assessment using voice recordings collected via mobile phones in decentralized, participant-driven environments. The analysis of vowel space plots confirmed that variations across different mobile devices do not compromise the ability to capture key articulatory characteristics of the vowel [a:], supporting the robustness of mobile voice data collection.

From acoustic, statistical, and practical perspectives, the findings confirm that COPD significantly impacts voice production, with many BLA and MFCC features exhibiting higher values in affected individuals. The identification of statistically significant features with large effect sizes highlights the potential of BLA and MFCC features as non-invasive digital biomarkers for COPD assessment. These biomarkers can potentially serve as a foundation for the development of decision-support tools for remote monitoring and early detection, offering a scalable and accessible alternative to traditional diagnostic methods. However, further research is required to validate these results in broader clinical settings, elucidate the causality aspects, and explore their integration into routine healthcare applications.

Data availability

The raw voice recordings cannot be made available due to ethical and general data protection regulations. However, an anonymized version of the feature dataset used in the present study can be made available from the corresponding author's institution upon reasonable request.

Glossary

AI: Artificial Intelligence

BH: Benjamini-Hochberg

BLA: Base Line Acoustic

BTH: Blekinge Institute of Technology

COPD: Chronic Obstructive Pulmonary Disease

DPI: Dry Powder Inhalers

FEV1: Forced Expiratory Volume in one second

FVC: Forced Vital Capacity

HC: Healthy Controls

HNR: Harmonic-to-noise Ratio

ICS: Inhaled Corticosteroid

MDI: Metered-dose Inhalers

ML: Machine Learning

MFCC: Mel-Frequency Cepstral Coefficient

MPT: Maximum Phonation Time

SW: Shapiro-Wilk

CRediT authorship contribution statement

Alper Idrisoglu: Writing – review & editing, Writing – original draft, Methodology, Conceptualization. Ana Luiza Dallora Moraes: Writing – review and editing. Abbas Cheddad: Writing – review and editing, Validation. Susanna Whitling: Writing – review and editing, Validation. Peter Anderberg: Writing – review and editing. Andreas Jakobsson: Writing – review and editing, Validation, Methodology. Johan Sanmartin Berglund: Writing – review and editing, Validation, Supervision.

Declaration of competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Excellence Center at Linköping –Lund in Information Technology (ELLIIT).

Ethics approval and consent to participate

The study was approved by the Swedish Ethical Review Authority in Umeå (DNR: 2020-01045) and was conducted in accordance with the principles of the Declaration of Helsinki. All participants provided written informed consent.

Acknowledgments

The authors appreciate the participants' strong interest, active engagement, and valuable contributions to the research.

References

1. Shastry A, Balasubramanium RK, Acharya PR. Voice Analysis in Individuals with Chronic Obstructive Pulmonary Disease. International Journal of Phonosurgery & Laryngology. 2014;4: 45–49. doi:10.5005/jp-journals-10023-1081

2. Pauwels RA, Buist AS, Calverley PMA, Jenkins CR, Hurd SS. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease. Am J Respir Crit Care Med. 2001;163: 1256–1276. doi:10.1164/ajrccm.163.5.2101039

3. Boers E, Barrett M, Su JG, Benjafield AV, Sinha S, Kaye L, et al. Global Burden of Chronic Obstructive Pulmonary Disease Through 2050. JAMA Network Open. 2023;6: e2346598. doi:10.1001/jamanetworkopen.2023.46598

4. Axelsson M, Backman H, Nwaru BI, Stridsman C, Vanfleteren L, Hedman L, et al. Underdiagnosis and misclassification of COPD in Sweden – A Nordic Epilung study. Respiratory Medicine. 2023;217. doi:10.1016/j.rmed.2023.107347

5. Aaron SD, Vandemheen KL, Whitmore GA, Bergeron C, Boulet L-P, Côté A, et al. Early Diagnosis and Treatment of COPD and Asthma — A Randomized, Controlled Trial. New England Journal of Medicine. 2024;390: 2061–2073. doi:10.1056/NEJMoa2401389

6. Mahdavi H, Rahbarpour S, Hosseini-Golgoo SM, Jamaati H. A single gas sensor assisted by machine learning algorithms for breath-based detection of COPD: A pilot study. Sensors and Actuators A: Physical. 2024;376: 115650. doi:10.1016/j.sna.2024.115650

