



PhD-FSTM-2024-097
The Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 18/12/2024 in Luxembourg
to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG EN BIOLOGIE

A dissertation by

Abir Elbéji

Born on 11 June 1998 in Tunis, (Tunisia)

Exploring Vocal Biomarkers For Disease Screening and Health Monitoring

Dissertation defense committee

Dr. Stéphane Bordas, Chairman
Full professor, University of Luxembourg

Dr. Vladimir Despotovic, Vice Chair
Scientist, Luxembourg Institute of Health

Dr. Guy Fagherazzi, dissertation supervisor
Director of the Department of Precision Health, Luxembourg Institute of Health (LIH)

Dr. Nicholas Cummins
Lecturer, King's College London

Dr. Dimitra Anastasiou
Lead R&T, Scientist, Luxembourg Institute of Science and Technology (LIST)



PhD-FSTM-2024-097

The Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 18/12/2024 in Luxembourg
to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG EN BIOLOGIE

A dissertation by

Abir Elbéji

Born on 11 June 1998 in Tunis, (Tunisia)

Exploring Vocal Biomarkers For Disease Screening and Health Monitoring

Dissertation defense committee

Prof. Dr. Stéphane Bordas, Chairman
Full professor in Computational Mechanics
Faculty of Science, Technology and Medicine,
Department of Engineering, University of Luxembourg

Dr. Nicholas Cummins, Member
Lecturer in AI for speech analysis
Department of Biostatistics and Health Informatics,
King's College London

Dr. Guy Fagherazzi, dissertation supervisor
Director of the Department of Precision Health,
Integrated Group Leader
Deep Digital Phenotyping Research Unit,
Luxembourg Institute of Health

Dr. Reka Toth, Expert in advisory capacity
Scientist
Bioinformatics Platform,
Luxembourg Institute of Health

Dr. Vladimir Despotovic, Vice Chair
Scientist in health data
Bioinformatics Platform
Luxembourg Institute of Health

Dr. Dimitra Anastasiou, Member
Lead R&T, Scientist
Luxembourg Institute of Science and Technology (LIST)

MD. Yael Bensoussan, Expert in advisory capacity
Otolaryngology
Health Department of Otolaryngology-Head and Neck Surgery
University of South Florida

Affidavit

I hereby confirm that the PhD thesis entitled “**Exploring Vocal Biomarkers For Disease Screening and Health Monitoring**” has been written independently and without any other sources than cited. All necessary ethical approvals have been obtained in accordance with law 2018 (incl. transitory regulations).

Luxembourg, 18/12/2024

Abir Elbéji

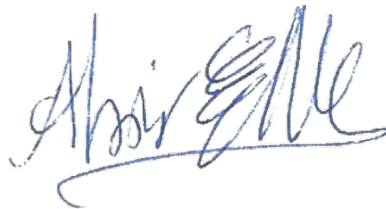
A handwritten signature in blue ink, appearing to read 'Abir Elbéji', with a stylized flourish underneath.

Table of Content

Affidavit.....	1
Table of Content.....	3
Acknowledgments.....	7
Abbreviations.....	9
Abstract.....	10
List of Figures.....	11
Tables.....	12
List of Publications.....	14
Chapter 1	
Objectives.....	15
Chapter 2	
Materials and Methods.....	17
Data Sources.....	18
Predi-COVID.....	18
Study Design.....	18
Data Collection.....	19
Colive Voice.....	19
Study Design.....	19
Data Collection.....	20
Data Processing.....	21
Audio Signal Analysis.....	22
Audio Features.....	22
Source Features.....	23
Formant (Filter) Features.....	23
Spectral Features.....	23
Prosodic Features.....	24
Audio Feature Extraction.....	25
Librosa.....	25
OpenSMILE.....	25
Pretrained Algorithms.....	26
Statistical Analyses.....	28
Chi-2.....	28
Mann U Whitney.....	28
One-way ANOVA.....	28
Bland-Altman.....	28
Feature Scaling and Normalization.....	28
StandardScaler Sklearn.....	29
Dimensionality Reduction.....	29
Feature Selection.....	29
Principal Component Analysis (PCA).....	29
Key AI Concepts.....	29
Machine Learning.....	30

Deep learning.....	31
Transfer Learning.....	33
Performance Evaluation.....	34
Cross-Validation.....	34
Evaluation Metrics.....	34
1. Accuracy.....	35
2. Precision.....	35
3. Recall (Sensitivity or True Positive Rate (TPR)).....	35
4. Specificity (True Negative Rate).....	35
5. F1 score.....	35
6. Area Under the Receiver Operating Characteristic Curve (AUC - ROC)...	35
7. Brier score.....	35
Essential Python libraries used.....	36
Pandas.....	36
NumPy.....	36
Scikit-learn.....	36
TensorFlow.....	36
Keras.....	36
PyTorch.....	37
IT Environment and hardware.....	37
Chapter 3	
Synopsis - General Introduction.....	39
Digital Health.....	40
Digital Biomarkers.....	42
Voice AI.....	42
Voice in Digital Health.....	43
Mechanisms of Voice Production.....	43
Vocal Biomarkers.....	44
Vocal Biomarkers as New Digital Clinical Endpoints.....	44
Applications of Vocal Biomarkers.....	44
Main vocal tasks and examples of disease-specific vocal biomarkers.....	45
Delineation of the PhD Scope.....	47
Objective 1: Symptom Detection and Monitoring.....	48
Contextualizing COVID-19 Symptoms and Fatigue.....	48
Defining and Understanding Fatigue.....	48
Biological Mechanisms of Fatigue in COVID-19.....	48
Persistent Challenges of Long COVID.....	48
Objective 2: Disease Screening.....	49
Diabetes Mellitus and its classification.....	49
Type 2 Diabetes (T2D).....	50
Challenges in managing T2D.....	50
Objective 3: Perceived Health Status Monitoring.....	50
Context and importance of telemonitoring.....	50
Understanding Respiratory Quality of Life (RQoL).....	50

Challenges in RQoL assessment.....	50
Chapter 4	
Vocal biomarkers for symptom detection and monitoring.....	53
Results.....	54
General Workflow.....	54
Study Population Characteristics.....	56
Model Performance and Vocal Biomarker Development.....	57
Discussion.....	57
Fatigue and voice.....	57
Implications of the study.....	58
Limitations of the study.....	59
Advancing fatigue assessment: Fatigue Severity Scale.....	60
Extending the impact beyond COVID-19.....	60
Conclusion.....	60
Chapter 5	
Vocal Biomarkers for Disease Screening.....	61
Results.....	62
General workflow.....	62
Study population characteristics.....	64
Algorithm performance.....	64
Performance stratification based on cofactors.....	65
Discussion.....	67
Voice Alterations and Underlying Mechanisms in T2D.....	67
Implications of the study.....	68
Limitations of the study.....	69
Integration into Voice-Based Screening.....	69
Conclusion.....	70
Chapter 6	
Vocal Biomarkers for Health Status Monitoring.....	71
Results.....	72
General workflow.....	72
Study population characteristics.....	74
Algorithm performance based on socio-demographic/clinical data.....	74
Algorithm performance based on voice recordings.....	75
Algorithm performance based on fused socio-demographic/clinical data and voice recordings.....	76
Discussion.....	78
Voice Alterations and Underlying Mechanisms in RQoL.....	78
Implications of the Study.....	78
Limitations of the Study.....	79
Conclusion.....	79
Chapter 7	
Conclusion and Perspectives.....	81
General Findings.....	82
Vocal Biomarkers for Symptom Monitoring: Fatigue in People with COVID-19.....	82

Vocal Biomarkers for Screening: T2D.....	82
Vocal Biomarkers for Perceived Health Status: RQoL.....	82
Perspectives.....	84
Research Perspectives.....	84
Challenges in Implementing Voice AI in Healthcare.....	85
Broader Implications and Future Research Directions.....	85
Final reflections.....	88
Chapter 8	
Contributions.....	89
Vocalive Platform for Vocal Biomarker Development.....	90
Project background.....	90
Overview of Vocalive.....	90
Services Provided by Vocalive.....	90
Workflow and Technical Aspects.....	91
Future Outlook and Impact.....	91
Participation in the TAILOR-Funded Exchange Program.....	92
Contribution to related work.....	95
Conferences.....	97
Oral communications.....	97
Poster presentations.....	97
References.....	98
Appendix of Original Papers.....	106

Acknowledgments

Pursuing a PhD has been a goal of mine ever since I was introduced to research work during my internships. Even during the lowest moments of this journey, I never once regretted my decision. It was never easy, and at times it felt like a rollercoaster, but sitting here now, writing the final part of my dissertation and expressing my gratitude to those who supported me is a testament to my perseverance. This journey would never have been as enjoyable or fruitful without all of you.

First and foremost, I would like to thank my PhD supervisor, **Dr. Guy Fagherazzi**, director of the Department of Precision Health and Leader of the Deep Digital Phenotyping Unit. Thank you for believing in me from day one, for welcoming me into your Lab, and for ensuring I received everything necessary for my intellectual and personal growth. I would never have reached this point without your constant support, always pushing me to give my best, allowing me to make mistakes and Learn from them. Thank you for showing me what a true leader looks like. Your guidance has been invaluable, and I feel incredibly fortunate to have had you as my supervisor. For that, I can never thank you enough.

I also want to extend my gratitude to **Dr. Vladimir Despotovic** and **Prof. Dr. Luis Leiva** for serving on my thesis monitoring committee and for your invaluable advice. Thank you for agreeing to assess my work and to serve on the defense committee, alongside **Dr. Nicholas Cummins**, **Dr. Dimitria Anastasiou**, **Dr. Reka Toth**, and **Dr. Yael Bensoussan**.

To my colleagues, I find it hard to express just how proud and grateful I am to have been part of such an amazing team. Working with you all has been an immense pleasure, sharing both the highs and lows along the way. **Dr. Charline** (yes, I need to formally address you now), I feel blessed that our paths crossed. You've been there for me every step of the way with your advice and unwavering support. I can never thank you enough (and a special thank you to **Cauchy**, the big fluff ball, whose pictures always brightened my day!). **Dulce**, I'm so glad I met you. I've enjoyed every moment and adventure we've shared. Your intelligence, kindness, and cheerfulness (and of course your clumsiness) have been a joy to experience. I know you'll achieve great things.

Mégane, I've never known anyone with more patience or such a desire to keep learning. You are truly inspiring. **Maurane**, you always bring a smile to my face, especially with your owl adventures. It's incredible to see you embarking on your own PhD journey, proof of your perseverance and your ability to overcome anything. **Kevser**, my dear "Ablam," your advice, and support have meant the world to me, making sure I never felt alone. I'm so thankful that I could share my deepest lows with you and always find comfort in your presence.

Hanin, it was such a pleasure to collaborate with you. Together, we made great things happen. **Aurélie** and **Gloria**, you were always there with helpful advice and a lending hand when I needed it. **Noémie**, **Mariana**, **Valeriya**, and **Milena**, I am really happy to have met you and look forward to working with you.

Thank you all from the bottom of my heart.

I would also like to thank my dear friend **Francesca**. It is wonderful to have you in my life, and I'm so happy that our relationship is more than just professional. You've always been a comforting presence, and your words helped me through every tough moment.

To my best friend **Oussema**, Bay, I know you don't like my English, but this is for you to let you know how grateful I am to have you by my side through it all.

To my housemates: **Walid**, ya ghali, thank you for always bringing joy and cheerfulness into my life. **Janika**, you bring a true sense of home with your presence, thank you so much for all the shared moments even the laziest ones (the best ones btw). **Sven**, I look forward to hearing all about your adventures when you return from your trip. **Dr. Felix**, you are the best PhD (and cook) in all senses, thank you for all the delicious meals you shared with us.

To "Bi3a M9at3a" or **Lila**, **Olfa**, **Madou**, **Nadhem**, **Borchani**, and **Slouma**, thank you for your amazing chats that always brought a smile to my face. I hope you'll always consider me one of you, as I do with all of you.

To **Kais**, I am forever grateful for your support and for constantly encouraging me to push beyond my limits. You were the one who opened the door to this field for me, and while the journey hasn't been without its challenges, I will always wish you nothing but the very best in every aspect of your life.

Finally, to my family, thank you all for your constant support. I hope I've made you proud and will continue to do so. **Dad**, thank you for always believing in me. **Issam**, my brother, I hope I remain a source of inspiration for you, as you always tell me I am. To my dearest **Mum** and my sister, **Haifa**, this achievement, and what you call success, is a testament to your endless support, love, and sacrifices. Mum, your tireless efforts, belief in me, and your presence through all the challenges have been my constant source of strength. Haifa, your encouragement and companionship have kept me grounded and motivated. I hope this milestone brings you as much joy and pride as you have brought into my life. I will always strive to make you proud, as you have shaped who I am today. And to **Houeida**, my sister, who I am insanely proud of, you've taught me so much, and I always look up to you. Your perseverance, assertiveness, and drive to achieve more are a constant source of inspiration. You amaze me every day. To **Pitouti**, I love you. Wenti outi, wenti zeda, wenti zeda wenti zeda... .

I love you all.

Now that I've thanked everyone, I feel truly blessed to have you all in my life. Please know that each one of you holds a special place in my heart that will never change.

Abbreviations

ADA	American Diabetes Association
AI	Artificial Intelligence
AUC	Aera Under the Curve
BMI	Body Mass Index
BYOL	Bootstrap Your Own Latent
CNN	Convolutional Neural Network
CFS	Chronic Fatigue Syndrome
COPD	Chronic obstructive pulmonary disease
COVID-19	COronaVirus Disease of 2019
FSS	Fatigue Severity Scale
KNN	K-nearest neighbor
LR	Logistic Regression
LSTM	Long Short Time Memory
NRI	Net Reclassification Index
PAID	Problem Areas in Diabetes Questionnaire
PCA	Principal Component Analysis
PHQ-9	Patient Health
RBDSQ	REM Sleep Behavior Disorder Screening Questionnaire
RF	Random Forest
RQoL	Respiratory Quality of Life
SD	Standard Deviation
SVM	Support Vector Machine
T2D	Type 2 Diabetes
WHO	World Health Organisation

Abstract

This PhD dissertation explores the development and application of vocal biomarkers as non-invasive tools for monitoring and screening in digital health, focusing on three key areas: symptom monitoring, disease screening, and health status assessment. Leveraging data from the Predi-COVID and Colive Voice studies, this research aims to demonstrate the feasibility and effectiveness of voice analysis in clinical and remote health settings.

Objective 1 investigates the use of vocal biomarkers for **symptom monitoring**, specifically focusing on detecting and monitoring fatigue in individuals recovering from COVID-19. Fatigue is one of the most persistent and debilitating symptoms of Long COVID, affecting up to 60% of patients 12 months post-infection. Using machine learning models, this study identified specific vocal features that correlate with fatigue, achieving high accuracy (up to 86% AUC) in distinguishing between fatigued and non-fatigued patients. The results underscore the potential of integrating voice-based fatigue monitoring into telehealth applications for real-time, remote patient management.

Objective 2 addresses the potential of vocal biomarkers for **screening Type 2 Diabetes**. With nearly 50% of Type 2 Diabetes cases remaining undiagnosed globally, there is a pressing need for accessible, cost-effective, and non-invasive screening methods. This research demonstrated that vocal biomarkers could effectively differentiate between individuals with and without Type 2 Diabetes, using data-driven models that consider various voice characteristics associated with Type 2 Diabetes risk factors. The study shows promise for deploying voice-based screening tools in resource-limited settings, offering a scalable solution to identify at-risk individuals early and enable preventive healthcare strategies.

Objective 3 focuses on monitoring **Respiratory Quality of Life** using a voice-based biomarker, integrating audio features with socio-demographic and clinical data. The best-performing models, using deep learning-based audio embeddings like BYOL-A, achieved 70.34% accuracy and an AUROC of 0.77. Advanced multimodal fusion techniques, including vector cross-attention, combined different voice tasks like vowel phonation and reading, boosting performance by up to 4.2%. The multimodal approach outperformed models relying on a single data source, highlighting the value of combining voice analysis with other health data for a more comprehensive and accurate assessment of respiratory health. These findings underscore the feasibility of non-invasive Respiratory Quality of Life monitoring, with potential applications in telehealth and personalized care.

This dissertation advances the field of digital health by establishing vocal biomarkers as realistic tools for symptom monitoring, disease screening, and health status assessment. Integrating advanced machine learning and audio signal processing techniques with diverse data sources provides a foundation for developing scalable, non-invasive, and cost-effective digital health solutions. Future research should focus on expanding these findings to diverse populations, enhancing algorithm robustness, and integrating these tools into clinical workflows to improve patient outcomes and healthcare accessibility.

List of Figures

Figure Title	Page
Figure 1 . Vocal biomarker pipeline, from data collection to implementation.	18
Figure 2 . Worldwide distribution of Colive Voice participants (09/2024).	20
Figure 3 . Overview of data dimensions in Colive Voice.	20
Figure 4 . Audio signal domains.	22
Figure 5 . Short-time Fourier transform (STFT) overview.	24
Figure 6 . Typical pipeline used in audio deep learning models.	32
Figure 7 . Transfer learning workflow for embedding extraction and fine-tuning.	33
Figure 8 . Cross-validation process for parameter tuning and model evaluation.	34
Figure 9 . Digital diabetes: perspectives for diabetes prevention, management, and research.	41
Figure 10 . Mechanisms of speech production.	43
Figure 11 . Key potential applications of vocal biomarkers are explored in this PhD thesis.	47
Figure 12 . Description of 60 persisting symptoms, 12 months after the acute infection.	49
Figure 13 . PhD scope.	51
Figure 14 . General pipeline for fatigue vocal biomarker development.	54
Figure 15 . VGG19 extracted features by gender and operating system.	58
Figure 16 . Feature extraction pipeline using VGG19 on mel-spectrograms.	59
Figure 17 . T2D status classification pipeline using voice data from Colive Voice USA participants.	62
Figure 18 . T2D physiological changes that affect the production of speech.	67
Figure 19 . Workflow of RQoL monitoring using early (feature-based) and late (model-based) fusion.	72
Figure 20 . Spectrograms of sustained vowel phonation of participants matched by age and gender.	75
Figure 21 . Performance of the best model	76
Figure 22 . Accuracy with the best-performing machine learning model for socio-demographic/clinical features only, voice features only and fused clinical and voice (multimodal) features.	77

Figure 23 . Data modalities and opportunities for multimodal biomedical AI.	86
Figure 24 . Multitask learning framework for speech analysis	92

Tables

Table Title	Page
Table 1. Characteristics of the deep audio embeddings.	26
Table 2. Machine learning algorithms used across PhD objectives.	31
Table 3. Vocal task Types: strengths and challenges.	45
Table 4. Examples of disease-specific vocal tasks and corresponding audio features.	46
Table 5. Characteristics of Predi-COVID participants by device type and gender.	56
Table 6. Results of the prediction models for fatigue status classification.	57
Table 7. Demographic and clinical characteristics of participants with and without T2D, stratified by gender.	64
Table 8. Results of the prediction models for T2D status classification.	65
Table 9. Performance stratification of voice-based T2D status detection algorithms.	66
Table 10. Demographic and clinical data of participants with normal and impaired RQoL	74
Table 11. Results of the prediction models for RQoL status classification based on socio-demographic/clinical data.	75
Table 12. Performance metrics of audio modalities and fusion techniques in RQoL prediction.	77

List of Publications

First Author Publications

Elbéji A, Zhang L, Higa E, Fischer A, Despotovic V, Nazarov PV, et al. Vocal biomarker predicts fatigue in people with COVID-19: results from the prospective Predi-COVID cohort study. *BMJ Open*. 2022;12: e062463.

Elbéji A, Pizzimenti M, Aguayo G, Fischer A, Ayadi H, Mauvais-Jarvis F, Riveline J-P, Despotovic V, Fagherazzi G. "A voice-based algorithm can predict type 2 diabetes status in USA adults: Findings from the Colive Voice study." Accepted in *Plos Digital Health*.

Second Author Publications

Higa E, **Elbéji A**, Zhang L, Fischer A, Aguayo GA, Nazarov PV, et al. Discovery and Analytical Validation of a Vocal Biomarker to Monitor Anosmia and Ageusia in Patients With COVID-19: Cross-sectional Study. *JMIR Med Inform*. 2022;10: e35622.

Fischer A, **Elbeji A**, Aguayo G, Fagherazzi G. Recommendations for Successful Implementation of the Use of Vocal Biomarkers for Remote Monitoring of COVID-19 and Long COVID in Clinical Practice and Research. *Interact J Med Res*. 2022;11: e40655.

Ayadi H, **Elbéji A**, Despotovic V, Fagherazzi G. Digital Vocal Biomarker of Smoking Status Using Ecological Audio Recordings: Results from the Colive Voice Study. *Digit Biomark*. 2024;8: 159–170.

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. *medRxiv*. 2023. p. 2023.11.11.23298300.

Despotovic V, **Elbéji A**, Nazarov PV, Fagherazzi G. Multimodal Fusion for Vocal Biomarkers Using Vector Cross-Attention. *Interspeech 2024*. ISCA: ISCA; 2024. pp. 1435–1439.

Other Publications

Fagherazzi G, Zhang L, **Elbéji A**, Higa E, Despotovic V, Ollert M, et al. A voice-based biomarker for monitoring symptom resolution in adults with COVID-19: Findings from the prospective Predi-COVID cohort study. *PLOS Digit Health*. 2022;1: e0000112.