7. Liu Y, Song J. Predictive analysis of the psychological state of charismatic leaders on employees' work attitudes based on artificial intelligence affective computing. Front Psychol. 2022;13. doi:10.3389/fpsyg.2022.965658

8. Kumar S, Bhagat V, Sahu P, Chaube MK, Behera AK, Guizani M, et al. A novel multimodal framework for early diagnosis and classification of COPD based on CT scan images and multivariate pulmonary respiratory diseases. Computer Methods and Programs in Biomedicine. 2024;243: 107911. doi:10.1016/j.cmpb.2023.107911

9. Gálvez-Barrón C, Pérez-López C, Villar-Álvarez F, Ribas J, Formiga F, Chivite D, et al. Machine learning for the development of diagnostic models of decompensated heart failure or exacerbation of chronic obstructive pulmonary disease. Sci Rep. 2023;13: 12709. doi:10.1038/s41598-023-39329-6

10. Idrisoglu A, Dallora AL, Cheddad A, Anderberg P, Jakobsson A, Sanmartin Berglund J. COPDVD: Automated Classification of Chronic Obstructive Pulmonary Disease on a New Developed and Evaluated Voice Dataset. Rochester, NY; 2024. doi:10.2139/ssrn.4713043

11. Farrús M, Codina-Filbà J, Reixach E, Andrés E, Sans M, Garcia N, et al. Speech-Based Support System to Supervise Chronic Obstructive Pulmonary Disease Patient Status. Applied Sciences. 2021;11: 7999. doi:10.3390/app11177999

12. Bringel KA, Leone DCMG, de C. Firmino JVL, Rodrigues MC, de Melo MDT. Voice Analysis and Neural Networks as a Clinical Decision Support System for Patients With Lung Diseases. Mayo Clinic Proceedings: Digital Health. 2024;2: 367–374. doi:10.1016/j.mcpdig.2024.06.006

13. Deo RC. Machine Learning in Medicine. Circulation. 2015;132: 1920–1930. doi:10.1161/CIRCULATIONAHA.115.001593

14. Kapetanidis P, Kalioras F, Tsakonas C, Tzamalis P, Kontogiannis G, Karamanidou T, et al. Respiratory Diseases Diagnosis Using Audio Analysis and Artificial Intelligence: A Systematic Review. Sensors. 2024;24: 1173. doi:10.3390/s24041173

15. Fagherazzi G, Fischer A, Ismael M, Despotovic V. Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice. Digit Biomark. 2021;5: 78–88. doi:10.1159/000515346

16. Gerratt BR, Kreiman J, Garellek M. Comparing Measures of Voice Quality from Sustained Phonation and Continuous Speech. Journal of Speech, Language, and Hearing Research. 2016;59: 994–1001. doi:10.1044/2016_JSLHR-S-15-0307

17. Idrisoglu A, Dallora AL, Anderberg P, Sanmartin Berglund J. Applied machine learning techniques to diagnose voice-affecting conditions and disorders: A systematic literature review. JMIR Medical Informatics. 2023; 18.

18. Babrak LM, Menetski J, Rebhan M, Nisato G, Zinggeler M, Brasier N, et al. Traditional and Digital Biomarkers: Two Worlds Apart? Digit Biomark. 2019;3: 92–102. doi:10.1159/000502000

19. Dorsey ER, Papapetropoulos S, Xiong M, Kieburtz K. The First Frontier: Digital Biomarkers for Neurodegenerative Disorders. Digit Biomark. 2017;1: 6–13. doi:10.1159/000477383

20. Robin J, Harrison JE, Kaufman LD, Rudzicz F, Simpson W, Yancheva M. Evaluation of Speech-Based Digital Biomarkers: Review and Recommendations. Digit Biomark. 2020;4: 99–108. doi:10.1159/000510820

21. Abdul ZKh, Al-Talabani AK. Mel Frequency Cepstral Coefficient and its Applications: A Review. IEEE Access. 2022;10: 122136–122158. doi:10.1109/ACCESS.2022.3223444

22. Węglarz K, Szczygieł E, Masłoń A, Blaut J. Assessment of breathing patterns and voice of patients with COPD and dysphonia. Respiratory Medicine. 2025;240: 108012. doi:10.1016/j.rmed.2025.108012

23. Abdulmajeed NQ, Al-Khateeb B, Mohammed MA. A review on voice pathology: Taxonomy, diagnosis, medical procedures and detection techniques, open challenges, limitations, and recommendations for future directions. Journal of Intelligent Systems. 2022;31: 855–875. doi:10.1515/jisys-2022-0058

24. Hassan MM, Hussein MT, Emam AM, Rashad UM, Rezk I, Awad AH. Is insufficient pulmonary air support the cause of dysphonia in chronic obstructive pulmonary disease? Auris Nasus Larynx. 2018;45: 807–814. doi:10.1016/j.anl.2017.12.002

25. Mohamed EE, El maghraby RA. Voice changes in patients with chronic obstructive pulmonary disease. Egyptian Journal of Chest Diseases and Tuberculosis. 2014;63: 561–567. doi:10.1016/j.ejcdt.2014.03.006

26. Saeed AM, Riad NM, Osman NM, Khattab AN, Mohammed SE. Study of voice disorders in patients with bronchial asthma and chronic obstructive pulmonary disease. Egypt J Bronchol. 2018;12: 20–26. doi:10.4103/ejb.ejb_34_17

27. Idrisoglu A, Dallora AL, Cheddad A, Anderberg P, Jakobsson A, Sanmartin Berglund J. COPDVD: Automated classification of chronic obstructive pulmonary disease on a new collected and evaluated voice dataset. Artificial Intelligence in Medicine. 2024;156: 102953. doi:10.1016/j.artmed.2024.102953

28. Verdolini K, Rosen CA, Branski RC, editors. Classification Manual for Voice Disorders–I. New York: Psychology Press; 2005. doi:10.4324/9781410617293

29. Tsanas A, Little MA, McSharry PE, Spielman J, Ramig LO. Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease. IEEE Transactions on Biomedical Engineering. 2012;59: 1264–1271. doi:10.1109/TBME.2012.2183367

30. Rogers HP, Hseu A, Kim J, Silberholz E, Jo S, Dorste A, et al. Voice as a Biomarker of Pediatric Health: A Scoping Review. Children. 2024;11: 684. doi:10.3390/children11060684

31. Berisha V, Sandoval S, Utianski R, Liss J, Spanias A. Characterizing the distribution of the quadrilateral vowel space area. The Journal of the Acoustical Society of America. 2014;135: 421–427. doi:10.1121/1.4829528

32. Honda K, Maeda S, Hashi M, Dembowski JS, Westbury JR. Human palate and related structures: their articulatory consequences. Proceeding of Fourth International Conference on Spoken Language Processing ICSLP '96. 1996. pp. 784–787 vol.2. doi:10.1109/ICSLP.1996.607480

33. Kent RD, Rountrey C. What Acoustic Studies Tell Us About Vowels in Developing and Disordered Speech. Am J Speech Lang Pathol. 2020;29: 1749–1778. doi:10.1044/2020_AJSLP-19-00178

34. Tai KY, Dhaliwal J, Balasubramaniam V. Leveraging Mann–Whitney U test on large-scale genetic variation data for analysing malaria genetic markers. Malar J. 2022;21: 79. doi:10.1186/s12936-022-04104-x

35. Wei J. The adoption of repeated measurement of variance analysis and Shapiro—Wilk test. Front Med. 2022;16: 659–660. doi:10.1007/s11684-021-0908-8

36. Khatun N. Applications of Normality Test in Statistical Analysis. Open Journal of Statistics. 2021;11: 113. doi:10.4236/ojs.2021.111006

37. González-Estrada E, Villaseñor JA, Acosta-Pech R. Shapiro-Wilk test for multivariate skew-normality. Comput Stat. 2022;37: 1985–2001. doi:10.1007/s00180-021-01188-y

38. Anderson MJ. Permutational Multivariate Analysis of Variance (PERMANOVA). Wiley StatsRef: Statistics Reference Online. John Wiley & Sons, Ltd; 2017. pp. 1–15. doi:10.1002/9781118445112.stat07841

39. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecology. 2001;26: 32–46. doi:10.1111/j.1442-9993.2001.01070.pp.x

40. Steyn HS. Non-parametric tests with effect sizes. Statistical consultation services. 2020 [cited 5 Dec 2024]. Available: http://natural-sciences.nwu.ac.za/sites/natural-sciences.nwu.ac.za/files/files/SDK/non-pts.pdf

41. Macbeth G, Razumiejczyk E, Ledesma RD. Cliff's Delta Calculator: A non-parametric effect size program for two groups of observations. Universitas Psychologica. 2011;10: 545–555.