Chapter 1

Objectives

This PhD thesis focuses on the exploration and application of vocal biomarkers in digital health, particularly within the domains of symptom detection, health status monitoring, and disease screening. The goal is to contribute to the growing body of research in vocal biomarkers by addressing specific challenges in these areas. The scope of this research is defined by three main objectives, each aligning with a key application of vocal biomarkers.

Objective 1: Symptom Detection and Monitoring

The first objective is to explore the potential of vocal biomarkers for detecting and monitoring symptoms of fatigue in COroNaVirus Disease of 2019 (COVID-19) patients. Fatigue is one of the most persistent and debilitating symptoms in Long COVID, affecting a significant proportion of patients even after recovery. Given the global nature of the pandemic and the need to reduce in-person healthcare interactions, COVID-19 represents an opportunistic choice for this study. Fatigue is particularly relevant as a starting point due to its prevalence in Long COVID and the potential for using vocal biomarkers to remotely monitor its persistence or resolution. This objective focuses on understanding the underlying biological mechanisms of fatigue and developing non-invasive monitoring techniques using vocal biomarkers using data from the Predi-COVID study.

Objective 2: Disease Screening

The second objective examines the potential of vocal biomarkers as a screening tool for Type 2 Diabetes (T2D). Given the complexity of T2D and the challenges associated with its diagnosis, particularly in resource-limited settings, this objective seeks to explore how vocal biomarkers can be used as a non-invasive, cost-effective screening method. The research involves developing a voice-based tool for T2D screening using Colive Voice data, potentially offering an alternative to traditional, invasive diagnostic methods. This work aims to address the critical need for accessible screening tools that can facilitate early detection and intervention.

Objective 3: Perceived Health Status Monitoring

The third objective aims to use vocal biomarkers to monitor Respiratory Quality of Life (RQoL) in a diverse population, including those with respiratory conditions like asthma and Chronic Obstructive Pulmonary Disease (COPD), and those without. Respiratory issues are closely linked to voice, making them a logical focus for vocal biomarker research. Effective monitoring of RQoL is essential for managing conditions like COPD and asthma, yet traditional methods are often cumbersome and invasive. This objective leverages data from the Colive Voice study to develop and validate vocal biomarkers that can accurately reflect changes in RQoL, enabling continuous, remote monitoring. The goal is to enhance patient management by providing a non-invasive, reliable method for tracking respiratory health, ultimately improving patient outcomes.

Chapter 2

Materials and Methods

A vocal biomarker is a distinctive feature or set of features from the audio signal of the voice linked to a clinical outcome. It can be used for patient monitoring, diagnosing conditions, assessing disease severity or progression, and supporting drug development[1]. To develop vocal biomarkers, which are vocal characteristics associated with clinical outcomes (see section [Vocal Biomarkers](#)), four main steps are necessary, as detailed in Figure 1. These steps include data collection, data processing, data analysis, and integrating the validated vocal biomarker into devices for clinical research, epidemiology, and clinical practices. In this chapter, we will detail each step within the context of the PhD thesis projects.

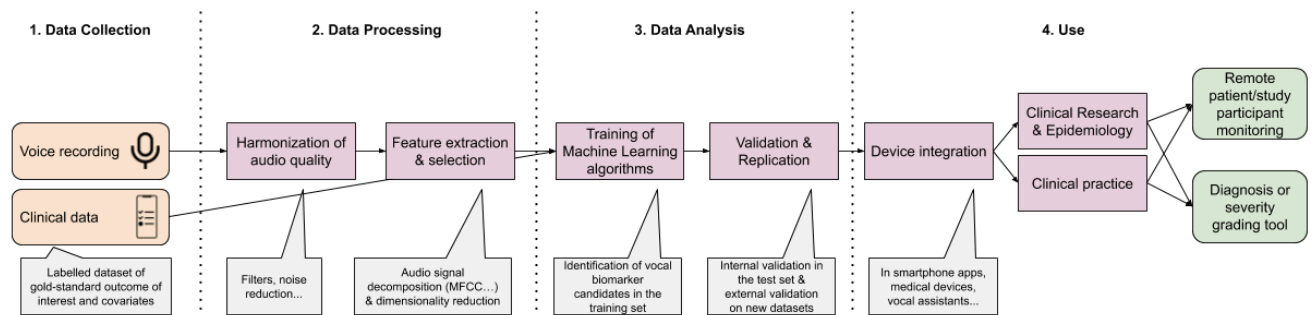


Figure 1. Vocal biomarker pipeline, from data collection to implementation (from Fagherazzi et al *Voice For Health: The Use Of Vocal Biomarkers From Research To Clinical Practice. Digital Biomarkers* 2021).

Data Sources

Two study datasets have been used in this thesis, both containing audio recordings and their corresponding clinical data annotations: Predi-COVID and Colive Voice.

Predi-COVID

The data from the Predi-COVID study was used for the first PhD objective. This analysis was conducted for the first published manuscript, entitled “Vocal Biomarker Predicts Fatigue in People with COVID-19: Results from the Prospective Predi-COVID Cohort Study”, which is detailed in [Chapter 4](#).

Study Design

Predi-COVID is a comprehensive prospective, hybrid cohort study designed to identify predictive factors for severe COVID-19 outcomes based on epidemiological, clinical, digital, and sociodemographic characteristics, as well as pathogen and host biomarkers[2]. Initiated in response to the rapid global spread of COVID-19, the study seeks to understand the variability in disease severity, particularly in Luxembourg. Participants include individuals who have tested positive for COVID-19, along with a subgroup from their households. The study aims to monitor symptoms, track long-term health consequences for up to a year, and to identify vocal biomarkers for remote patient monitoring. Enrolled participants undergo a detailed follow-up protocol, involving both digital and face-to-face assessments, leveraging mobile applications for real-time data capture, and providing biological samples at specified intervals to deepen the understanding of the disease’s progression and impact. Trial

registration number NCT04380987. Approved by the National Research Ethics Committee of Luxembourg (study number 202003/07) in April 2020.

Data Collection

The patients were tracked remotely for up to a year using the CoLive LIH smartphone app to collect voice data. To ensure a minimum quality level, participants were asked to record in a quiet environment while maintaining a certain distance from the microphone, and an audio example of the required recording was provided.

All participants in this study were invited to perform two vocal tasks:

- Type 1: Participants were asked to read the 25th article of the Declaration of Human Rights, in their preferred language among those supported (French, German, English, and Portuguese).
- Type 2: Participants were asked to hold the [a] vowel phonation without breathing for as long as they could.

Colive Voice

The data from the Colive Voice study was used for the second and third PhD objectives. This data supported the analysis for the second manuscript entitled “A Voice-Based Algorithm Can Predict Type 2 Diabetes Status in USA Adults: Findings from the Colive Voice Study”. Additionally, it was used for the third paper, where I am the second author, entitled “Digital Voice-Based Biomarker for Monitoring Respiratory Quality of Life: Findings from the Colive Voice Study.” These studies are detailed in [Chapter 5](#) and [Chapter 6](#), respectively.

Study Design

In 2021, the Luxembourg Institute of Health launched Colive Voice, a global, multilingual research initiative focused on the identification of vocal biomarkers for screening and monitoring of various chronic diseases and common health symptoms. Colive Voice operates as a comprehensive vocal biomarker screening platform, gathering audio recordings from individuals aged 15 and above worldwide, without restrictions on their health status or conditions, in languages including English, French, German, and Spanish. Participants engage in standardized vocal tasks, and the data collected is enriched with clinical and demographic annotations. The project is officially registered with ClinicalTrials.gov (NCT04848623) and received ethical approval from the National Research Ethics Committee of Luxembourg (study number 202103/01) in March 2021. All participants have given their informed consent to participate in the study.

To date (09/2024), there are a total of up to 7000 participants in Colive Voice, coming from 5 continents (Figure 2), among which 62% are females. The language distribution in Colive Voice included French (46%), English (36%), Spanish, and German, with Portuguese and Arabic being added in the second version of Colive Voice. This is one of the largest audiobanks available for health research consisting of exclusively real-life voice recordings.

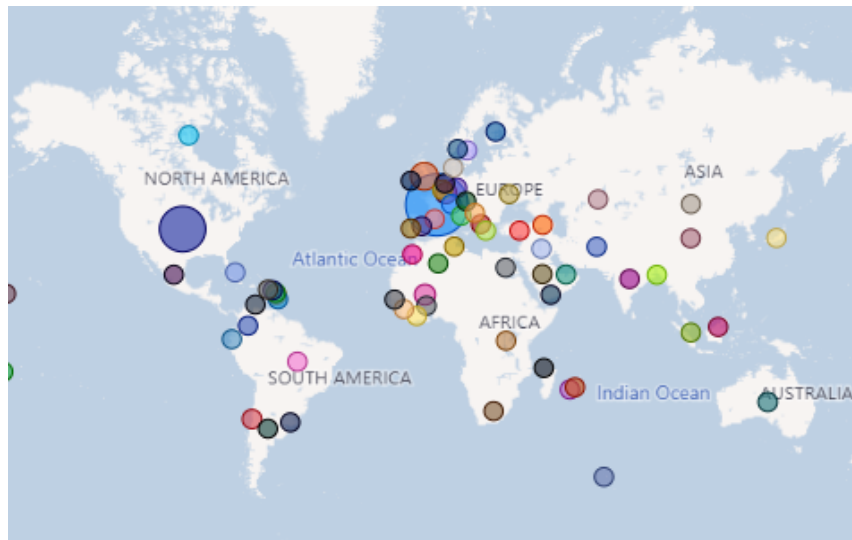


Figure 2. Worldwide distribution of Colive Voice participants (09/2024).

Data Collection

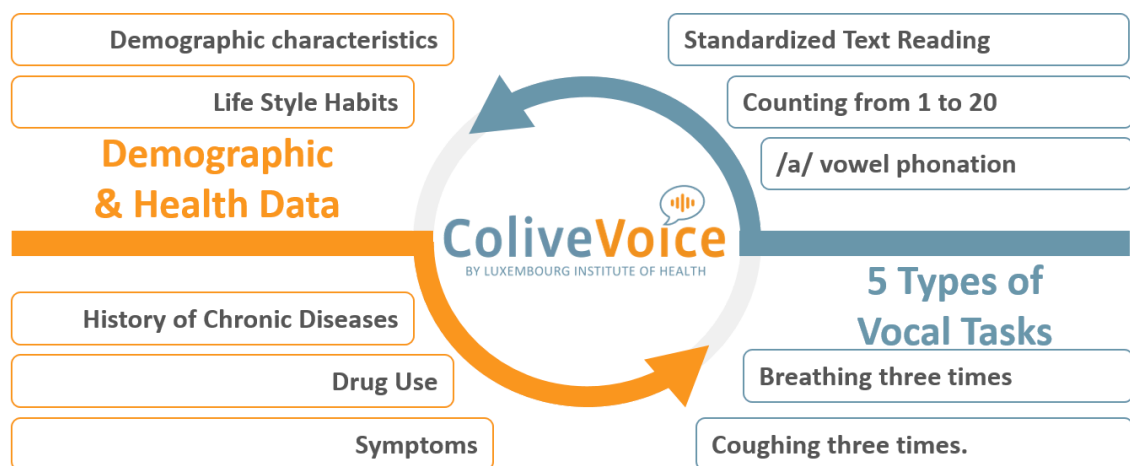


Figure 3. Overview of Data Dimensions in Colive Voice.

Colive Voice participants are invited to complete a comprehensive questionnaire to gather a diverse range of information as shown in Figure 3:

- **Demographic characteristics:** Age, gender, education level, and mother tongue.
- **Lifestyle habits:** Smoking status and alcohol intake.
- **Anthropometric data:** Body mass index (BMI).
- **Symptoms:** Information collected through questionnaires such as the Patient Health Questionnaire-9 (PHQ-9), Fatigue Severity Scale (FSS), Vocal Quality of Life Index-11 (VQ-11), REM Sleep Behavior Disorder Screening Questionnaire (RBDSQ), Problem Areas in Diabetes (PAID), and stress score.
- **Drug use:** Includes painkillers, cancer treatments, diabetes treatment, and contraceptive pills.
- **History of chronic diseases:** Includes cancer, endocrine diseases (e.g., diabetes, thyroid disorders), depression diagnosis, cardiovascular diseases (e.g., hypertension,

stroke), neurological diseases (e.g., migraine), digestive diseases, and respiratory diseases.

Data Processing

To ensure artificial intelligence (AI) readiness of the collected audio recordings in both Predi-COVID and Colive Voice studies, they undergo a series of data processing steps, required due to uncontrolled conditions for audio data acquisition. This involves data cleaning, standardization, and proper formatting for effective training and deployment of AI models. These steps address several challenges, grouped as follows:

Hardware

- Device Variations: Differences in devices and microphone quality, may introduce noise or affect recording quality.
- Hum: Low-frequency stationary noise, often caused by electromagnetic interference from electrical devices.
- Hiss: High-frequency stationary noise, typically caused by low-quality audio microphones.
- Crackling: Non-stationary noise that can arise from electrical interference, mechanical issues, or abrupt changes in amplitude.

Software

- Integrated Audio Processing: Software-driven compression, noise reduction, and automatic gain control performed on the recording device can inadvertently alter the audio quality.

Environmental

- Background Noise: Includes nature sounds (e.g., birds), urban sounds (e.g., cars), domestic sounds (e.g., TV, radio), and human activity (e.g., multiple speakers).
- Sudden Noises: Interruptions such as notification sounds, doors closing, or unexpected loud noises during recording.

Recording compliance

- Adherence to voice tasks: Not following the specified voice tasks correctly.
- Lead and trail silence and noise: Unnecessary silence and noise at the beginning and end of recordings.
- Corrupted audio files.
- Silence or noisy recordings: Recordings with only excessive silence or background noise.

A processing pipeline is implemented to harmonize and prepare the recordings for subsequent steps. This pipeline includes harmonizing audio quality across the studies through peak normalization, trimming lead-trail silence, converting stereo to mono, resampling the sample rate to 16 kHz, and performing noise reduction and quality checks (partially manual to include as many audio recordings as possible).

The pipeline is further enhanced by quality annotation of Colive Voice data, ensuring the harmonization and AI readiness of the audio data for developing vocal biomarkers. Since the technical solutions are IP protected (see section [Vocalive Platform for Vocal Biomarker Development](#) for further details), they will not be detailed in this dissertation.

Audio Signal Analysis

Sound is described as a vibration that travels through a medium (air, water, etc.) as an audible compression wave. These physical vibrations may be translated to an electrical signal by employing a transducer, such as a microphone, and displayed as a function of time. This signal is then digitized to perform different preprocessing steps and apply machine learning algorithms. The audio waves are characterized by physical characteristics such as frequency, amplitude, and direction.

A normal human can hear sound vibrations in the range of 20 Hz to 20 kHz. Signals that create such audible vibrations qualify as an audio signal[3].

Early signal processing techniques were mainly using time-domain operations such as correlation, convolution, inner product, and signal averaging. While the time-domain operations provided some information about the signal they were limited in their ability to extract the frequency content of a signal (Figure 4). The introduction of Fourier theory addressed this issue by enabling the analysis of signals in the frequency domain. However, the Fourier technique provided only the global frequency content (represented by the spectrum (Magnitude (Frequency))) of a signal and not the time occurrences of those frequencies. Hence neither time-domain nor frequency-domain analysis were sufficient enough to analyze signals with time-varying frequency content. To overcome this difficulty and to analyze the nonstationary signals effectively, techniques that could provide joint time and frequency information were needed[3].

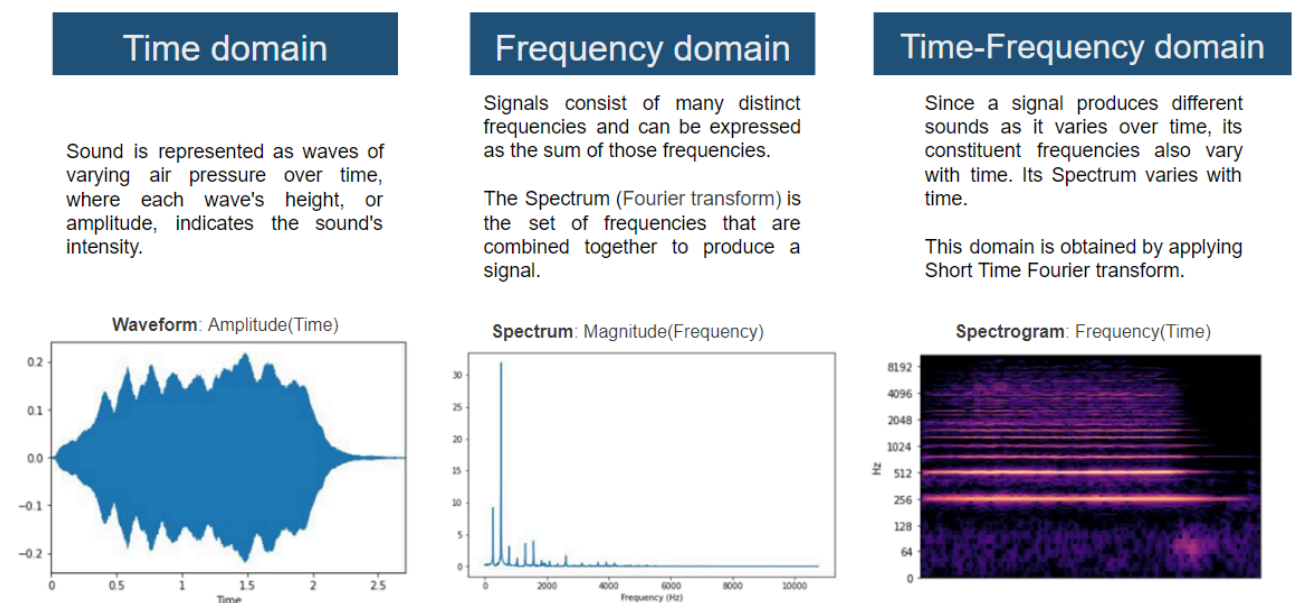


Figure 4. Audio signal domains.

Audio Features

Audio features can be broken down into four categories that describe different aspects of sound[4][5].

Source Features

Source features provide insights into the origin of voice production, specifically how air flows from the lungs through the glottis. They can directly characterize this airflow using glottal features such as Jitter, which measures frequency variations of the voice, or indirectly by capturing movements of the vocal folds through voice quality features like Shimmer, which assesses amplitude fluctuations. These measurements help in analyzing the mechanical aspects of voice production and the quality of the sound produced.

Formant (Filter) Features

Formant features focus on the resonant frequencies of the vocal and nasal tract. The features such as the first frequency (F1), second mean frequency (F2), and their variability metrics (F1 variability and F2 variability) provide a basis for identifying the vocal tract's shape and configuration during speech. These formant frequencies are crucial for vowel sounds and help in distinguishing between different speech sounds.

Spectral Features

Spectral features describe the speech spectrum, capturing the frequency distribution of the speech signal at a particular moment in a high-dimensional representation. Frequently used spectral features include Power Spectral Density (PSD) and Mel Frequency Cepstral Coefficients (MFCCs). PSD describes the distribution of the power of a signal across various frequencies, revealing the energy content at each frequency component within the signal. MFCCs, on the other hand, provide a representation of the short-term power spectrum of sound, encapsulating important characteristics of the speech's timbre.

Spectrograms

A spectrogram is a visual representation of the spectrum of frequencies in a sound or other signal as they vary with time. It is typically created by applying a Short-Time Fourier Transform[6] to small segments of the signal to determine the frequency component at each time segment (Figure 5). This results in a 2D graphical representation where one axis represents time, the other represents frequency, and the color or intensity of each point represents the amplitude of a particular frequency at a particular time.

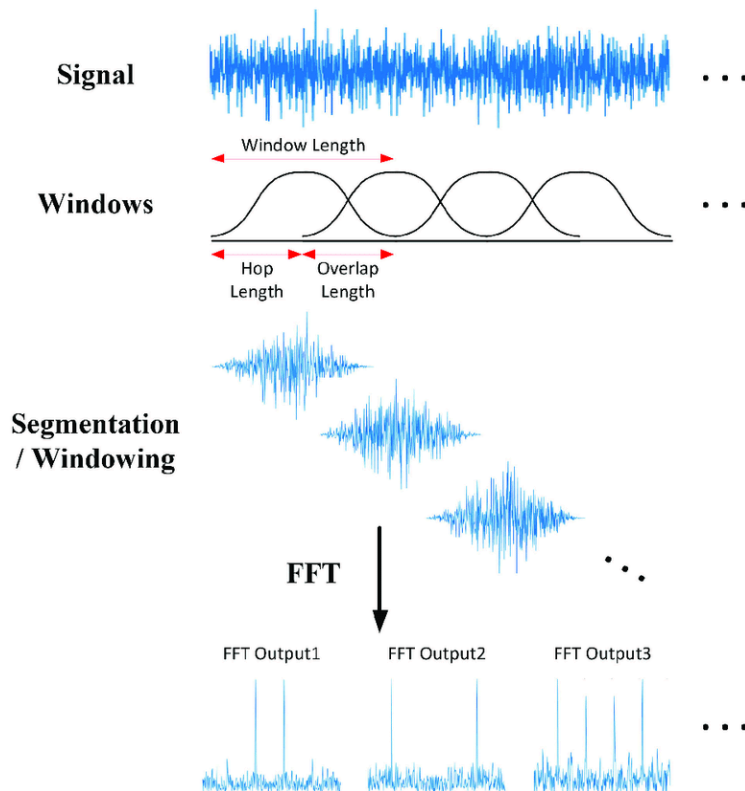


Figure 5. Short-time Fourier transforms (STFT) overview (from Jeon, H. et al. Area-Efficient Short-Time Fourier Transform Processor for Time-Frequency Analysis of Non-Stationary Signals.).

Mel-Spectrograms

A Mel-spectrogram is an enhanced version of the basic spectrogram, where the frequency scale is converted to the Mel scale, a perceptual scale that more closely aligns with human auditory responses than linear frequency scales. The Mel-spectrogram emphasizes perceptually important frequency bands and is particularly effective for speech and music processing. It is commonly used as a feature in machine learning models for speech recognition, music genre classification, and other audio analysis applications.

Prosodic Features

Prosodic features encapsulate the rhythmic and intonational aspects of speech[5]. This includes measurements like the speaking rate, the pitch (auditory perception of tone), loudness, and energy dynamics. These features help in analyzing the emotional and expressive qualities of speech, as well as its natural flow and articulation patterns.

Each category provides a unique perspective on the acoustic properties of speech, contributing to a comprehensive analysis of voice and speech disorders or variations in normal speech patterns.

Audio Feature Extraction

Once the audio recordings are processed, audio features are extracted for AI algorithms and statistical analyses. These features help describe the data within the target groups or outcomes. They can be either low/moderate-dimensional handcrafted features (see Synopsis Introduction for more details) or high-dimensional features such as audio embeddings from pre-trained models on audios or images (transfer learning). More details and definitions will be provided in the following sections.

The following feature extraction tools were tested in all the projects conducted within this PhD thesis:

Librosa

Librosa is an open-source Python library widely used for music and audio analysis[7]. It provides the tools to analyze and extract meaningful data from audio signals, particularly designed to handle music and speech. This library simplifies the process of handling audio data and extracting audio features with a high-level interface for analyzing a wide range of audio signals.

The main utilities of Librosa include:

Audio Processing: Librosa supports various functionalities for basic audio processing, such as loading audio files, resampling, and audio conversion.

Feature Extraction: It provides extensive support for extracting audio features like Mel-frequency cepstral coefficients (MFCCs), Mel-spectrograms, spectral contrast, chroma features, and zero-crossing rate, which are crucial for tasks like music genre classification or speech recognition.

OpenSMILE

OpenSMILE is a versatile, open-source software tool designed for extracting, processing, and interpreting large-scale audio features relevant in the fields of speech and music processing, emotion recognition, and beyond[8]. Developed at the Technical University of Munich, it is highly regarded for its robust feature extraction capabilities and flexibility in handling various audio analysis tasks.

The main utilities of OpenSMILE include:

Feature Extraction: OpenSMILE excels at extracting a wide array of audio features from continuous signal streams. It can process both real-time and recorded audio, making it suitable for many applications, from interactive systems to batch analysis of large datasets.

Configurable Architecture: It offers a configurable architecture that allows users to define and customize feature sets through configuration files, making it adaptable to specific research needs and experimental settings.

Two main feature sets provided by OpenSMILE are ComParE 2016 and eGeMAPSv02:

ComParE 2016: This feature set was introduced for the Computational Paralinguistics Challenge (ComParE) at the Interspeech 2016 conference[9]. It includes 6373 attributes per second, encompassing a broad spectrum of signal descriptors, including energy, spectral, cepstral, and voice quality features. This extensive set is designed to tackle a wide range of computational paralinguistics tasks such as emotion, and sentiment analysis.

eGeMAPSv02: The extended Geneva Minimalistic Acoustic Parameter Set version 0.2 (eGeMAPSv02) focuses on a selected set of 88 voice parameters that have proven useful in affective computing[10]. This set is streamlined to facilitate effective voice research and is often used in emotion recognition studies. It includes frequency-based features, energy, and cepstral parameters designed to capture the emotional state conveyed by speech.

Pretrained Algorithms

1. Audio Embeddings

Audio embeddings are compact, high-dimensional representations of audio signals that capture essential characteristics of the sound, such as pitch, timbre, and rhythm, in a form suitable for machine learning models. These embeddings are typically generated by deep learning models (Table 1), which process raw audio data and transform it into a numerical representation that retains the most important features while reducing the complexity of the original signal.

Embeddings	Pre-training	Encoder	Input	Output
Representation	Dataset	Architecture	Sr (kHz)	Embedding
VGGish	YouTube-8M	VGG based	16	128-d/frame
YAMNet	YouTube-8M	Mobilenet_v1	16	128-d/frame
OpenL3	AudioSet (envir.)	Audio sub network of L3 -Net	48	6.144-d/frame
BYOL-A	AudioSet	CNN	16	1.024-d/frame
BYOL-S	AudioSet-Speech	CNN	16	1.024-d/frame
Wav-2-Vec	Multiple sets	CNN	16	Config dependent

Table 1. Characteristics of the deep audio embeddings.

VGGish is a convolutional neural network model that transforms audio into a compact embedding. It is trained on a large-scale dataset derived from YouTube videos[11]. The model is based on the VGG architecture, which was originally developed for image recognition tasks but has been adapted for audio. These embeddings capture the salient features of the audio and are often used as input features for other machine learning tasks, such as audio classification and event detection.

YAMNet is a deep convolutional neural network that predicts audio events from a wide range of classes[12]. It is based on the MobileNet_v1 architecture, making it efficient for mobile

environments. The model is pre-trained on the YouTube-8M dataset and can recognize 521 audio event classes, making it useful for applications that require audio context recognition in real time.

OpenL3 is an audio embedding model that is part of the L3-Net family, specifically designed to provide an auditory version of the successful image-based Look, Listen, and Learn architecture[13]. It is trained on the AudioSet dataset focused on environmental sounds. OpenL3 processes audio signals into embeddings that can be used for various tasks, such as audio classification, similarity comparison, and clustering.

BYOL-A (Bootstrap Your Own Latent for Audio) is a model that applies the self-supervised learning principles of the original BYOL architecture to the audio domain[14]. It uses a convolutional neural network to generate embeddings without needing labeled data. BYOL-A is pre-trained on the AudioSet and learns powerful, generalizable audio representations that are useful for a broad range of audio analysis tasks.

BYOL-S adapts the BYOL approach specifically to the domain of speech processing[15]. Like BYOL-A, it leverages a convolutional neural network but is trained on a subset of the AudioSet that focuses specifically on speech-related sounds. This specialization allows BYOL-S to capture nuanced features in speech that are crucial for applications in speech recognition and related tasks.

Wav2Vec is a framework designed for speech processing that learns robust representations of raw audio by directly modeling the raw waveform, rather than traditional handcrafted features[16]. Wav2Vec uses a convolutional neural network and is pre-trained on multiple large-scale speech datasets. It operates by predicting future audio samples from the past context, thereby learning features that are useful for speech recognition tasks. The flexibility of its configuration makes it adaptable to a wide range of speech and audio processing tasks.

2. Image Embeddings

An innovative application of Mel-spectrograms as inputs for image-pretrained algorithms is explored. By treating Mel-spectrograms as analogous to natural images, the pre-trained network's robust feature-extraction capabilities are harnessed. This methodology not only enhances the efficiency of the analysis by using pre-existing, well-optimized models but also offers a cross-domain adaptability that is particularly beneficial in scenarios where audio data is complex and labeled datasets are limited. Thus, this technique could improve the accuracy and effectiveness of audio analysis in the research, providing a strong example of how machine learning tools can be adapted across different domains to achieve superior analytical outcomes.

VGG19 is a convolutional neural network architecture designed for image classification, consisting of 19 layers, including 16 convolutional layers and 3 fully connected layers[17]. It uses small 3x3 convolutional filters and max-pooling layers to capture detailed features in the input data. Pre-trained on the large image dataset ImageNet, VGG19 is widely used in transfer learning applications. In voice analysis, VGG19 can extract deep audio features by converting audio signals into spectrograms, making it valuable for tasks such as voice biomarker identification and health status prediction.

Statistical Analyses

To further describe the clinical data and the extracted audio features, and to analyze the voice characteristics of the target groups, various statistical analyses were conducted. This section details the most commonly used statistical analyses in this dissertation.

Chi-2

The Chi-Square test is a statistical method used to determine if there is a significant association between categorical variables[18]. It compares the observed frequencies in each category to the expected frequencies if there are no associations between the variables. The test calculates a chi-square statistic, which follows a chi-square distribution, allowing for the determination of the p-value to assess the significance of the observed differences.

Mann U Whitney

The Mann-Whitney U test, also known as the Wilcoxon rank-sum test, is a non-parametric test used to compare differences between two independent groups[21]. It assesses whether the distributions of the two groups differ significantly by ranking all the values from both groups together and then comparing the sum of ranks between the groups. It is often used as an alternative to the t-test when the assumption of normality is not met.

One-way ANOVA

One-way Analysis of Variance (ANOVA) is a statistical test used to compare the means of three or more independent groups to determine if there is a significant difference among them[22]. It evaluates the variance within each group and the variance between the groups to calculate an F-statistic. If the F-statistic is significantly larger than expected by chance, it indicates that at least one group's mean is different from the others. Post-hoc tests are typically used to identify which specific groups differ.

Bland-Altman

The Bland-Altman plot, also known as the Tukey mean-difference plot, is a graphical method to assess the agreement between two quantitative measurements[23]. It plots the difference between the two measurements against their average. The plot includes lines for the mean difference and limits of agreement (typically set at ± 1.96 standard deviations from the mean difference), allowing for visual assessment of bias and variability. It is widely used in method comparison studies to evaluate the consistency between two measurement techniques.

Feature Scaling and Normalization

Feature scaling and normalization are crucial preprocessing steps in machine learning. They ensure that features contribute equally, improving algorithm performance and convergence, particularly for gradient descent-based models. Scaling helps prevent features with larger ranges from dominating the model, reducing bias, and enhancing model accuracy. Additionally, techniques like Principal Components Analysis (PCA) perform better with scaled data, as they are sensitive to the scale of the input features. Overall, scaling and normalization lead to more stable, efficient, and interpretable models.

StandardScaler Sklearn

StandardScaler (Sklearn) is a tool that standardizes features by removing the mean and scaling to unit variance, transforming data so each feature has a mean of zero and a standard deviation of one[24]. This is useful for machine learning algorithms sensitive to feature scales, such as logistic regression and PCA, ensuring consistent variance across all features.

Dimensionality Reduction

For high-dimensional audio features or embeddings, dimensionality reduction is essential to avoid overfitting, feature collinearity, and poor generalization. To achieve this, techniques such as feature selection (e.g., SelectKBest) or PCA are conducted, depending on the data type and the importance of maintaining interpretability and explainability.

Feature Selection

SelectKBest (Sklearn) is a feature selection tool that selects the top k features based on their importance to the outcome variable[25]. This method uses statistical tests to score each feature and select the highest-scoring ones. It is primarily used for handcrafted features to enhance interpretability and explainability in the model, making it easier to understand which features are most influential in predicting the target variable. This approach is especially useful when dealing with a large number of features, allowing for more efficient and effective modeling.

Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space while retaining most of the original variability[26]. It does this by identifying the principal components, which are the directions in the data that capture the most variance. PCA is mainly used for embeddings, as they are high-dimensional vectors with less explainability. By reducing the number of dimensions, PCA helps to avoid overfitting, improve generalization, and simplify the model without losing significant information.

Key AI Concepts

For vocal biomarker identification and development, we train AI algorithms using extracted, scaled, and selected features or reduced embeddings. This involves several approaches:

- **Machine Learning with Handcrafted Features:** We feed machine learning algorithms with extracted handcrafted audio features for classification or regression tasks. These features capture specific aspects of the audio signal, such as pitch, tone, and rhythm, which are relevant for identifying vocal biomarkers.
- **Deep Learning with Mel-Spectrograms:** Mel-spectrograms, which represent the audio signal in both time and frequency domains, are used to train deep learning models. These visual representations of sound allow convolutional neural networks (CNNs) to learn complex patterns and features that are crucial for accurate biomarker identification.

- **Transfer Learning:** This technique leverages pre-trained algorithms to incorporate cross-domain knowledge, enhancing the model's performance with limited data. There are two main strategies:
 - **Feature Extraction:** Feeding extracted audio or image embeddings from pre-trained models into machine learning or deep learning algorithms. This allows the use of rich, high-level features that have been learned from large datasets.
 - **Fine-tuning:** Adjusting the parameters of the pre-trained models by retraining them on our specific dataset to improve their performance and adapt them to the nuances of our audio data. This method is particularly effective when domain-specific data is limited but the pre-trained models have been trained on large, diverse datasets.

Machine Learning

Machine learning is a discipline within AI that emphasizes the creation of algorithms capable of modifying their parameters and improving their performance based on empirical data[27]. This field incorporates a variety of statistical, probabilistic, and optimization techniques that enable computers to learn from past experiences or historical data without explicit programming for each task.

The machine learning algorithms used in this research, as shown in Table 2, were implemented using the Scikit-learn (Sklearn) library[28], which provides a robust and efficient framework for model development and evaluation.

Machine Learning Algorithm	Characteristics	PhD Objectives
Logistic Regression [29]	A linear classifier that models the probability of a binary outcome using the logistic function is often used for its simplicity and interpretability.	1, 2, 3
Support Vector Machine [30]	A supervised learning model that constructs hyperplanes in a multidimensional space to separate different classes, is particularly effective in high-dimensional spaces.	1, 2, 3
Random Forest [31]	An ensemble learning method that builds a multitude of decision trees during training and outputs the mode of the classes or mean prediction of the individual trees, enhancing generalization.	1, 2, 3
K-Nearest Neighbors [32]	A non-parametric, instance-based learning algorithm that classifies a sample based on the majority class among its k nearest neighbors in the feature space.	1
Voting Classifier [33]	An ensemble model that combines the predictions of multiple base classifiers to improve overall performance, using either hard voting (majority rule) or soft voting (average probabilities).	1
XGboost [34]	An advanced gradient boosting algorithm that builds sequential decision trees, optimizing performance through regularization techniques and parallel processing.	3

Table 2. Machine learning algorithms used across PhD objectives.

Deep learning

Deep learning is a sophisticated subset of machine learning characterized by networks with multiple layers, known as deep neural networks[35]. These networks are capable of learning from vast amounts of unstructured data through the use of multiple levels of abstraction. Deep learning methodologies are integral to cutting-edge applications in areas such as computer vision, speech recognition, and natural language processing.

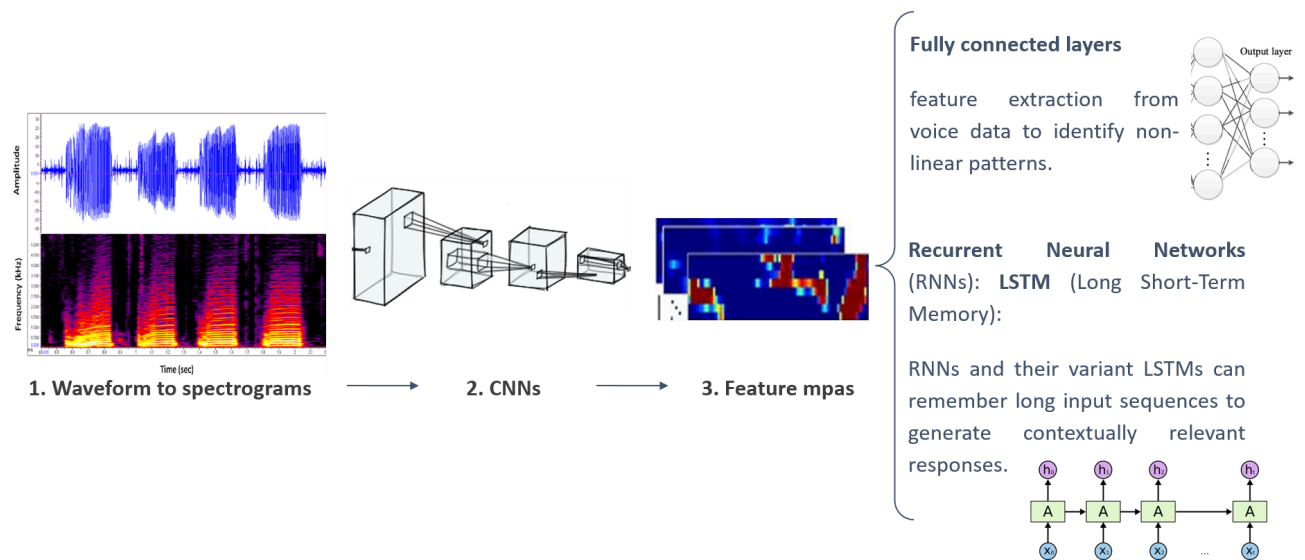


Figure 6. Typical pipeline used in audio deep learning models.

Figure 6 illustrates a typical pipeline used in audio deep learning models, where spectrograms are employed to represent audio signals. The process includes the following steps:

1. **Waveform to Spectrograms:** Raw audio waveforms are converted into spectrograms, which visually represent the frequency spectrum of the audio signal over time.
2. **Convolutional Neural Networks (CNNs):** The spectrograms are then fed into CNNs to extract high-level features. CNNs are effective in capturing local patterns in the spectrograms, such as pitch and tone.
3. **Feature Maps:** The CNNs produce feature maps that highlight important aspects of the audio signal. These maps can be further processed using different neural network architectures:
 - **Recurrent Neural Networks (RNNs) and LSTM (Long Short-Term Memory):** The feature maps are passed through RNNs or their variant LSTMs. These networks are capable of remembering long input sequences and generating contextually relevant responses, making them suitable for capturing temporal dependencies in audio data.
 - **Fully Connected Layers:** Alternatively, the feature maps can be fed into fully connected layers to identify non-linear patterns in the data and make final predictions or classifications based on the extracted audio features.

This pipeline demonstrates the combination of CNNs for spatial feature extraction and either RNNs/LSTMs for temporal sequence modeling or fully connected layers for pattern recognition, providing a comprehensive approach to audio analysis in deep learning applications.

Transfer Learning

Transfer learning is a powerful strategy in machine learning that involves applying knowledge gained from one task (Task A) to a different but related task (Task B). As depicted in Figure 7, a pre-trained network, which has learned general features from a large dataset in Task A, is reused to either extract embeddings or be fine-tuned for the new task. This approach accelerates the development of models for Task B by leveraging the learned representations, reducing the need for extensive new data and computational resources. The pre-trained model can either be fine-tuned or used to generate feature embeddings for the new data, resulting in improved performance and faster convergence, especially in deep learning applications.

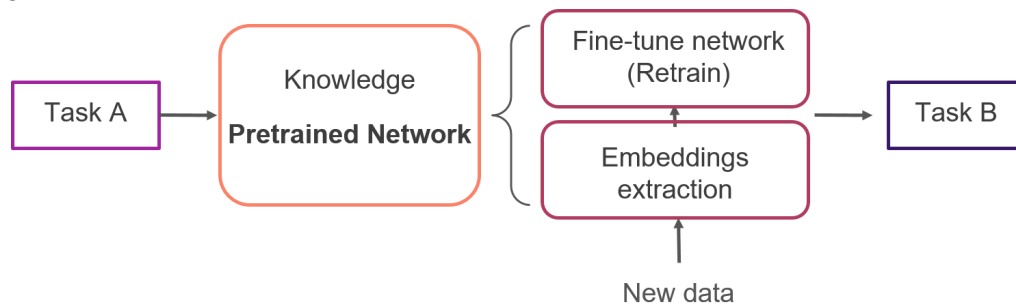


Figure 7. Transfer learning workflow for embedding extraction and fine-tuning.

Performance Evaluation

Cross-Validation

Cross-validation is a statistical method used to assess a predictive model's generalizability and avoid overfitting[36]. The process involves partitioning the dataset into 'k' folds, often 5 or 10, training the model on 'k-1' folds, and testing it on the remaining fold. This cycle is repeated 'k' times, with each fold used once as the test set. Performance metrics are calculated for each iteration and averaged to provide an overall performance measure (Figure 8). Stratified cross-validation, a variant of this method, ensures that each fold has a representative distribution of the target variable, enhancing the reliability of the results[37]. The latter approach has been adopted in this PhD thesis.

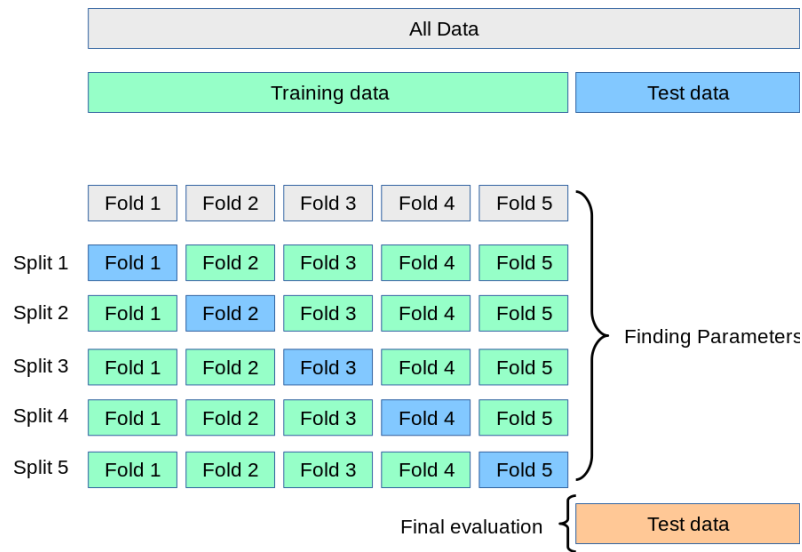


Figure 8. Cross-validation process for parameter tuning and model evaluation (from ScikitLearn).