42. Rovetta A. A Framework to Avoid Significance Fallacy. Cureus. 15: e40242. doi:10.7759/cureus.40242

43. Meissel K, Yao ES. Using Cliff's Delta as a Non-Parametric Effect Size Measure: An Accessible Web App and R Tutorial. Practical Assessment, Research, and Evaluation. 2024;29. doi:10.7275/pare.1977

44. Kala Z. Global Sensitivity Analysis of Structural Reliability Using Cliff Delta. Mathematics. 2024;12: 2129. doi:10.3390/math12132129

45. Barnes T, Moore SC, Osatuke K. Testing Significance Tests: A Simulation with Cliff's Delta, t-tests, and Mann-Whitney U. National Center for Organizational Development, Department of Veteran Affairs. 2018 [cited 17 Feb 2025]. Available: https://nces.ed.gov/FCSM/pdf/H3_Barnes_2018FCSM.pdf

46. Bais F, Neut J van der. Adapting the Robust Effect Size Cliff's Delta to Compare Behaviour Profiles. Survey Research Methods. 2022;16: 329–352. doi:10.18148/srm/2022.v16i2.7908

47. Marepalli GS, Kollu PK, Inavolu MD. Early Detection of Chronic Obstructive Pulmonary Disease in Respiratory Audio Signals Using CNN and LSTM Models. 2024 IEEE International Conference on Contemporary Computing and Communications (InC4). 2024. pp. 1–6. doi:10.1109/InC460750.2024.10648991

48. Petrizzo D, Popolo PS. Smartphone Use in Clinical Voice Recording and Acoustic Analysis: A Literature Review. Journal of Voice. 2021;35: 499.e23-499.e28. doi:10.1016/j.jvoice.2019.10.006

49. Awan SN, Bahr R, Watts S, Boyer M, Budinsky R, null null, et al. Validity of Acoustic Measures Obtained Using Various Recording Methods Including Smartphones With and Without Headset Microphones. Journal of Speech, Language, and Hearing Research. [cited 21 May 2024]. doi:10.1044/2024_JSLHR-23-00759

50. Hojat M, Xu G. A Visitor's Guide to Effect Sizes – Statistical Significance Versus Practical (Clinical) Importance of Research Findings. Adv Health Sci Educ Theory Pract. 2004;9: 241–249. doi:10.1023/B:AHSE.0000038173.00909.f6

51. Tomczak M, Tomczak E. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. 2014 [cited 9 Dec 2024]. Available: https://www.wbc.poznan.pl/publication/413565

52. Kraemer HC, Morgan GA, Leech NL, Gliner JA, Vaske JJ, Harmon RJ. Measures of clinical significance. Journal of the American Academy of Child & Adolescent Psychiatry. 2003;42: 1524–1529.

53. Kahnert K, A. Jörres R, Behr J, Welte T. The Diagnosis and Treatment of COPD and Its Comorbidities. Dtsch Arztebl Int. 2023;120: 434–444. doi:10.3238/arztebl.m2023.027

54. Di Marco F, Balbo P, De Blasio F, Cardaci V, Crimi N, Girbino G, et al. Early management of COPD: where are we now and where do we go from here? A Delphi consensus project. COPD. 2019;Volume 14: 353–360. doi:10.2147/COPD.S176662

55. Ho T, Cusack RP, Chaudhary N, Satia I, Kurmi OP. Under- and over-diagnosis of COPD: a global perspective. Breathe. 2019;15: 24–35. doi:10.1183/20734735.0346-2018