Evaluation Metrics

The projects related to this PhD thesis focused on binary classification tasks. The performance of the developed algorithms was evaluated using all the listed metrics for each specific task or project.

To evaluate these classification models, several metrics derived from the confusion matrix are commonly used. The confusion matrix itself is a table used to describe the performance of a classification model on a set of test data for which the true values are known. It is organized into four different outcomes of predictions:

True Positives (TP): Correctly predicted positive observations.

False Positives (FP): Incorrectly predicted as positive.

True Negatives (TN): Correctly predicted negative observations.

False Negatives (FN): Incorrectly predicted as negative.

1. Accuracy

Accuracy measures the overall correctness of the model and is defined as the ratio of correctly predicted observations to the total observations.

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

2. Precision

Precision (or Positive Predictive Value) measures the accuracy of positive predictions. It is defined as the ratio of correctly predicted positive observations to the total predicted positives.

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

3. Recall (Sensitivity or True Positive Rate (TPR))

Recall measures the ability of a model to find all the relevant cases (all actual positives). It is defined as the ratio of correctly predicted positive observations to all observations in actual class.

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

4. Specificity (True Negative Rate)

Specificity measures the proportion of actual negatives that are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition).

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

5. F1 score

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. It is particularly useful when the class distribution is uneven.

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

6. Area Under the Receiver Operating Characteristic Curve (AUC - ROC)

The ROC curve is a graphical representation of a classifier's performance across all classification thresholds, plotting the True Positive Rate (Recall) against the False Positive Rate (1 - Specificity) at various threshold settings. The AUC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1) and provides an aggregate measure of performance across all possible classification thresholds.

$$AUC = \int_0^1 TPR(x) dx$$

7. Brier score

The Brier Score is a metric used to evaluate the accuracy of probabilistic predictions and the calibration of the algorithms. It measures the mean squared difference between the predicted probability and the actual outcome. The score ranges from 0 to 1, where a lower Brier Score indicates better model performance and more accurate calibration of the predicted probabilities.

Essential Python libraries used

Pandas

Pandas is an open-source Python library for data manipulation and analysis[38]. It provides data structures like DataFrames and Series for efficient handling of structured data. Pandas supports operations such as merging, reshaping, selecting, and handling missing data. Widely used in both academic and commercial settings, pandas excels in fast, efficient operations on large datasets. In this PhD thesis, our structured datasets from Predi-COVID and Colive Voice are manipulated using DataFrames with Pandas, enabling efficient data processing.

NumPy

NumPy, short for Numerical Python, is essential for scientific computing in Python, providing support for large, multi-dimensional arrays and matrices, along with numerous mathematical functions to operate on these arrays[39]. Its performance is enhanced by efficient array operations implemented in C. NumPy that underpins many other Python scientific libraries, such as Pandas, making it indispensable for numerical computations. In this PhD thesis, the audio signals imported from the Librosa library, the audio embeddings, and the inputs for tools like OpenSMILE are all handled as NumPy arrays, ensuring efficient numerical processing and manipulation.

Scikit-learn

Scikit-learn, often abbreviated as Sklearn, is a Python module that integrates a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems[40]. Built upon NumPy, it provides tools for data mining, analysis, classification, regression, clustering, model selection, and preprocessing. In this PhD thesis, Scikit-learn has been used extensively for developing and evaluating machine learning models, enabling effective model selection, data preprocessing, fine-tuning algorithms, implementing classification algorithms, and conducting cross-validation to ensure the accuracy and reliability of the developed algorithms.

TensorFlow

TensorFlow, developed by the Google Brain team, is an open-source library for high-performance numerical computation[41]. Widely used for machine learning and deep learning applications, it is known for its flexibility and extensive capabilities. TensorFlow supports both CPU and GPU devices and is used in various applications, from beginner tutorials to large-scale commercial and research projects.

Keras

Keras is a high-level neural networks API written in Python that runs on top of TensorFlow[42]. It enables fast experimentation and prototyping through its modularity, minimalism, and extensibility. Keras is popular for its user-friendliness and is suitable for building various neural network models. Specifically, the pre-trained algorithm VGG19, used in this PhD thesis, is part of the Keras API.

PyTorch

PyTorch, primarily developed by Facebook's AI Research lab, is widely used for applications such as computer vision and natural language processing[43]. Known for its flexibility and speed, especially in research and prototype development, PyTorch offers dynamic computational graphing, providing high flexibility and intuitive operations for deep learning tasks. It is favored in the academic and research sectors due to its ease of use and seamless handling of tensor operations with GPU acceleration. In this PhD thesis, the pre-trained algorithms on audio data were primarily based on PyTorch, leveraging its robust capabilities for deep learning.

IT Environment and hardware

The work for the related PhD projects was conducted locally on a DELL Precision 5550. This workstation is equipped with an Intel Core i7-10850H CPU, 32GB of RAM, and a 1TB SSD, providing robust performance for data processing and model training tasks. Additionally, some computationally intensive tasks were performed on the GPU servers from the Luxembourg Institute of Health (LIH), equipped with 8 GPUs (NVIDIA RTX A6000 with 48 GB GDDR6 memory), 80 CPUs (Intel Xeon Gold 5218R @ 2.10GHz), and total RAM of 500 GB, enhancing the speed and efficiency of deep learning model training.

Chapter 3

Synopsis - General Introduction

The following synopsis provides a comprehensive overview of vocal biomarkers, illustrating how these voice technologies, along with the methods for training and implementing them, are integrated into the broader field of Digital Health. It highlights the development and use of digital biomarkers, the role of Voice AI, and their significant implications in healthcare.

Digital Health

Digital health is the field of knowledge and practice associated with the development and use of digital technologies to improve health[44,45]. It expands the concept of eHealth to include digital technology and encompasses a wider range of smart devices and connected equipment. Commonly understood areas of digital health include AI, big data, blockchain, health data, health information systems, the infodemic, the Internet of Things (IoT), interoperability, and telemedicine.

The COVID-19 pandemic has created an urgent need for coordinated mechanisms to respond across health sectors. Digital health solutions have been identified as promising approaches to address this challenge[46]. By automating or transferring actions that patients can perform on their own, digital health and AI-based technologies have the potential to relieve clinicians. Using telemonitoring solutions to enable self-surveillance and remote monitoring of symptoms might therefore help to improve and personalize care delivery.

Digital health represents a pivotal evolution in the integration of information and communications technologies (ICT) within healthcare, extending the scope of eHealth to encompass a broader array of digital technologies and consumer involvement[47]. This field includes advanced tools such as AI, big data, blockchain, and the IoT, all aimed at enhancing the delivery, access, and efficiency of healthcare services. The emergence of digital health has been significantly influenced by seminal contributions from early pioneers such as Frank, who introduced the foundational concepts based on internet capabilities[48], and Eysenbach, who broadened the definition of digital health (or “eHealth” as he termed it) to emphasize networked healthcare delivery and the importance of adapting mindsets within healthcare frameworks[49].

A primary benefit of digital health is the empowerment of patients through technologies that facilitate enhanced self-management and active participation in their healthcare. This is supported by digital platforms like health portals, mobile apps, and social media. Additionally, the capability for remote monitoring and telehealth through wearable devices such as smartwatches, smart socks, and continuous glucose monitors, along with other tools like artificial pancreas systems, smart blood pressure monitors, and smart ECGs, exemplifies how digital health can extend care beyond traditional settings (Figure 9). These devices enable patients to manage their conditions from home while still receiving personalized care[50–52] (Figure 9).

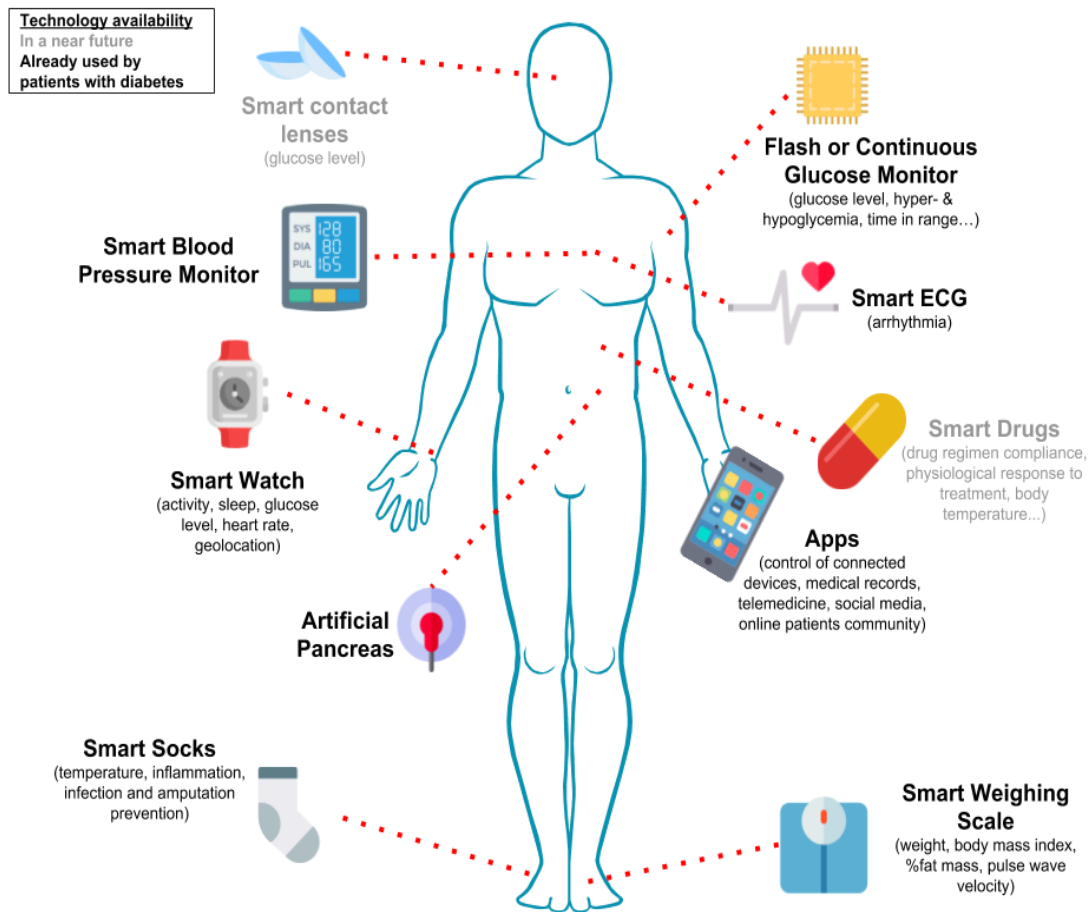


Figure 9. Digital diabetes: perspectives for diabetes prevention, management and research. (from Fagherazzi et al. ([Diabetes & Metabolism, 2018](#))).

The application of large-scale data analytics and AI in digital health supports personalized medicine and improves clinical decision-making through enhanced risk stratification and diagnostic accuracy[51]. Digital health also addresses barriers to healthcare access, including socioeconomic, disability, and language barriers, through mechanisms such as remote consultations and automated translation services[51]. Efficiency gains are remarkable as well, with digital health interventions reducing unnecessary healthcare visits and associated costs, thereby streamlining healthcare delivery[51].

However, the rapid advancement and implementation of digital health technologies, especially during the COVID-19 pandemic, underscore the urgency for effective integration and coordinated response mechanisms to health crises[46,51]. Challenges such as the thoughtful assessment of new technologies, the integration of comprehensive care records, and the avoidance of increased clinician workload remain critical considerations[53]. Therefore, the trajectory of digital health demands a meticulous approach to address these technical, organizational, and patient-centered factors to fully realize its potential in transforming healthcare delivery.

Digital Biomarkers

As defined by the European Medicines Agency (EMA), digital biomarkers are quantifiable, objective measures derived from digital devices that reflect physiological and/or behavioral states, offering insights into biological or pathological processes, or responses to interventions. This conceptualization is supported by regulatory bodies such as the Food and Drug Administration (FDA), aligning with traditional biomarker definitions as indicators of biological processes[54].

The utility of digital biomarkers spans various clinical and research applications. According to the Biomarkers, EndpointS, and other Tools (BEST) Resource by the FDA [54], biomarkers can be categorized into seven primary types:

1. **Susceptibility/Risk Biomarkers:** predict the likelihood of developing a condition in asymptomatic individuals.
2. **Predictive Biomarkers:** forecast a patient's response to a specific therapeutic intervention.
3. **Diagnostic Biomarkers:** detect or verify the presence of a disease or its specific subtypes.
4. **Prognostic Biomarkers:** indicate the probable progression of a disease, its recurrence, or the occurrence of a clinical event in diagnosed patients.
5. **Pharmacodynamic/Response Biomarkers:** measure the alteration in levels of these biomarkers in response to a therapeutic intervention to assess the biological impact of the treatment.
6. **Monitoring Biomarkers:** monitor the state of a disease or exposure to certain substances.
7. **Safety Biomarkers:** indicate the potential adverse effects of a medical intervention.

As digital health evolves, various innovative technologies, including voice AI, are being explored for their potential applications and benefits in healthcare.

Voice AI

As the primary means of communication, voice plays an important role in daily life. Voice also conveys personal information such as social status, personal traits, and the emotional state of the speaker[55]. Voice AI refers to AI systems that understand and generate human speech. These technologies leverage machine learning to interpret spoken language and respond or take action based on that input.

Voice AI applications are increasingly diverse, including:

- **Speech recognition:** Converts spoken language into text.
- **Speech synthesis:** Generates spoken language from text to produce natural-sounding speech.
- **Natural language understanding:** Interprets the meaning behind spoken inputs to respond appropriately.
- **Voice cloning:** Creates a digital replica of a person's voice from a small sample.
- **Emotion recognition:** Analyze vocal patterns and detect emotions in a speaker's voice.

- **Vocal biomarkers:** Identify voice-based features linked to health conditions for monitoring and diagnostic purposes.

Voice in Digital Health

Within the scope of digital health, voice is a promising source of digital data since it is rich, user-friendly, inexpensive to collect, and non-invasive. It may be used to remotely monitor health-related problems when integrated into innovative telemonitoring or telemedicine technologies. The complexity of neural processing involved in speech production makes speech sensitive to slight changes in the physiological condition and pathophysiological state of a speaker[56].

Mechanisms of Voice Production

The first step to producing voice is creating an airflow that comes from the contraction of the lungs. When the pressure on the vocal fold exceeds a certain threshold, the vocal folds are going to enter into a self-sustained vibration. These oscillations change the density of air passing through the vocal folds over time, which creates a wave. This wave is then modulated by the shape and the position of the mouth and the tongue, creating the voice[56]. The sound produced by vocal fold vibration is modulated in the vocal tract, which includes key speech articulators such as the velum, tongue, lips, and lower jaw. Finally, speech sounds are radiated from the lips[57].

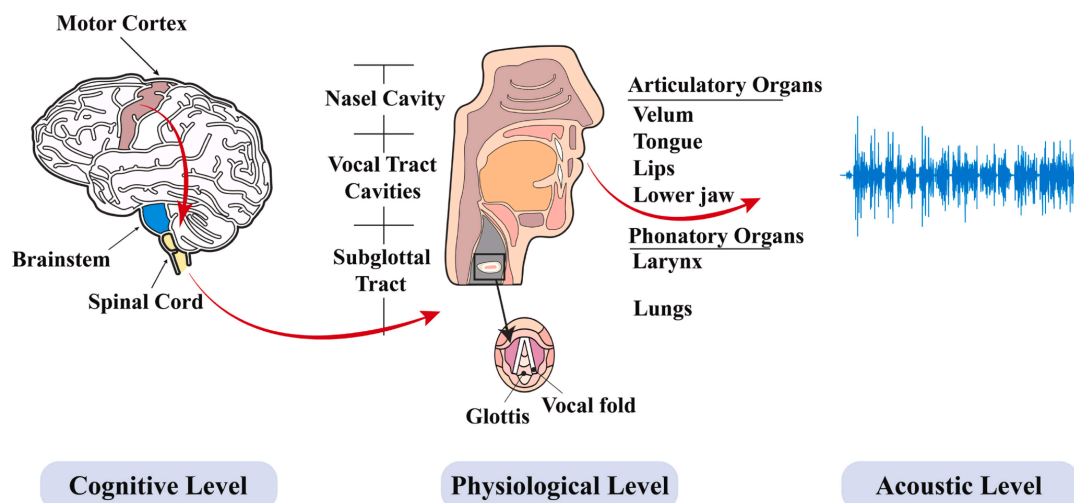


Figure 10. Mechanisms of speech production (from Almaghrabi, et al. Bio-acoustic features of depression: A review. Biomedical Signal Processing and Control).

As illustrated in Figure 10, speech production operates on three main levels: cognitive planning, physiological (muscle actions), and acoustic (sound generation). The vocalization plan originates in the brain, and motor nuclei in the brainstem and spinal cord coordinate the necessary muscle movements. The resulting vibrations from the vocal folds are shaped by the vocal tract and articulators, and the final speech sounds are emitted through the lips.

Vocal Biomarkers

A vocal biomarker is a signature, a feature, or a combination of features from the audio signal of the voice that is associated with a clinical outcome and can be used to monitor patients, diagnose a condition, grade the severity or stages of a disease, or for drug's development[58]. Current research on vocal biomarkers explores a wide range of applications, but the focus has primarily been on neurodegenerative diseases[59] and mental health[60].

Vocal Biomarkers as New Digital Clinical Endpoints

Vocal biomarkers, derived from the subtle nuances in voice, represent a powerful yet underused resource in healthcare, with increasing efforts to integrate them into clinical practice.

The push to incorporate voice biomarkers into routine clinical workflows is driven by their potential to enhance the reliability of diagnostics, predict health outcomes, and personalize patient care. This integration ensures that voice AI technologies are accessible, ethically sourced, and clinically valid. Additionally, voice biomarkers are gaining traction as digital endpoints in clinical trials, where they are used to assess the efficacy of medical treatments or interventions over time. By systematically collecting and analyzing voice data, researchers can track patient outcomes and responses to therapies, offering a non-invasive, continuous monitoring method that reduces participant burden and expands access to diverse populations.

Despite challenges related to standardization and data privacy, the adoption of voice biomarkers as digital clinical endpoints promises to revolutionize healthcare, making it more personalized, accessible, and effective.

Applications of Vocal Biomarkers

Symptom detection and monitoring: Voice biomarkers are increasingly used to detect subtle changes that may indicate underlying health conditions. These can include monitoring symptoms of chronic diseases such as Parkinson's or detecting signs of respiratory infections.

Diagnosis: Beyond symptom monitoring, voice biomarkers can aid in the diagnosis of conditions by identifying specific vocal signatures associated with different diseases, such as cardiovascular or neurological disorders.

Screening: Within preventive medicine, voice analysis can be employed to screen for potential health issues before they become apparent, aiding in early intervention strategies.

Perceived health status monitoring: This application is tailored for individual health management, where people can use voice analysis tools to monitor their perceived general health status, tracking changes that might indicate health improvements or deteriorations.

Disease progression/severity monitoring: Voice biomarkers can help monitor disease progression and severity by regularly analyzing voice data. This enables healthcare providers to track how a disease evolves, adjust treatment plans, and intervene at critical points.

Treatment effectiveness: Voice biomarkers can evaluate treatment effectiveness by analyzing changes in voice patterns before, during, and after therapy. This real-time feedback allows for more personalized treatment plans and ensures effective interventions.

Main vocal tasks and examples of disease-specific vocal biomarkers

Vocal biomarkers are assessed through various vocal tasks categorized into verbal, vowel/syllable, and nonverbal vocalizations[58]. Each type of vocal task can reveal different audio features relevant to specific diseases, but also some challenges to account for (Table 3).

- **Verbal Tasks:** These include isolated words, short sentence repetition, reading passages, and running speech. These tasks are essential for evaluating how well the vocal system can handle complex speech tasks and can reveal issues with articulation, fluency, and other speech characteristics.
- **Vowel/Syllable Tasks:** These tasks involve sustained vowel phonation and diadochokinetic tasks (rapid repetition of syllables). They assess the stability and control of vocal fold vibration, respiratory support, and neuromuscular coordination.
- **Nonverbal Vocalizations:** Tasks such as coughing and breathing provide insight into the respiratory system's function and the coordination between respiration and phonation.

Vocal Task Type	Strengths	Challenges
Verbal Tasks	Detailed insights into articulation, fluency, and speech patterns; useful for cognitive and neurological assessments	Susceptible to linguistic and cultural variability; higher cognitive load for some patients
Vowel/Syllable Tasks	Assesses vocal fold vibration stability, respiratory support, and neuromuscular coordination; useful for detecting voice disorders	Lacks complexity of natural speech, limiting ability to detect subtle cognitive-related issues
Nonverbal Vocalizations	Insight into respiratory function; easy to collect; relevant for respiratory conditions (COPD, asthma)	Provides limited information on higher-level speech or cognitive functions

Table 3. Vocal task types: strengths and challenges.

Different diseases manifest distinct vocal features, which can be better-captured thanks to specific vocal tasks. Table 4 below shows some examples of diseases, the associated vocal tasks, and the relevant audio features:

Disease	Vocal Tasks	Relevant Audio Features
Parkinson's Disease [61]	Sustained Phonation, Reading text, Spontaneous Speech, diadochokinetic task	Reduced Pitch Range, Increased Jitter and Shimmer, Reduced Speech Rate, Altered Formant Frequencies
Depression [57]	Sustained Phonation, Reading text	Increased Jitter and Shimmer, Reduced HNR, Slower Speech Rate, Altered Formant Frequencies
Voice Disorders [62]	Sustained Phonation, Reading text	Increased Jitter and Shimmer, Reduced HNR, Changes in Pitch and Intensity

Table 4: Examples of disease-specific vocal tasks and corresponding audio features.

Delineation of the PhD Scope

This PhD thesis focuses on a subset of key potential applications of vocal biomarkers, as shown in Figure 11. The related case studies covered in this work were selected based on their potential to advance priority topics of the vocal biomarker research field. Each of the following objective of the PhD thesis aligns with a particular application of vocal biomarkers:

- **Objective 1** focuses on symptom detection and monitoring by exploring fatigue symptoms in COVID-19 patients. COVID-19 provides an opportunistic use case for the study of vocal biomarkers due to the global health need for remote monitoring solutions that can minimize physical contact while still ensuring effective patient management.
- **Objective 2** targets screening by examining how vocal biomarkers can be used to screen for T2D. Exploring T2D aligns with the broader goal of developing non-invasive, cost-effective screening tools, especially in resource-limited settings where access to traditional diagnostic methods is challenging.
- **Objective 3** addresses perceived health status monitoring by assessing RQoL in the general population with participants with various respiratory diseases. RQoL is directly linked to respiratory function, which naturally influences vocal characteristics. As respiratory diseases are prevalent, this application demonstrates the potential of vocal biomarkers for scalable, non-invasive health monitoring, allowing continuous assessment without the need for clinical visits.

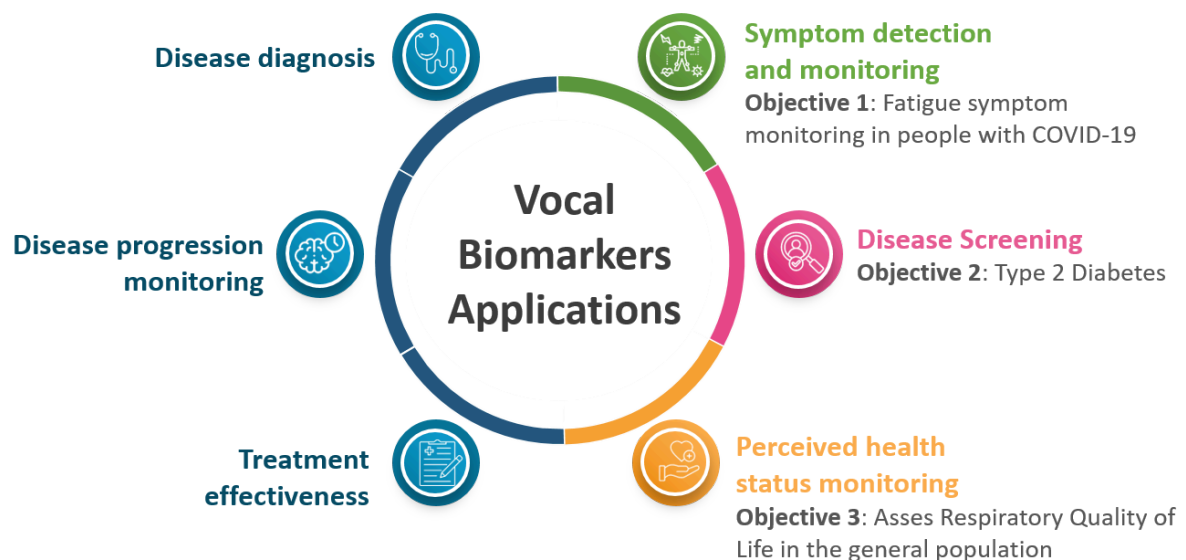


Figure 11. Key potential applications of vocal biomarkers explored in this PhD thesis

Objective 1: Symptom Detection and Monitoring

Contextualizing COVID-19 Symptoms and Fatigue

COVID-19, induced by the SARS-CoV-2 virus, manifests a spectrum of symptoms from mild to severe. The World Health Organization (WHO) outlines clinical criteria for suspected infection, including acute onset of fever and cough, complemented by other symptoms like fatigue, headache, and dyspnea. Fatigue emerges as one of the most frequently reported symptoms, persisting even during and beyond the recovery phase, with up to 60% of patients reporting chronic fatigue 12 months later[63]. This persistence significantly affects the quality of life and complicates disease management strategies[64], supporting the need for long-term monitoring solutions for these patients.

Defining and Understanding Fatigue

Fatigue is characterized as an overwhelming sensation of tiredness, lack of energy, and exhaustion that sleep or rest does not alleviate[65]. It is subjectively experienced and can manifest in acute or chronic forms, significantly influencing daily activities and overall quality of life. The clinical dimensions of fatigue are multifaceted, encompassing physical, cognitive, and emotional aspects[65,66]. Physically, it might manifest as diminished physical activity and extended recovery periods[67]. Cognitively, it impairs concentration, induces memory lapses, and reduces mental endurance[68]. Emotionally, fatigue can diminish motivation and interest in previously enjoyed activities[69].

Biological Mechanisms of Fatigue in COVID-19

The persistence of fatigue in Long COVID may stem from diverse biological mechanisms, including cerebral and peripheral pathologies, as well as psychosocial factors[70]. Chronic inflammation at the neuromuscular junctions and brain may induce enduring fatigue[70,71]. Additionally, damage to muscle fibers and sarcolemma, alongside skeletal muscle injuries, contributes to the onset of post-COVID fatigue, suggesting a complex interplay of factors affecting the trajectories of patients' recovery.

Persistent Challenges of Long COVID

As the pandemic progresses, reports of prolonged symptomatology post-COVID-19, termed 'Long COVID syndrome,' have increased[72]. This condition is associated with a variety of symptoms impacting multiple organs and presenting significant challenges in healthcare[73]. Fatigue remains a critical, persistent symptom irrespective of the initial disease severity. Studies indicate that up to 60% of patients may experience chronic fatigue twelve months post-recovery[73]. Chronic fatigue has been documented as the most prevalent long-term symptom following the acute phase of COVID-19[73–77] (Figure 12), emphasizing the need for effective monitoring and management strategies.

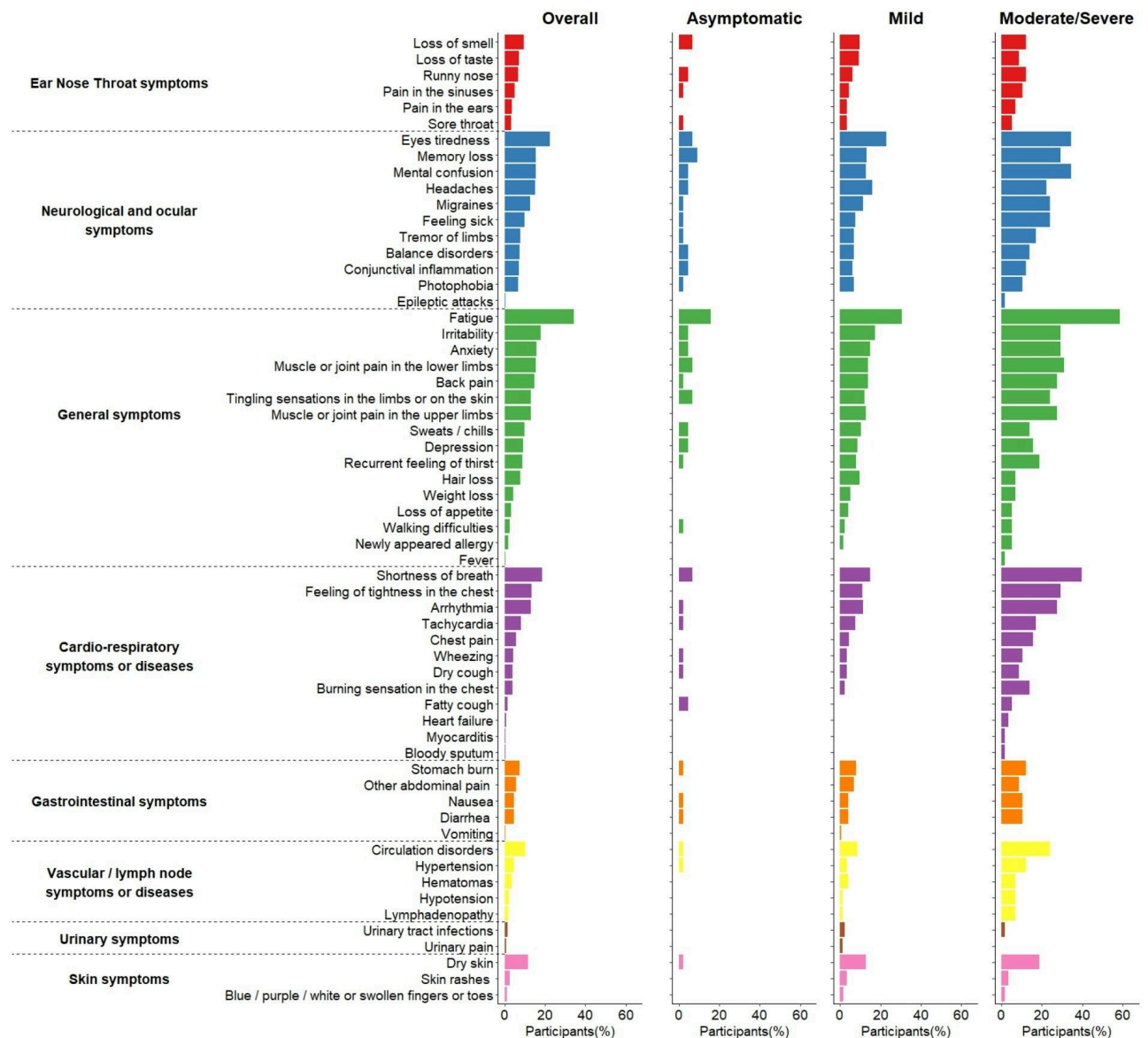


Figure 12. Description of 60 persisting symptoms, 12 months after the acute infection (from Fischer A. et al. Long COVID Symptomatology After 12 Months and Its Impact on Quality of Life According to Initial Coronavirus Disease 2019 Disease Severity.)

Objective 2: Disease Screening

Diabetes Mellitus and its classification

Diabetes mellitus is a complex group of metabolic diseases characterized by chronic hyperglycemia, which arises from issues in insulin secretion, insulin action, or both[78]. The importance of insulin as an anabolic hormone means that metabolic abnormalities in carbohydrates, lipids, and proteins can have significant impacts. Insufficient insulin levels or insulin resistance in target tissues such as skeletal muscles, adipose tissue, and the liver result from defects at various points in the insulin signaling pathway. The American Diabetes Association (ADA) classifies diabetes into type 1, type 2, other specific types, and

gestational diabetes mellitus (GDM), each with specific pathophysiologies and clinical implications.

Type 2 Diabetes (T2D)

T2D, previously known as non-insulin-dependent diabetes, is the most prevalent form of diabetes[78,79]. It is primarily characterized by hyperglycemia, insulin resistance, and relative insulin deficiency. The development of T2D results from a complex interplay of genetic, environmental, and behavioral factors. This form of diabetes is particularly concerning due to its common occurrence, subtle onset, and frequently delayed diagnosis, which are even more pronounced in resource-poor settings.

Challenges in managing T2D

Individuals with T2D are at an increased risk of both acute and long-term complications, leading often to premature mortality. This is worsened by the high prevalence of undiagnosed cases, with almost one in every two people with T2D remaining undiagnosed worldwide[80]. As a result, these individuals cannot begin treatment or preventive measures to avoid or delay complications, particularly in resource-limited environments where traditional screening methods, such as invasive blood analysis, are costly and challenging to implement[81]. Alternative methods, such as questionnaires and risk scores like the ADA risk assessment[82], are used, but these too have limitations, including susceptibility to biases and errors.

Objective 3: Perceived Health Status Monitoring

Context and importance of telemonitoring

Monitoring chronic respiratory diseases or other conditions affecting breathing is fundamental in respiratory healthcare. Telemonitoring solutions reduce clinician workload, decrease hospital admissions, and enable timely interventions. Remote monitoring is crucial for identifying clinically relevant deterioration in RQoL and can serve as a prognostic tool for conditions like COPD and asthma. For instance, a recent study found that a 4-point decline in RQoL over one year, as assessed by the St George's Respiratory Questionnaire (SGRQ)[83], was linked to higher rates of hospitalization and mortality. This highlights the need for continuous monitoring to identify early signs of deterioration and facilitate timely management.

Understanding Respiratory Quality of Life (RQoL)

RQoL measures a patient's well-being and daily functioning with their respiratory health, typically assessed through questionnaires like SGRQ, CRQ[84], BPQ[85], and VQ11[86]. These tools quantify the impact of respiratory diseases on patients' lives, providing insights into symptom severity, physical functioning, and emotional well-being. Accurate RQoL assessment is vital for effective disease management and improving patient outcomes.

Challenges in RQoL assessment

Traditional RQoL assessment methods face several challenges. Questionnaires depend on patients' self-reports, which can be subjective and biased. Completing detailed questionnaires can be time-consuming and burdensome, leading to potential

non-compliance or inaccurate reporting. These methods may also require clinical visits and invasive equipment, which are not always practical for continuous monitoring. Addressing these challenges can make RQoL monitoring more reliable and practical, offering a non-invasive, cost-efficient solution for remote monitoring and better disease management.

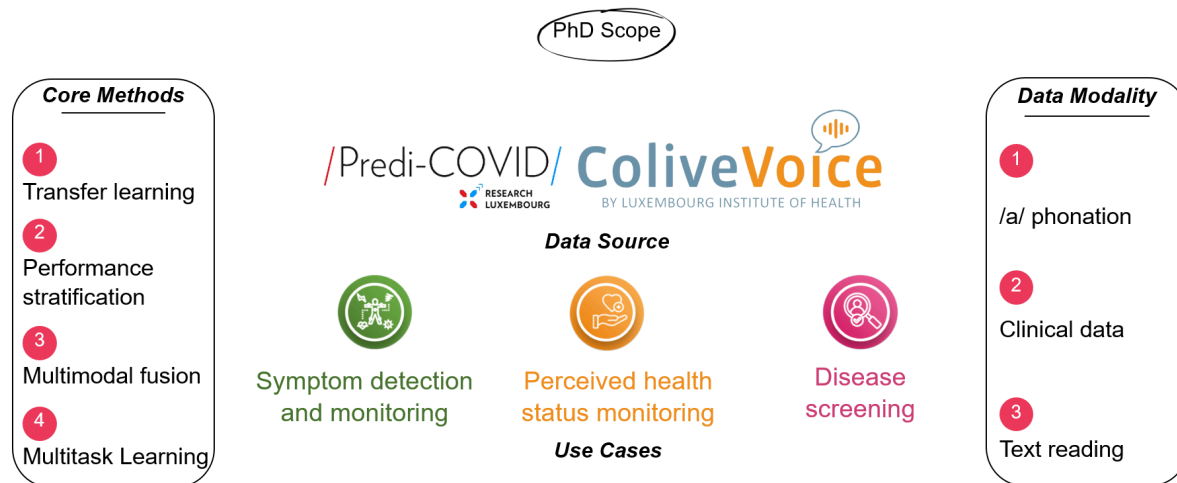


Figure 13. PhD scope

The scope of this PhD research, as shown in Figure 13, connects the foundational ideas discussed in the synopsis with the specific objectives and methods explored in the upcoming chapters. Using data from the Predi-COVID and Colive Voice studies, the PhD scope focuses on three key areas: symptom detection and monitoring, perceived health status monitoring, and disease screening. Advanced techniques namely transfer learning, performance stratification, multimodal fusion, and multitask learning were applied while using various data modalities, including /a/ phonation, clinical data, and text reading.

This broad overview leads into the detailed examination of each objective in the next chapters, where the use of these methods and data sources will be explained with the specific research papers and findings.

Chapter 4

Vocal biomarkers for symptom detection and monitoring

In this chapter, Objective 1 is addressed, focusing on symptom detection and monitoring through the exploration of fatigue symptoms in COVID-19 patients. The results and discussion presented here detail the findings from the related research paper[87], highlighting the development and application of vocal biomarkers for this purpose.

Results

General Workflow

We followed a comprehensive workflow to derive a vocal biomarker for fatigue using data from participants in Predi-COVID[2], as shown in Figure 14. Stratification by the operating system was implemented to account for potential differences in audio quality and processing between Android and iOS devices, which could introduce variability in the recordings and impact the analysis. Additionally, we stratified by gender to address known gender-based differences in vocal characteristics, ensuring that our findings were not biased by these inherent variations. This stratification was essential to reduce heterogeneity and improve the reliability and robustness of the vocal biomarker across diverse participant groups.

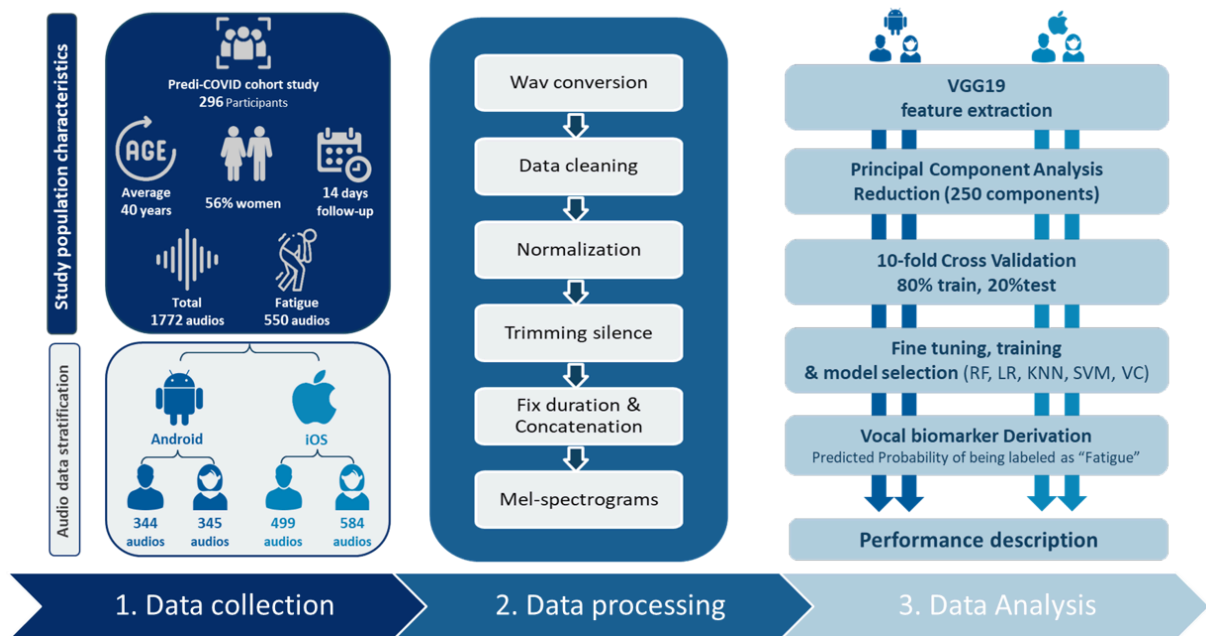


Figure 14. General pipeline for fatigue vocal biomarker development (from Elbéji A., et al. Vocal biomarker predicts fatigue in people with COVID-19: results from the prospective Predi-COVID cohort study.)

Audio recordings in 3gp (Android) and m4a (iOS) formats were converted to .wav, excluding those shorter than 2 seconds. Two audio types were considered: text reading and /a/ vowel phonation. Poor-quality recordings were excluded using DBSCAN-based clustering[88]. Peak normalization and trimming of silences were applied to the remaining high-quality recordings. The /a/ vowel phonation and text reading recordings from the same participant on the same day were then concatenated. Participants were categorized into 'fatigue' and 'no fatigue' groups based on self-reported assessments, with descriptive statistics used for characterization.

Voice features were extracted using the pre-trained VGG19, reduced to 250 PCA components explaining 97% and 99% of the variance in the data for iOS and Android audio sets, and evaluated with 10-fold cross-validation across four models (logistic regression, k-nearest neighbors, support vector machine, and soft voting classifier). To ensure methodological rigor and prevent potential data leakage, the cross-validation was stratified at the patient level, ensuring that no recordings from the same individual appeared simultaneously in both the training and validation sets. Metrics such as AUC, accuracy, F1-score, precision, recall, and Brier score were used to assess performance. The best model's predicted probability of fatigue served as the final vocal biomarker, intended as a quantitative measure to monitor fatigue. Results are detailed in the subsequent sections.

Study Population Characteristics

We analysed a total of 296 Predi-COVID participants who reported their fatigue status (“I feel well” as “No Fatigue” and “I am fatigued”/“I don’t feel well” as “Fatigue”) on the same day as the audio recordings during the 14 days of follow-up. This resulted in a total of 1772 audio recordings collected with an average of 6 days of reporting fatigue status and providing audio recordings. Their mean age was 40 years (SD=13), of which 165 (56%) were women. No significant differences were observed in age, gender, body mass index, smoking status, antibiotic usage, and asthma prevalence between users of Android and iOS devices ($p>0.05$), indicating a balanced representation across the primary device platforms used for data collection (Table 5).