56. Ferrer CA, Torres D, González E, Calvo JR, Castillo E. Effect of different jitter-induced glottal pulse shape changes in periodicity perturbation measures. Sixteenth Annual Conference of the International Speech Communication Association. 2015. Available: https://www.researchgate.net/profile/Diana-Torres-Boza/publication/281557525_Effect_of_Different_Jitter-Induced_Glottal_Pulse_Shape_Changes_in_Periodicity_Perturbation_Measures/links/55edbaac08aef559dc4289d4/Effect-of-Different-Jitter-Induced-Glottal-Pulse-Shape-Changes-in-Periodicity-Perturbation-Measures.pdf

57. Cataldo E, Soize C. A stochastic mechanical model to generate jitter in the production of voiced sounds. In: Papadrakakis M, Papadopoulos V, Stefanou (eds.) G, editors. Proceedings of UNCECOMP 2015. The Island of Crete, Greece; 2015. pp. 1–14. Available: https://hal.science/hal-01158279

58. Li G, Hou Q, Zhang C, Jiang Z, Gong S. Acoustic parameters for the evaluation of voice quality in patients with voice disorders. Annals of Palliative Medicine. 2021;10: 13036–13136. doi:10.21037/apm-20-2102

59. Vasilakis M, Stylianou Y. A mathematical model for accurate measurement of jitter. Fifth International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications. 2007. Available: https://library.oapen.org/bitstream/handle/20.500.12657/54827/9788855180276.pdf?sequence=1#page=20

60. Zewoudie AW, Luque J, Hernando Pericás FJ. Jitter and shimmer measurements for speaker diarization. VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop: proceedings: November 19-21, 2014: Escuela de Ingeniería en Telecomunicación y Electrónica Universidad de

Las Palmas de Gran Canaria: Las Palmas de Gran Canaria, Spain. 2014. pp. 21–30. Available: https://upcommons.upc.edu/handle/2117/27438

61. Upadhya SS, Cheeran AN, Nirmal JH. Statistical comparison of Jitter and Shimmer voice features for healthy and Parkinson affected persons. 2017 second international conference on electrical, computer and communication technologies (ICECCT). IEEE; 2017. pp. 1–6. Available: https://ieeexplore.ieee.org/abstract/document/8117853/

62. Segizbaeva MO, Aleksandrova NP. Inspiratory muscle resistance to fatigue during exercise and simulated airway obstruction. Hum Physiol. 2014;40: 683–689. doi:10.1134/S0362119714050120

63. Fitting JW. Respiratory muscles in chronic obstructive pulmonary disease. Swiss Medical Weekly. 2001;131: 483–486.

64. Persson A. The acoustic characteristics of Swedish vowels. Phonetica. 2024;81: 599–643. doi:10.1515/phon-2024-0011

65. Chiaramonte R, Bonfiglio M. Acoustic analysis of voice in Parkinson's disease: a systematic review of voice disability and meta-analysis of studies. Rev Neurol. 2020;70: 393–405.