		All	m4a		3gp		P-values (m4a, 3gp)
			Female	Male	Female	Male	
Participants (N)	Total	296	107	80	51	58	-
Age (years)	mean (SD)	40.3 (12.6)	38.8 (13.4)	42.9 (12.7)	37.8 (11.6)	41.5 (11.3)	0.28
Body Mass Index (kg/m²)	mean (SD)	24.1 (4.7)	24.6 (5.5)	26.5 (4.1)	24.1 (3.8)	26.6 (4.17)	0.95
Antibiotic (%)	No	265 (90%)	93 (87%)	73 (91%)	44 (86%)	55 (95%)	0.87
	Yes	31 (10%)	14 (13%)	7 (9%)	7 (14%)	3 (5%)	
Asthma (%)	No	284 (96%)	104 (97%)	75 (94%)	47 (92%)	58 (100%)	0.82
	Yes	12 (4%)	3 (3%)	5 (6%)	4 (8%)	0 (0%)	
Smoking (%)	Never	199 (67%)	77 (72%)	51 (64%)	36 (71%)	35 (60%)	0.41
	Former smoker	53 (18%)	19 (18%)	20 (25%)	9 (18%)	13 (22%)	
	Current smoker	44 (15%)	11 (10%)	9 (11%)	6 (11%)	10 (18%)	
Audio recordings	Total	1772	584	499	345	344	<0.001
	No Fatigue	1222 (69%)	394 (67%)	370 (74%)	190 (55%)	268 (78%)	
	Fatigue	550 (31%)	190 (33%)	129 (26%)	155 (45%)	76 (22%)	
Mean (SD) and maximum of audio recording per participant in the 14-day follow-up period	mean (SD)	6 (5)	6 (5)	6 (5)	6 (5)	6 (5)	-
	max	16	14	16	15	14	

Table 5. Characteristics of Predi-COVID participants by device type and gender

Model Performance and Vocal Biomarker Development

For Android users, the best model achieved a weighted Area Under the Curve (AUC) of 86% for females and 82% for males, with a mean Brier Score of 0.15 and 0.12 respectively. For iOS users, the performance of the best model showed an AUC of 79% for females and 85% for males, with a mean Brier Score of 0.17 and 0.12 respectively (Table 6).

Audio format	Gender	Algorithm	Accuracy	Precision	Recall	F1-score	Weighted AUC
3gp (Android)	Female	LR	0.77	0.77	0.77	0.77	0.85
		KNN	0.72	0.73	0.72	0.72	0.76
		SVM	0.80	0.80	0.80	0.80	0.86
		VC	0.78	0.78	0.78	0.78	0.86
	Male	LR	0.78	0.79	0.78	0.79	0.81
		KNN	0.83	0.83	0.83	0.79	0.84
		SVM	0.84	0.83	0.84	0.83	0.82
		VC	0.84	0.84	0.84	0.84	0.82
m4a (iOS)	Female	LR	0.72	0.72	0.72	0.72	0.75
		KNN	0.68	0.65	0.68	0.65	0.67
		SVM	0.79	0.79	0.79	0.79	0.79
		VC	0.77	0.76	0.77	0.76	0.78
	Male	LR	0.73	0.74	0.73	0.73	0.80
		KNN	0.89	0.89	0.89	0.88	0.81
		SVM	0.85	0.84	0.85	0.84	0.85
		VC	0.89	0.89	0.89	0.88	0.85

Table 6. Results of the prediction models for Fatigue status classification

The vocal biomarker effectively differentiated between fatigued and non-fatigued COVID-19 participants, demonstrating the feasibility of using vocal analysis as a non-invasive tool to monitor symptoms remotely. The performance of these models illustrates the potential of integrating digital health technologies into clinical practice, particularly in the management of pandemic-related conditions.

Discussion

Fatigue and voice

Prior research underscores the impact of fatigue on voice quality and speech dynamics. Fatigue influences key vocal characteristics such as pitch, word duration[2,89], and the timing of articulated sounds, which can significantly alter speech production[90]. Subtle phonetic changes correlated with fatigue levels have been quantitatively analysed, revealing potential diagnostic applications of vocal analysis for fatigue detection[89,90].

Given the respiratory and vocal symptoms experienced by COVID-19 patients, the process of voice production may be compromised. Dysphonia occurs in up to 49% of patients[89–91], particularly those experiencing fatigue and respiratory distress. This highlights the importance of vocal analysis in these cases[91]. The research indicates that circadian rhythms[92] also affect speech dynamics, suggesting that vocal analysis could be effectively used to monitor fatigue in real-time, especially in settings that demand high vigilance[90].

Furthermore, Long COVID[90,93] exhibits many similarities with Chronic Fatigue Syndrome (CFS)[94]. Studies reveal overlapping biological mechanisms between these conditions, indicating that insights into one disorder can enhance understanding of the other[95]. CFS alters vocal quality but also affects phonation and articulation linked to neuromuscular, hormonal, and autonomic dysfunctions[96]. Using the Chalder Fatigue Scale, Voice Handicap Index, Voice Rating Scale and acoustic analyses, a strong correlation between vocal quality and fatigue severity has been shown, with gender-specific differences in how fatigue affects vocal parameters. This body of work highlights the systemic nature of chronic fatigue syndrome and its impact beyond conventional symptoms, affecting vocal characteristics.

Implications of the study

Clinical Implication

This study underscores the potential utility of vocal biomarkers in identifying fatigue among COVID-19 patients, providing a foundation for telemedicine applications aimed at enhancing patient care. The ability to monitor fatigue remotely through voice analysis not only helps in managing COVID-19 but also has implications for other chronic conditions where fatigue is a predominant symptom.

Technical implication

The study also highlighted the importance of preventing data heterogeneity which is considered contamination leading to challenges in building reliable and robust algorithms. This contamination is caused by two factors: first, the significant differences in how men and women experience and report fatigue[96,97], and second, the variability in microphone quality between different smartphone platforms (iOS and Android) used by participants, which have a direct impact on the quality of the recorded audios (Figure 15).

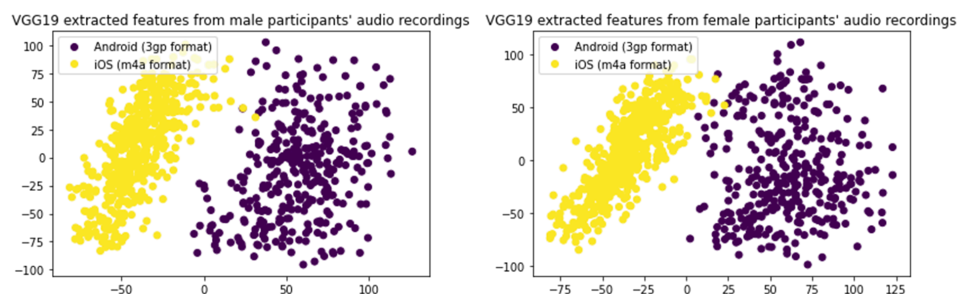


Figure 15. VGG19 extracted features by gender and operating system

In addition to leveraging advanced and innovative technologies for algorithm development, this approach involves converting audio recordings into mel-spectrograms, which are two-dimensional representations of the audio signal's frequency content over time. While mel-spectrograms are not traditionally considered images, they share structural similarities with image data, which allows us to use image-based pre-trained models like VGG19 (Figure 16). The decision to use VGG19 stems from its robust performance in extracting complex patterns in 2D data, despite being originally designed for image recognition. By applying transfer learning from VGG19, we benefit from its capability to capture intricate features in a manner that significantly reduces the need for large datasets, shortens training times, and enhances model performance compared to training models from scratch. Although models pre-trained specifically on spectrograms were also considered (e.g. VGGish[11]), VGG19 demonstrated superior performance in our case, possibly due to its ability to capture high-level abstractions in 2D representations like mel-spectrograms. This decision underscores the flexibility and power of transfer learning in cases where traditional spectrogram-based models may not perform as well.

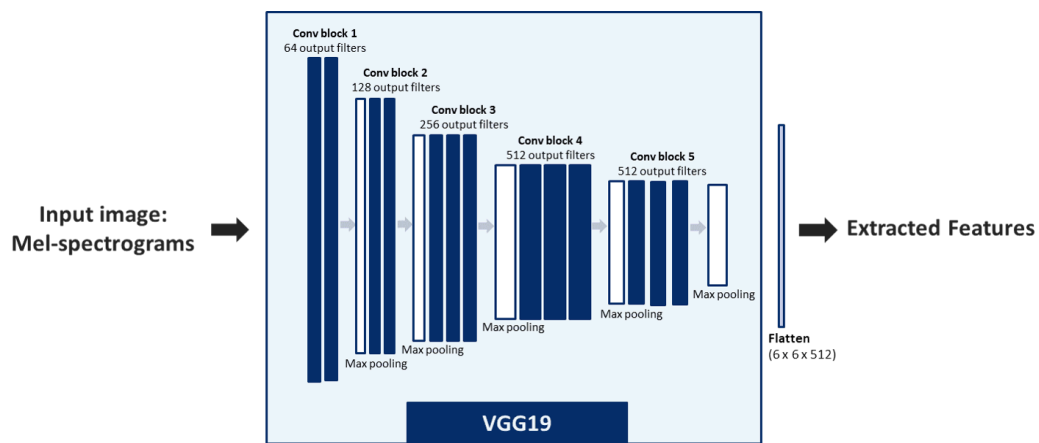


Figure 16. Feature extraction pipeline using VGG19 on mel-spectrograms

Limitations of the study

While the findings are promising, the limitations of the study must be acknowledged. The algorithms were based on a rather limited dataset, primarily from one geographic location (Luxembourg), which could affect the generalizability of the results. The use of binary fatigue classification (fatigued vs. non-fatigued) could also fail to capture the full spectrum of fatigue severity and multidimensionality.

Furthermore, although the data were stratified by gender and device type to limit biases, the diversity in spoken languages could have influenced voice feature consistency and model performance. The absence of a comparable dataset for external validation and the reliance on qualitative self-reported measures of fatigue, along with the lack of individual time-series voice analysis, could compromise the depth and accuracy of our results. These factors point to crucial areas for improvement, emphasizing the need for more robust data collection and analysis to enhance the reliability of vocal biomarker technology in future research.

Advancing fatigue assessment: Fatigue Severity Scale

Moving forward, a key enhancement would be the transition from the subjective binary (fatigued vs. non-fatigued) to a more nuanced scale such as the Fatigue Severity Scale (FSS)[98]. The FSS includes sub-dimensions like physical fatigue, mental fatigue, and the impact of fatigue on daily functioning. Such a scale would allow for a detailed, quantitative evaluation of fatigue, capturing its complexity and variability across individuals. This comprehensive assessment could also improve our understanding of how voice characteristics correlate with different levels of fatigue severity, and voice may likely be more sensitive to certain sub-dimensions, such as mental fatigue, than others. This is particularly valuable in longitudinal settings, where tracking these subtleties is crucial for understanding the progression of conditions associated with fatigue and for tailoring interventions that are responsive to changes in a patient's condition over time. The FSS is available in the Colive Voice study.

Extending the impact beyond COVID-19

Fatigue is a prevalent symptom in many chronic diseases[11,98,99], including multiple sclerosis[100] and cancer[101], underscoring the broader relevance of this study beyond COVID-19. The significant variability in how fatigue manifests across these conditions can be disabling. Vocal biomarkers offer a potential universal tool for assessing fatigue across various diseases, enhancing diagnostic accuracy, and facilitating more comprehensive disease management. It is also crucial to validate this work in people with Long COVID, where fatigue remains the most frequent persisting symptom[74–77].

Further research should focus on integrating these biomarkers into clinical workflows to develop real-time monitoring tools that provide immediate feedback to healthcare providers and patients[102]. Additionally, expanding the dataset to include a more diverse range of populations and incorporating longitudinal data will enhance the robustness and applicability of the predictive models, ensuring they are effective across different demographic and disease contexts.

Conclusion

The development of a vocal biomarker for fatigue in COVID-19 patients represents a significant step forward in the remote monitoring capabilities of health technology. This study provides a basis for further exploration into the use of voice as a digital biomarker, highlighting its potential to improve patient outcomes through non-invasive monitoring techniques. Future enhancements in this field will likely focus on refining the performance of fatigue vocal biomarkers and expanding their use to other diseases and conditions beyond COVID-19.

Chapter 5

Vocal Biomarkers for Disease Screening

For Objective 2, this chapter focuses on screening by examining how vocal biomarkers can be used to detect T2D. The following sections detail the results and discussion from the related research paper, demonstrating the potential of voice analysis in early diabetes screening.

Results

General workflow

Figure 17 shows the workflow to develop a voice-based T2D status classifier using data from USA participants in the Colive Voice study. Stratification by gender was used to reduce heterogeneity and address gender-based biases.

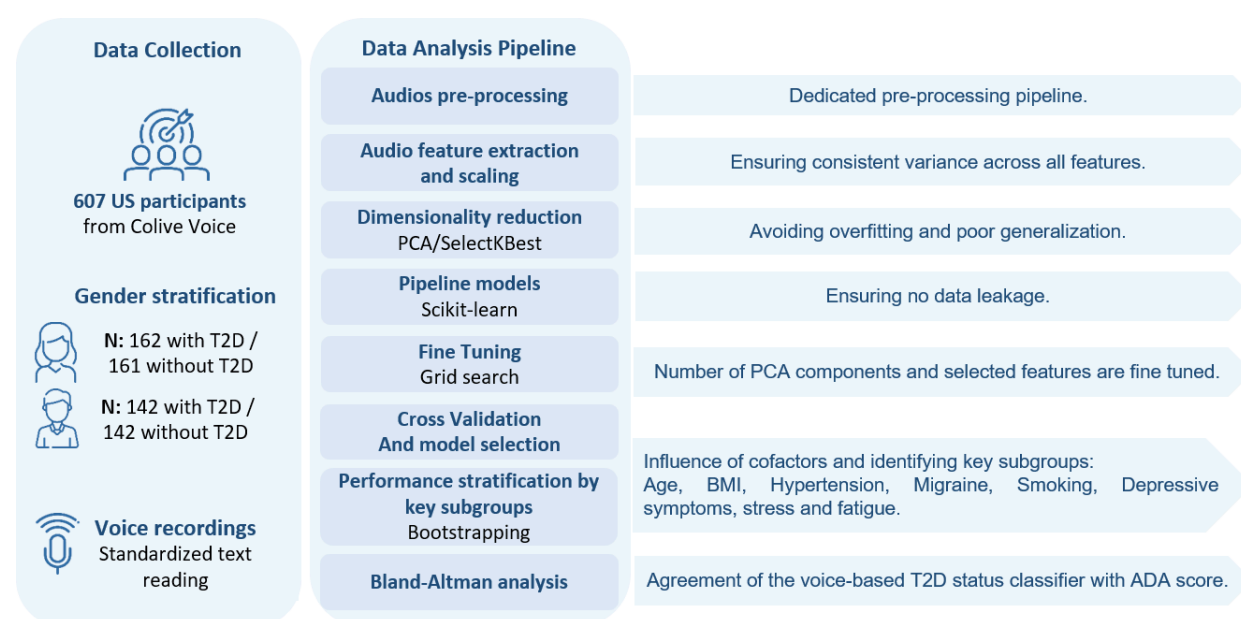


Figure 17. T2D status classification pipeline using voice data from Colive Voice USA participants

The collected voice recordings of standardized text readings were pre-processed using a dedicated pipeline (see section [Data Processing](#)). Following preprocessing and quality checking, audio features were extracted and scaled to ensure consistent variance across all features. Dimensionality reduction was applied using PCA for BYOL-S embeddings and SelectKBest for OpenSMILE features due to their differing nature. BYOL-S embeddings, being high-dimensional and non-explainable, PCA was applied to reduce dimensionality while retaining most of the variance, ensuring that the model generalizes effectively without overfitting on irrelevant noise or redundant information. In contrast, OpenSMILE features are interpretable, handcrafted audio features (e.g., pitch, energy, and formants), so SelectKBest was used to focus on the most relevant ones for the task, enhancing model performance and generalization.

We trained several models using Scikit-learn, with hyperparameter-tuning performed via grid search to optimize the selection of the number of PCA components and selected features in

addition to algorithm hyperparameters. 5-fold stratified Cross-validation was used for model selection and validation, ensuring the robustness of our approach.

The performance of the best-performing algorithm was stratified by key subgroups, including demographics, comorbidities, lifestyle factors, and symptoms. To reinforce confidence in our performance metrics and facilitate comparisons, we employed a bootstrapping technique. Specifically, 1000 subsamples were generated for each combination of comorbidity and its status, with sampling conducted with replacement from the original dataset. This process, repeated for each comorbidity, enabled us to recalculate the metrics for each subsample, ensuring robust performance estimates and accounting for variability in the presence or severity of these conditions. Bootstrapping is crucial in this context as it provides more reliable estimates by simulating the variability inherent in small or imbalanced datasets.

Additionally, Bland-Altman analysis was conducted to compare the agreement of the voice-based T2D status classifier with the ADA risk score. Detailed results are presented in subsequent sections.

Study population characteristics

The analysis included 323 females and 284 males as shown in Table 7, with a predominantly white demographic (76.5% of females and 71.8% of males with T2D). Significant differences were observed between groups in age, BMI, and prevalence of hypertension, which were all higher among participants with T2D ($p < 0.001$ for all). The prevalence of diagnosed depression was also higher in the T2D group compared to those without, particularly among females (61.7% vs. 45.3%). We analysed standardized text voice recordings from these participants.

	Female group			Male group		
T2D status	Without T2D	With T2D	P-value	Without T2D	With T2D	P-value
Participants (N)	161	162	-	142	142	-
Age (year)	40.0 (13.5)	49.5 (12.1)	<0.001	41.6 (14.0)	47.6 (13.4)	<0.001
Body Mass Index (kg/m ²)	28.0 (7.3)	35.8 (8.9)	<0.001	26.6 (5.5)	32.8 (8.5)	<0.001
Ethnicity: White	118 (73.3%)	124 (76.5%)	0.28	110 (77.5%)	102 (71.8%)	0.59
Ethnicity: Black	20 (12.4%)	21 (13.0%)		10 (7.0%)	12 (8.5%)	
Ethnicity: Other	23 (14.3%)	17 (10.5%)		22 (15.5%)	28 (19.7%)	
Fatigue Severity Scale	32.3 (13.4)	40.3 (12.3)	<0.001	31.3 (12.8)	40.3 (12.3)	<0.001
Perceived stress (% yes)	38 (23.6%)	49 (30.3%)	0.48	29 (20.4%)	38 (26.7%)	0.16
Smoking (% yes)	28 (17.4%)	19 (11.7%)	0.22	32 (22.5%)	34 (23.9%)	0.24
Migraine (% yes)	33 (20.5%)	43 (26.5%)	0.25	16 (11.3%)	19 (13.4%)	0.72
Thyroid disease (% yes)	0 (0%)	37 (22.8%)	<0.001	0 (0%)	10 (0.7%)	<0.01
Hypertension (% yes)	18 (11.2%)	81 (50.0%)	<0.001	18 (12.7%)	83 (58.5%)	<0.001
Diagnosed depression (% yes)	73 (45.3%)	100 (61.7%)	<0.01	45 (31.7%)	69 (48.6%)	<0.01
HbA1c (%)	-	7.14 (1.8)	-	-	7.20 (1.7)	-
Diabetes treatment (% yes)	-	126 (77.8%)	-	-	114 (80.3%)	-
Diabetes duration (year)	-	8.9 (7.3)	-	-	9.1 (7.6)	-

Table 7. Demographic and clinical characteristics of participants with and without T2D, stratified by gender

Algorithm performance

The MLP classifiers trained with BYOL-S/CvT embeddings showed the best performance for both genders (Table 8). For predicting T2D in females, the classifier achieved a sensitivity of

0.67±0.11, a specificity of 0.66±0.04, an AUC of 0.71±0.07, and a Brier score of 0.31. In males, the performance metrics included a sensitivity of 0.73±0.03, a specificity of 0.70±0.02, an AUC of 0.75±0.05, and a Brier score of 0.22. The predicted probability of having T2D was subsequently used for sensitivity analysis with the ADA risk score.

	Features	Dimensionality reduction	Classifier	Accuracy	Specificity	Sensitivity	AUC
Female group	OpenSMILE ComParE 2016 (6373)	200 selected features	LR	0.60 (0.03)	0.60 (0.03)	0.62 (0.07)	0.62 (0.02)
			MLP Classifier	0.63 (0.02)	0.61 (0.02)	0.74 (0.02)	0.66 (0.02)
			SVM RBF	0.57 (0.02)	0.57 (0.02)	0.63 (0.03)	0.61 (0.01)
	Byol-S embeddings (2048)	PCA, n_components= n_samples	LR	0.67 (0.04)	0.68 (0.04)	0.65 (0.11)	0.70 (0.06)
			MLP Classifier	0.67 (0.04)	0.66 (0.04)	0.67 (0.11)	0.71 (0.07)
			SVM RBF	0.66 (0.04)	0.65 (0.07)	0.67 (0.11)	0.71 (0.05)
Male group	OpenSMILE ComParE 2016 (6373)	100 selected features	LR	0.56 (0.02)	0.55 (0.01)	0.58 (0.05)	0.61 (0.05)
			MLP Classifier	0.61 (0.05)	0.61 (0.06)	0.63 (0.06)	0.64 (0.05)
			SVM RBF	0.57 (0.05)	0.57 (0.06)	0.54 (0.05)	0.57 (0.05)
	Byol-S embeddings (2048)	PCA, n_components= 100	LR	0.69 (0.04)	0.66 (0.07)	0.72 (0.03)	0.73 (0.06)
			MLP Classifier	0.71 (0.02)	0.70 (0.02)	0.73 (0.03)	0.75 (0.05)
			SVM RBF	0.70 (0.04)	0.64 (0.05)	0.76 (0.03)	0.78 (0.05)

Table 8. Results of the prediction models for T2D status classification

Performance stratification based on cofactors

To study the influence of important clinical factors on the performance of the predictive algorithms of T2D status classification, a performance analysis was conducted. This analysis was segmented by demographics (age and BMI), comorbidities (migraine and hypertension), and lifestyle factors and symptoms (smoking, depressive symptoms, stress, and fatigue). For this, we conducted a bootstrapping technique to enhance the reliability of the performance metrics, and to facilitate robust comparisons across different subgroups.

As shown in Table 9 regarding demographics, we observed differences based on age. Females aged 60 and above demonstrated superior performance metrics with a specificity of 0.74 ± 0.12 , sensitivity of 0.74 ± 0.07 , and AUC of 0.74 ± 0.07 . In contrast, younger females (below 60) showed lower scores in specificity and sensitivity (0.65 ± 0.04), and AUC (0.65 ± 0.03), suggesting that age is a critical factor in the algorithm's performance.

			Females			Males		
			Specificity	Sensitivity	AUC	Specificity	Sensitivity	AUC
Demographics	Age	<60 y	0.65 (0.04)	0.65 (0.04)	0.65 (0.03)	0.70 (0.04)	0.74 (0.04)	0.72 (0.03)
		≥ 60y	0.74 (0.12)	0.74 (0.07)	0.74 (0.07)	0.70 (0.11)	0.70 (0.10)	0.70 (0.07)
	Body Mass Index	<25 kg/m ²	0.68 (0.06)	0.58 (0.12)	0.63 (0.07)	0.70 (0.06)	0.78 (0.09)	0.74 (0.05)
		≥ 25 kg/m ²	0.65 (0.05)	0.68 (0.04)	0.67 (0.03)	0.69 (0.05)	0.72 (0.04)	0.71 (0.03)
Comorbidities	Hypertension	Present	0.76 (0.11)	0.75 (0.05)	0.75 (0.06)	0.72 (0.11)	0.76 (0.05)	0.74 (0.06)
		Absent	0.65 (0.04)	0.61 (0.05)	0.63 (0.03)	0.69 (0.04)	0.70 (0.05)	0.70 (0.03)
	Migraine	Present	0.86 (0.07)	0.75 (0.07)	0.80 (0.05)	0.67 (0.12)	0.71 (0.11)	0.69 (0.09)
		Absent	0.62 (0.04)	0.65 (0.04)	0.65 (0.04)	0.70 (0.04)	0.74 (0.04)	0.72 (0.03)
Lifestyle factors and symptoms	Smoking	Present	0.60 (0.09)	0.53 (0.12)	0.57 (0.07)	0.74 (0.09)	0.76 (0.07)	0.75 (0.06)
		Absent	0.67 (0.04)	0.69 (0.04)	0.68 (0.03)	0.69 (0.04)	0.72 (0.04)	0.71 (0.03)
	Depressive symptoms	Severe	0.75 (0.05)	0.71 (0.05)	0.73 (0.03)	0.71 (0.07)	0.71 (0.06)	0.71 (0.04)
		Mild	0.58 (0.05)	0.61 (0.06)	0.60 (0.04)	0.69 (0.05)	0.75 (0.05)	0.72 (0.03)
	Stress	Present	0.76 (0.07)	0.62 (0.07)	0.69 (0.05)	0.69 (0.09)	0.77 (0.07)	0.72 (0.06)
		Absent	0.63 (0.04)	0.70 (0.04)	0.66 (0.03)	0.70 (0.04)	0.72 (0.04)	0.71 (0.03)
	Fatigue	Severe	0.68 (0.06)	0.68 (0.05)	0.68 (0.04)	0.71 (0.06)	0.73 (0.05)	0.72 (0.04)
		Mild	0.65 (0.05)	0.66 (0.06)	0.65 (0.04)	0.69 (0.05)	0.73 (0.06)	0.71 (0.04)

Table 9. Performance stratification of voice-based T2D status detection algorithms

In terms of comorbidities, hypertension was identified as a significant enhancer of algorithm performance across both genders. The presence of hypertension improved sensitivity to 0.75 ± 0.05 for females and 0.76 ± 0.05 for males, indicating an increased efficiency of the algorithm in detecting T2D among individuals with hypertension.

The analysis also highlighted the significant influence of the presence of severe depressive symptoms, particularly in females, enhancing both specificity (0.75 ± 0.05) and sensitivity

(0.71 ± 0.05). This suggests that depressive symptoms may modify some voice features that contribute to better discrimination between people with T2D and those without T2D, especially in women.

Conversely, smoking exhibited divergent effects on sensitivity between genders, with an increase in males (0.76 ± 0.07) and a decrease in females (0.53 ± 0.12). This gender-specific response underscores the need for tailored approaches in the algorithm's application.

These findings emphasize the importance of incorporating these variables into the development and ongoing refinement of screening tools. By doing so, we can ensure more accurate and gender-specific healthcare strategies for the management and diagnosis of T2D, making these tools not only more effective but also more personalized.

Discussion

Voice Alterations and Underlying Mechanisms in T2D

Exploring the interface between T2D and voice quality, significant research has highlighted that T2D can induce vocal changes through complications like neuropathy[103] or poor glycemic control[104] (Figure 18). Early findings indicated that individuals with T2D often experience phonatory symptoms such as vocal straining[105,106] and hoarseness[105,107], particularly those with poor glycemic control and neuropathy. These patients often have reduced maximum phonation times, indicating neuromuscular and respiratory alterations[105][106].

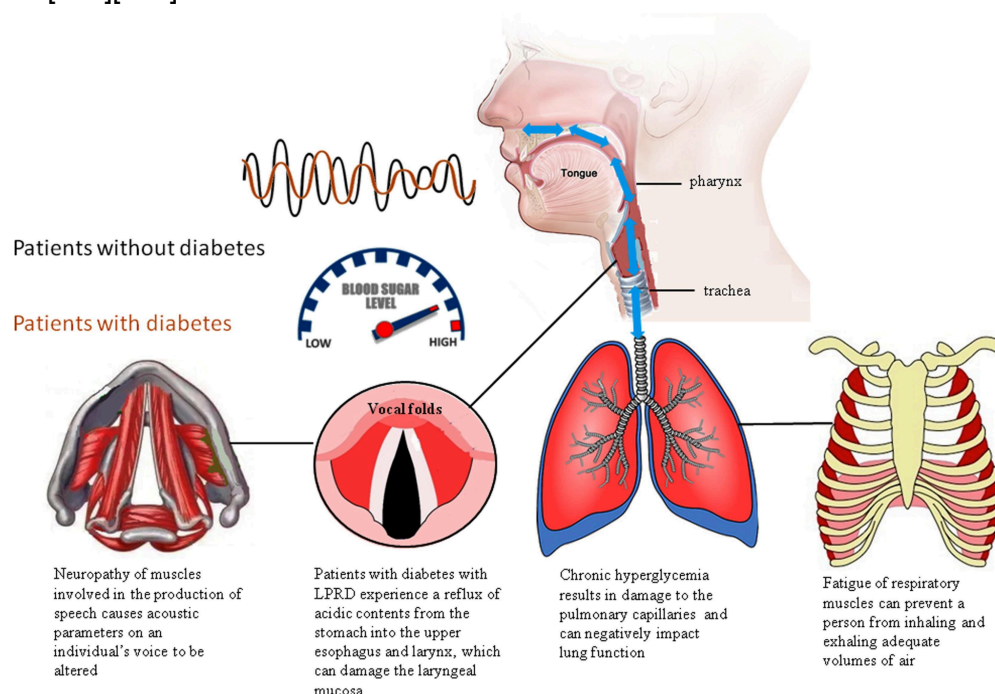


Figure 18. T2D physiological changes that affect the production of speech (from Saghiri, M., et al. Scoping review of the relationship between diabetes and voice quality. *Diabetes Research and Clinical Practice*, 2023).

Additionally, a recent study showed promising results in using voice analysis to predict T2D status, with reported accuracies of 89% for women and 86% for men[106,108]. However, a closer examination of the results reveals significant limitations. The voice-only models demonstrated much lower sensitivity (0.58 for women and 0.52 for men), highlighting the limited ability of voice features alone to identify T2D contrary to their claim. Additionally, the study's performance improvements largely relied on incorporating demographic features such as age and BMI, raising concerns about the true contribution of voice features. The study also had a small sample size (n=267), an imbalanced dataset skewed towards participants without T2D (72% of the study population), and collected numerous voice samples from each individual (total of 18,465 recordings), potentially inflating performance metrics and increasing the risk of overfitting.

Implications of the study

Feasibility of Predicting T2D Status Using Vocal Biomarkers

Our study underscores the feasibility of employing vocal biomarkers as an innovative approach for detecting T2D status, contributing significantly to the evolving field of digital health diagnostics. By harnessing sophisticated voice-based algorithms, we demonstrated that a rapid, non-invasive voice recording could effectively predict T2D status, providing a potential user-friendly first-line T2D screening method.

Technical implication: Benchmarking Voice Features

In a novel technical exploration, we compared traditional voice features extracted using the OpenSMILE toolkit (ComParE2016) against advanced hybrid BYOL-S/CvT voice embeddings. This comparison revealed that while traditional acoustic features provide valuable insights, the hybrid embeddings offer superior performance in capturing subtle voice variations influenced by T2D.

Highlighting T2D complexity: Influence of Cofactors on the Voice-Based T2D Status Classifier

Given the "black box" nature of embeddings and pre-trained algorithms, we focused on analyzing clinical data to identify key risk factors and symptoms of T2D that affect voice. This stratification allowed us to capture the complexity of T2D and its interaction with various demographic and physiological factors. The study highlighted how age and comorbidities like hypertension influence the predictive performance of voice AI tools, underscoring the need for a multifaceted approach in developing AI screening tools to enhance specificity, sensitivity, and personalized healthcare.

Sensitivity Analysis of the Voice-Based Classifiers to ADA Risk Score

A critical component of our study was the sensitivity analysis which confirmed a strong agreement of up to 96% between the predictions made by our voice-based algorithms and the ADA risk score. This analysis not only validates the effectiveness of our model but also reinforces the potential of voice-based tools to serve as reliable adjuncts to established at-risk T2D screening methodologies, enhancing early detection and preventive strategies in diabetes care.

Limitations of the study

Limited Sample Size and Diversity with a Focus on Early-Stage T2D Detection

While our study used the largest dataset available for voice-based T2D detection, it is primarily limited by its sample size and the homogeneity of its participants which are mainly English-speaking individuals from the USA self-reporting their T2D status. Furthermore, the variability in T2D duration among our subjects poses significant challenges, particularly affecting the reliability of early-stage diabetes detection. Early detection is crucial as it allows for timely intervention, potentially preventing the progression of the disease and its associated complications. Therefore, targeting individuals with early-stage T2D or prediabetes is vital for enhancing the effectiveness and accuracy of voice-based screening tools.

Expanding the dataset to include a more diverse array of participants from various linguistic and cultural backgrounds is essential. Additionally, focusing on including subjects at the onset of diabetes or prediabetes will provide critical insights into the subtle changes in voice patterns that occur early in the disease's progression. Another future direction is to rely on more objective measurements such as glycemic control or variability parameters, as well as clinical diagnosis of T2D. This strategic emphasis will not only improve the model's diagnostic precision across different populations but also ensure that the tool is robust and universally applicable, thereby making it a reliable component in global diabetes screening and prevention strategies.

Missing T2D Family History and Physical Activity Data for ADA Score Calculation

The Colive Voice study did not collect data on T2D family history and physical activity, which limits our ability to benchmark against the ADA diabetes risk score fully. However, since the ADA score is primarily influenced by age and BMI (both of which are accurately available in our study) the impact of this limitation is somewhat mitigated. Age and BMI contribute significantly to the total score, accounting for up to 45% (5 out of 11 points). In contrast, the combined influence of physical activity and family history is less substantial, contributing less than 20% of the total score. Therefore, the accurate data for age and BMI ensures that the majority of the score's variability is captured. Future work will aim to collect comprehensive data, including physical activity and family history, to further enhance the accuracy of our assessments.

Integration into Voice-Based Screening

The complex interplay of diabetes-related physiological changes and voice quality highlights the potential of voice-based algorithms as non-invasive tools for T2D screening. By leveraging digital health technologies, voice analysis offers a scalable, cost-effective approach that enhances early detection and management, especially in resource-constrained settings. By improving engagement and accessibility, voice-based screening can reach a broader population, potentially detecting more cases and making diabetes care more inclusive and effective.

Conclusion

This study provides promising insights into the use of voice-based algorithms for T2D screening, emphasizing the need for further research and validation to refine these tools for broader clinical and public health applications. The integration of detailed demographic and health-related parameters will be crucial in enhancing the accuracy and effectiveness of these innovative screening methods, potentially transforming the approach to diabetes care globally.

Chapter 6

Vocal Biomarkers for Health Status Monitoring

In this chapter, Objective 3 is explored, focusing on health status monitoring by assessing RQoL in the general population, including participants with various respiratory diseases. The chapter presents the results and discussions from the associated research[109], highlighting the use of vocal biomarkers in monitoring and improving respiratory health.

Results

General workflow

We followed a comprehensive workflow to develop a vocal biomarker for predicting RQoL using data from the Colive Voice study (Figure 19). Participants provided sustained vowel phonation and reading recordings, which were preprocessed to remove noise, convert to mono, trim silences, remove DC offset, and normalize peaks. Socio-demographic data (BMI and smoking habits) and clinical data (night coughing, chest pain, sore throat, and associated diseases such as asthma and COPD) were encoded using one-hot encoding.

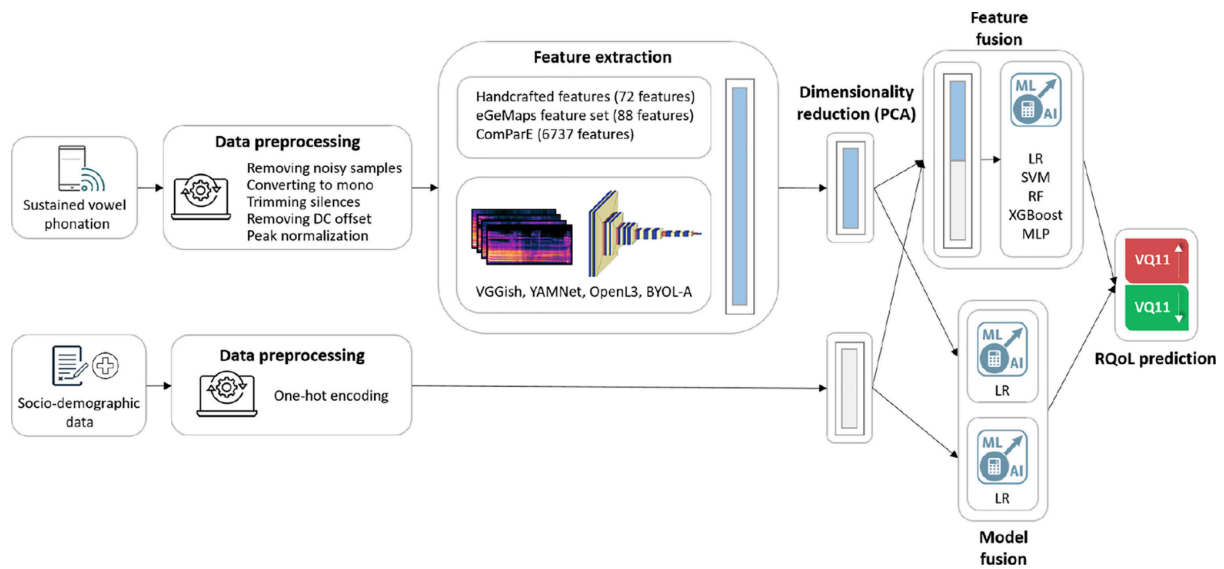


Figure 19. Workflow of RQoL monitoring using early (feature-based) and late (model-based) fusion. (from Despotovic, V., Elbéji, A., et al. Digital voice-based biomarker for monitoring the respiratory quality of life: Findings from the colive voice study.)

Participants were stratified into impaired and normal RQoL groups based on the VQ11 questionnaire scores. To ensure comparability, recordings from the normal RQoL group were matched by age and gender with the impaired RQoL group, creating a balanced subset. Audio features were extracted using a combination of handcrafted features, standard feature sets (eGeMAPS and ComParE), and deep learning-based embeddings (VGGish, YAMNet, OpenL3, and BYOL-A). Dimensionality reduction techniques, such as PCA, were applied to the audio embeddings to reduce their dimensionality and ensure robust generalization. The first 23 principal components were retained to explain most of the variance in the data.

Several machine learning models, including Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and

Multilayer Perceptron (MLP), were trained using the extracted features. The hyperparameters of the models were tuned via grid search to optimize hyperparameters and the number of PCA components. The performance of these models was evaluated using metrics such as accuracy, sensitivity, specificity, area under the receiver operating characteristic curve (AUROC), and Brier score. Stratified 5-fold cross-validation was used to ensure reliable and robust performance estimates.

Both audio and socio-demographic features were integrated into the models. Given the larger size of the audio feature vectors compared to the socio-demographic features, PCA was applied to audio embeddings to equalize the dimensions of the features before fusion. The final model's performance was assessed, and its ability to predict RQoL based on voice recordings and socio-demographic data was validated. The detailed results and implications of the study are discussed in subsequent sections.

Study population characteristics

The analysis included a total of 1908 participants (Table 9), equally divided into two groups (954 participants in each group) based on their RQoL as measured by the VQ11 questionnaire: normal RQoL (VQ11 < 22) and impaired RQoL (VQ11 ≥ 22). The mean VQ11 score was significantly different between the groups, with the normal RQoL group averaging a score of 15±3 and the impaired RQoL group averaging 28.3±6.1 (p < 0.0001). As the groups were matched by age and gender, no significant differences were observed in these variables.

		Total			Normal Respiratory Quality of life (VQ11<22)			Impaired Respiratory Quality of Life (VQ11≥22)			p value
Participants		1908			954 (50%)			954 (50%)			NA
Mean VQ11 score		21.6 (8.2)			15 (3)			28.3 (6.1)			<0.0001
Gender		F	M	O	F	M	O	F	M	O	1
		1280 (67.1%)	608 (31.9%)	20 (1%)	640 (67.1%)	304 (31.9%)	10 (1%)	640 (67.1%)	304 (31.9%)	10 (1%)	
Age		42.4 (14.2)			42.4 (14.1)			42.5 (14.2)			0.948
BMI [kg/m²]	Underweight	66 (3.5%)			35 (3.7%)			31 (3.2%)			<0.0001
	Normal weight	792 (41.5%)			490 (51.3%)			302 (31.7%)			
	Overweight	466 (24.4%)			224 (23.5%)			242 (25.4%)			
	Obesity	601 (30.6%)			205 (21.5%)			379 (39.7%)			
Smoking status	Not at all	1533 (80.4%)			806 (84.5%)			727 (76.2%)			<0.0001
	Less than daily	98 (5.1%)			50 (5.2%)			48 (5%)			
	Daily	277 (14.5%)			98 (10.3%)			179 (18.8%)			
Day coughing	No	1181 (61.9%)			704 (73.8%)			477 (50%)			<0.0001
	Transient	597 (31.3%)			235 (24.6%)			362 (38%)			
	Frequent	130 (6.8%)			15 (1.6%)			115 (12%)			
Night coughing	No	1414 (74.1%)			802 (84%)			612 (64.2%)			<0.0001
	Transient	396 (20.8%)			137 (14.4%)			259 (27.1%)			
	Frequent	98 (5.1%)			15 (1.6%)			83 (8.7%)			
Chest pain	Yes	191 (10%)			43 (4.5%)			148 (15.5%)			<0.0001
Sore throat	Yes	190 (10%)			71 (7.4%)			119 (12.5%)			0.0002
Asthma	Yes	306 (16%)			118 (12.4%)			188 (19.7%)			<0.0001
COPD	Yes	73 (3.8%)			18 (1.9%)			55 (5.8%)			<0.0001

Table 10. Demographic and clinical data of participants with normal and impaired RQoL.

Algorithm performance based on socio-demographic/clinical data

We first evaluated the relevance of socio-demographic and clinical data in predicting RQoL. Using only these data, including BMI, smoking habits, day and night coughing, chest pain,

sore throat, asthma, and COPD, we created a baseline model with 23 features. The best performance was achieved using an LR classifier, yielding an AUROC of 0.70 and an accuracy of 64.1% (Table 11). Feature importance analysis with an RF model identified BMI and coughing symptoms as the most significant predictors.

ML model	Accuracy [%]	Sensitivity [%]	Specificity [%]	AUC	Brier score
LR	64.1 (1.71)	59.64 (3.98)	68.55 (4.86)	0.70 (0.03)	0.22 (0.01)
SVM	62.79 (2.11)	54.4 (6.48)	71.17 (4.74)	0.67 (0.03)	0.23 (0.01)
RF	61.43 (2.02)	53.25 (4.8)	69.59 (5.57)	0.66 (0.03)	0.24 (0.01)
XGBoost	62.26 (2.41)	53.25 (4.7)	71.27 (5.44)	0.66 (0.03)	0.24 (0.01)
MLP	63.84 (2.17)	56.39 (3.94)	71.27 (4.37)	0.70 (0.03)	0.22 (0.01)

Table 11. Results of the prediction models for RQoL status classification based on socio-demographic/clinical data

Algorithm performance based on voice recordings

We furthermore investigated the potential of using voice recordings as digital biomarkers for RQoL. Various sets of audio features were extracted, including handcrafted features, eGeMAPS, ComParE, and deep audio embeddings (VGGish, YAMNet, OpenL3, BYOL-A). PCA was applied to reduce feature dimensionality. The BYOL-A deep audio embeddings provided the best performance, with an AUROC of 0.70 and an accuracy of 65.57%, outperforming other feature extraction methods by over 2%.