Appendix

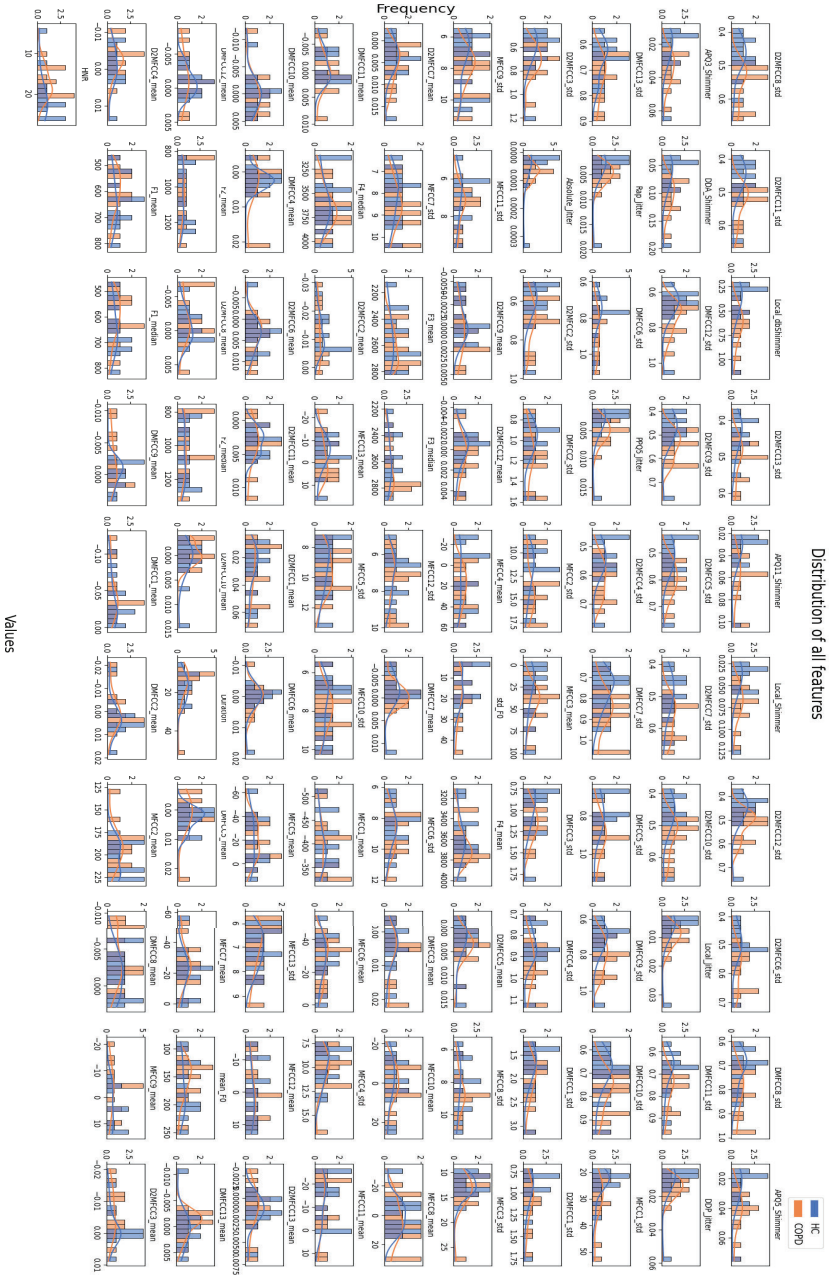


Figure A1. The distribution of all feature values for COPD and HC groups.

Chronic Obstructive Pulmonary Disease (COPD) is a leading cause of morbidity and mortality worldwide, with high underdiagnosis rates due to limitations in current diagnostic methods such as spirometry. This doctoral thesis explores the potential of voice as a digital biomarker to support the assessment of COPD, guided by the principles of Applied Health Technology (AHT), which emphasizes interdisciplinary collaboration and real-world applicability.

The research includes four interconnected studies. Study I presents a systematic literature review of machine learning (ML) applications for voice-affecting disorders, identifying COPD as underrepresented in current research. Study II addresses this gap by collecting a new dataset of vowel [a:] recordings from Swedish-speaking COPD patients and healthy controls once a week in self-determined quiet settings. Voice features, including baseline acoustic (BLA) parameters and Mel-Frequency Cepstral Coefficients (MFCCs), were extracted and used to train three ML classifiers: CatBoost (CB), Random Forest (RF), and Support Vector Machine (SVM). CB demonstrated the highest test accuracy at 78%.

Study III investigates the effects of signal segmentation on model performance and shows that certain temporal segments of voice recordings contain more informative patterns, enhancing classification outcomes by increasing accuracy to 85%. Study IV applies statistical and practical significance tests to compare voice features between COPD and healthy groups. A total of 34 features, including shimmer measures and higher-order MFCC derivatives, were found to meaningfully differentiate the groups.

This thesis reframes the human voice as a source of clinically relevant data, demonstrating how it can be digitized, analyzed, and interpreted using ML to aid COPD assessment. The results indicate that voice-based analysis can provide an accessible, non-invasive, and scalable complement to existing diagnostic tools. By integrating technical, clinical, and ethical perspectives, the thesis contributes new knowledge and practical methodologies that align with AHT's goal of creating value-driven, user-centered healthcare solutions. The findings support future development of mobile and remote voice-based screening tools for COPD and other conditions.

Printed by Media-Tryck, Lund 2025 NORDIC SWAN ECOLABEL 3041 0903

Blekinge Institute of Technology
Doctoral Dissertation Series No. 2025:07
ISSN: 1653-2090
ISBN: 978-91-7295-503-5