Spectrograms of sustained vowel phonation from participants with normal and impaired RQoL highlighted significant differences (Figure 20). Participants with impaired RQoL exhibited interrupted phonation with voice breaks and increased energy in higher frequency bands, indicating aperiodic noise and glottal constriction.

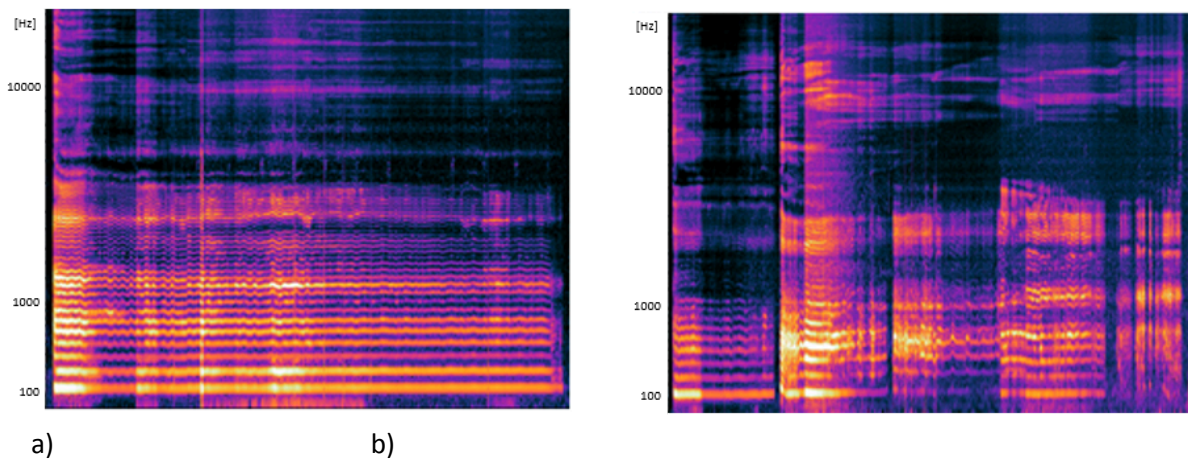


Figure 20. Spectrograms of sustained vowel phonation of participants matched by age and gender (male, age 67), with a) normal RQoL (VQ11 score: 21); and b) impaired RQoL (VQ11 score: 46) (from Despotovic, V., Elbéji, A., et al. Digital voice-based biomarker for monitoring the respiratory quality of life: Findings from the colive voice study.)

Algorithm performance based on fused socio-demographic/clinical data and voice recordings

Combining socio-demographic/clinical data with voice features significantly improved the predictive performance. Using early fusion (feature-based fusion), the best performance was achieved with the combination of BYOL-A audio embeddings and socio-demographic/clinical features, yielding an accuracy of 70.34% and an AUROC of 0.77. This multimodal model also achieved the lowest average Brier score of 0.19, indicating good calibration.

As shown in Figure 21, Specificity was generally higher than sensitivity, meaning the models were better at predicting normal RQoL than impaired RQoL. The confusion matrix for the best-performing model showed a higher number of false negatives compared to false positives. The ROC curve for this model demonstrated its superior performance.

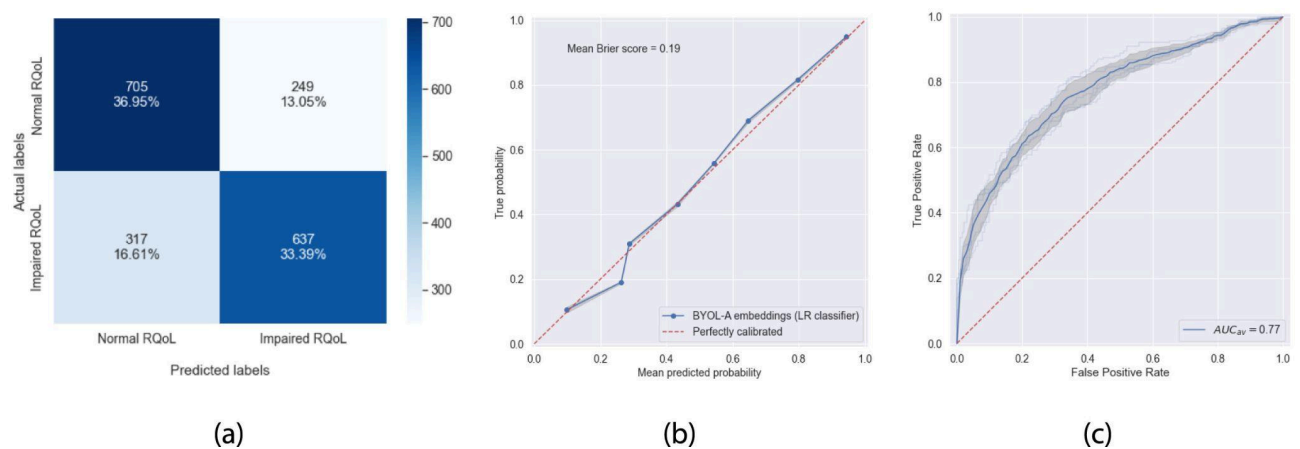


Figure 21. Performance of the best model: (a) Confusion matrix; (b) Probability calibration curve; and (c) ROC curve. (from Despotovic, V., Elbéji, A., et al. Digital voice-based biomarker for monitoring the respiratory quality of life: Findings from the colive voice study.)

The Net Reclassification Improvement (NRI) measure confirmed that adding vocal biomarkers to socio-demographic/clinical data improved predictive capability. The largest NRI improvement of 0.19 was observed for eGeMAPS features modeled with an SVM.

Overall, the integration of voice features with socio-demographic and clinical data enhances the prediction of RQoL, showcasing the potential of multimodal data fusion in improving health outcomes (Figure 22).

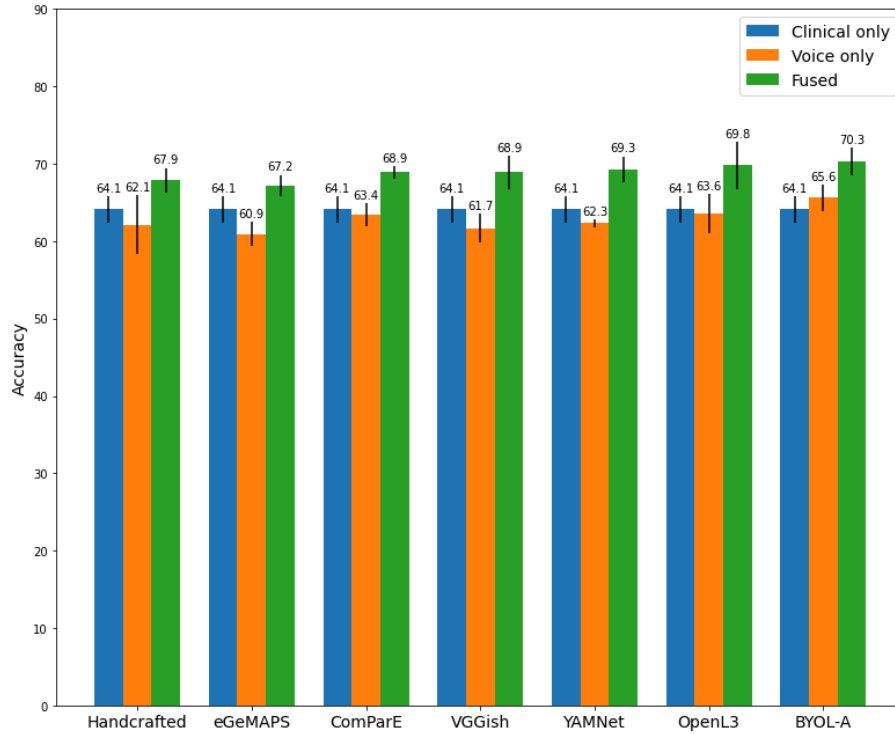


Figure 22. Accuracy with the best-performing machine learning model for socio-demographic/clinical features only, voice features only and fused clinical and voice (multimodal) features. (from Despotovic, V., Elbéji, A., et al. Digital voice-based biomarker for monitoring the respiratory quality of life: Findings from the colive voice study.)

In addition to early fusion techniques, we also explored a more advanced multimodal fusion approach using vector cross-attention, which integrates voice tasks from sustained vowel phonation and reading recordings. This method allowed for better modeling of inter-task relationships, improving prediction accuracy. The vector cross-attention model outperformed early fusion approaches, providing a relative improvement of up to 4.2% in accuracy and yielding more robust vocal biomarkers for RQoL prediction (Table 12).

Modality	Accuracy [%]	Sensitivity [%]	Specificity [%]	AUROC	Runtime
Vowel phonation (/a/)	63.8 (59.0-68.7)	55.4 (48.4-62.3)	72.3 (65.7-78.7)	0.69 (0.64-0.74)	0.2 s/epoch
Reading	61.9 (56.9-66.8)	58.3 (51.2-65.6)	65.5 (58.5-72.0)	0.65 (0.59-0.70)	0.2 s/epoch
Early fusion	65.2 (60.3-70.0)	59.7 (52.8-66.6)	70.7 (64.0-77.0)	0.70 (0.65-0.75)	0.3 s/epoch
Fusion (SHCA)	66.1 (61.3-70.8)	59.5 (52.5-66.5)	72.6 (66.1-78.9)	0.71 (0.66-0.76)	14.6 s/epoch
Fusion (MHCA)	65.4 (60.5-69.9)	60.4 (53.3-67.3)	70.4 (63.7-76.5)	0.70 (0.65-0.75)	17.7 s/epoch

Table 12. Performance metrics of audio modalities and fusion techniques in RQoL prediction. (from Despotovic, V., Elbéji, A., et al. Digital voice-based biomarker for monitoring the respiratory quality of life: Findings from the colive voice study.)

Discussion

Voice Alterations and Underlying Mechanisms in RQoL

The production of voice is directly influenced by the respiratory system. Airflow originates from the lungs and travels through the respiratory tract, passing through the larynx where it causes the vocal folds to vibrate. The sound is then shaped as it moves through the oral and nasal cavities. Respiratory diseases can disrupt this process, leading to noticeable changes in voice. Research has demonstrated that in conditions like COPD[110], inspiratory vocal fold closure, which contributes to refractory breathlessness, is a frequent occurrence. Altered breathing patterns and voice changes are closely linked to diminished lung function in COPD patients[111], likely due to respiratory and muscle impairment[112]. Acoustic features of speech have been shown to differ significantly between periods of COPD exacerbation and stability[113] and can even signal exacerbations up to seven days before symptoms emerge[114], making them a potential early warning tool for COPD.

Similarly, decreased voice-related quality of life, persistent coughing, and laryngeal dysfunction affect up to 88% of patients with severe asthma[115]. Abnormal vocal fold movements, driven by muscle tension in the vocal folds and larynx, further complicate this condition[115]. Vocal signatures derived from voice recordings have shown promise in detecting asthma worsening, offering a non-invasive alternative to traditional lung function measures[116].

Implications of the Study

Feasibility of Using Vocal Biomarkers for RQoL Prediction

This study underscores the feasibility of employing Vocal Biomarkers to monitor RQoL in the general population. Voice recordings offer a non-invasive, easy-to-use method for data collection, facilitating remote patient monitoring without the need for clinical visits. Our results showed that vocal biomarkers outperformed socio-demographic and clinical predictors by approximately 1.5% in accuracy, confirming their potential as a surrogate for traditional clinical measures.

Highlighting the Complementarity of Multimodal Features

The integration of voice features with socio-demographic and clinical data revealed that these modalities provide complementary insights, leading to improved performance. Using early fusion, we achieved a substantial performance boost, with the combined model reaching an accuracy of 70.34% and an AUROC of 0.77. This improvement underscores the value of multimodal data fusion in providing a more robust and comprehensive assessment of RQoL.

Benchmarking Voice Features

Among the various audio features tested, the BYOL-A deep audio embeddings provided the best performance. These general-purpose audio representations, extracted from a self-supervised model trained on extensive audio data, outperformed other features. However, the limited interpretability of deep audio embeddings poses a challenge for clinical applications. Therefore, a trade-off between performance and interpretability must be considered when selecting audio features.

Sensitivity and Calibration of the Multimodal Model

To evaluate not only the accuracy of class label predictions but also the associated probabilities, we used the Brier score. The well-calibrated model, indicated by a solid average Brier score and a nearly linear calibration curve, ensures that the predicted probabilities align closely with the actual outcomes.

Additionally, a more advanced multimodal fusion technique using vector cross-attention was introduced, based on our recent work published at Interspeech 2024. This approach integrates two voice tasks, sustained vowel phonation, and reading, allowing the model to focus on key features across these different modalities. By capturing inter-task relationships, this approach improved the accuracy of RQoL predictions by up to 4.2% compared to single-task models. The cross-attention mechanism, which enables the model to selectively learn from different modalities, demonstrates the potential of using multiple voice tasks to develop more robust and comprehensive vocal biomarkers.

The integration of such advanced fusion techniques can significantly enhance the precision of health status monitoring systems, moving beyond single-modality limitations. Future research should continue exploring these multimodal approaches, as they offer valuable insights into complex health conditions like respiratory diseases. The vector cross-attention model also highlights the growing importance of combining both technical innovation and clinical relevance in developing scalable, accurate, and non-invasive voice biomarkers.

Limitations of the Study

A major strength of this study is the real-world data collection via a mobile app, demonstrating the feasibility of using digital voice-based biomarkers for remote monitoring of RQoL. This approach supports personalized and timely treatment without relying on costly, invasive, or cumbersome equipment.

However, crowdsourced data collection poses challenges, including the risk of low-quality responses and varying audio recording conditions. We mitigated these risks by using a validated questionnaire and a proprietary data processing pipeline to harmonize recordings. Nonetheless, the possibility of some low-quality data cannot be entirely excluded.

The integration of voice AI into healthcare for RQoL monitoring offers significant potential. By leveraging advancements in digital health technologies, voice analysis could enhance the early detection and ongoing management of respiratory conditions. This scalable and cost-effective solution is particularly valuable in resource-constrained settings, supporting remote monitoring and personalized care.

Conclusion

In this study, we developed a digital voice-based biomarker for monitoring RQoL in the general population. Our results confirm that vocal biomarkers can serve as viable surrogates for standard clinical measures from questionnaires, with the highest potential realized in a multimodal setup. The best performance was achieved with a model combining BYOL-A deep audio embeddings and socio-demographic/clinical data, reaching an accuracy of over 70.8% and an AUROC of 0.77. This represents a performance boost of over 5% compared to using voice features alone.

Additionally, we explored an advanced multimodal fusion technique using vector cross-attention, which integrates multiple voice tasks, such as sustained vowel phonation

and reading—by capturing inter-task relationships. This approach further improved the accuracy of RQoL predictions by up to 4.2%, showcasing the power of combining diverse vocal modalities to develop more robust and comprehensive vocal biomarkers.

The proposed approach facilitates rapid screening and represents a step towards scalable, non-invasive, and low-cost solutions for remote monitoring of respiratory health status. Future research should focus on further validating these findings across diverse populations and exploring additional multimodal integrations, such as cross-attention mechanisms, to enhance predictive accuracy and clinical applicability.

Chapter 7

Conclusion and Perspectives

General Findings

Vocal Biomarkers for Symptom Monitoring: Fatigue in People with COVID-19

The first objective of this thesis explores the application of voice analysis for symptom monitoring and detection, focusing specifically on fatigue in individuals with COVID-19. Fatigue is one of the most commonly reported symptoms across various chronic diseases. Effective monitoring and detection of this symptom can significantly enhance the quality of life for those affected.

With a use case among people with COVID-19, vocal biomarkers demonstrated a high level of accuracy in differentiating between fatigued and non-fatigued participants. The use of voice analysis as a non-invasive tool was proven feasible, with models achieving weighted AUC values of up to 86%. This suggests that vocal biomarkers could be integrated into digital devices, such as smartphones, providing a practical method for real-time, remote symptom monitoring. This integration holds the potential to revolutionize telemedicine applications, especially for managing pandemic-related conditions.

Vocal Biomarkers for Screening: T2D

T2D is a growing global health concern, often referred to as a silent epidemic. The prevalence of T2D is projected to increase significantly, with billions of cases expected by 2045. 50% of individuals with T2D remain undiagnosed due to the reliance on invasive, costly, and logistically challenging blood glucose tests. This underscores the urgent need for accessible, non-invasive screening tools.

This thesis investigates the feasibility of using voice analysis to detect T2D status, leveraging data from the Colive Voice study, which included participants from the USA. The study demonstrates that voice, a rich and ecological non-invasive tool, could effectively classify individuals with and without T2D. Performance stratification of the classifiers revealed key subgroups where the predictive models performed optimally.

T2D is a multifaceted disease influenced by various risk factors, comorbidities, lifestyle factors, and symptoms, all of which contribute to unique voice characteristics in affected individuals. The study's findings highlight the potential of voice-based AI tools to serve as a first-line, non-invasive screening method. This could be particularly beneficial in resource-limited environments, offering a scalable and cost-effective solution for early detection and preventive healthcare strategies.

Vocal Biomarkers for Perceived Health Status: RQoL

The third objective of this thesis focuses on assessing perceived health status through the monitoring of RQoL using vocal biomarkers. RQoL is a crucial measure of well-being in individuals with chronic respiratory conditions like COPD and asthma, where early detection of health deterioration can lead to timely interventions and improved outcomes.

Using data from the Colive Voice study, which collected multilingual audio recordings and socio-demographic data, this research demonstrates the potential of vocal biomarkers to predict RQoL in a diverse population. Various audio features, including handcrafted features and deep learning-based embeddings, were analyzed to develop predictive models for RQoL. The study found that the integration of voice features with socio-demographic and clinical data significantly enhanced predictive performance, achieving an accuracy of 70.34% and an AUROC of 0.77.

Furthermore, advanced multimodal fusion techniques were applied, particularly the use of vector cross-attention to combine different voice tasks, such as sustained vowel phonation and reading. This approach allowed the model to capture relationships between different vocal modalities, improving the accuracy of RQoL predictions by up to 4.2% compared to single-task models. The cross-attention mechanism demonstrated the value of leveraging multiple voice tasks to develop more robust and comprehensive vocal biomarkers.

The findings highlight the feasibility of using voice analysis as a non-invasive tool for monitoring RQoL. Vocal biomarkers, when combined with other clinical data, provide a comprehensive and practical approach for continuous, remote monitoring. This approach is particularly valuable for resource-constrained settings, where it offers a scalable, cost-effective solution for respiratory health management. The potential to integrate such models into digital health applications can facilitate rapid screening and personalized care, supporting better management of chronic respiratory diseases.

A common thread across the three objectives of this thesis is the transformative potential of vocal biomarkers as non-invasive, scalable tools for healthcare. Whether applied to symptom monitoring in COVID-19 patients, screening for T2D, or assessing RQoL, the integration of voice analysis into clinical practice holds the promise of improving early detection, continuous monitoring, and personalized care. Each use case highlights the power of voice as a rich source of information, capable of capturing physiological and pathological changes across a range of health conditions. While these applications focus on different diseases and symptoms, they collectively emphasize the potential of vocal biomarkers to revolutionize telemedicine and healthcare, particularly in resource-limited settings where traditional methods may be impractical. Moreover, the findings from all three objectives demonstrate the importance of combining voice data with other health indicators to enhance predictive accuracy and ensure a more comprehensive understanding of an individual's health status. Together, these contributions lay the groundwork for future advancements in digital health, where vocal biomarkers could be integrated into a wide range of telemedicine and healthcare applications.

Perspectives

Research Perspectives

As vocal biomarkers evolve into a promising tool for symptom monitoring, disease screening, and health status assessment, several research perspectives emerge that can drive future advancements across all applications covered in this thesis:

Nuanced Symptom Evaluation and Longitudinal Studies

Across fatigue monitoring, T2D screening, and RQoL assessment, moving beyond binary classifications is essential. In fatigue monitoring, adopting more detailed scales such as the Fatigue Severity Scale (FSS) could provide richer insights into how varying levels of fatigue impact vocal characteristics. Similarly, early-stage detection of T2D would benefit from identifying subtle vocal changes that appear before more pronounced symptoms develop. Longitudinal studies are crucial for understanding the progression of these symptoms or conditions, providing real-time insights that can guide personalized interventions and improve treatment outcomes.

Enhancing Generalizability Through Diverse Populations

Expanding datasets to include participants from diverse geographic, linguistic, and cultural backgrounds is key to ensuring the generalizability of vocal biomarkers. Fatigue, T2D, and RQoL are influenced by various demographic and cultural factors, and validating biomarkers across these variables is essential for global applications. Cross-linguistic and cultural validation will ensure that vocal biomarkers remain effective and reliable regardless of population-specific nuances in vocal characteristics.

Multimodal Data Integration for Comprehensive Health Insights

Combining vocal biomarkers with additional health data, such as physical activity, glycemic control, spirometry results, and environmental factors, will enhance the accuracy of predictive models. This multimodal approach will create a more holistic view of health, improving screening and monitoring outcomes. For instance, integrating comorbidities, lifestyle factors and glycemic control with vocal biomarkers in T2D screening could yield more accurate risk assessments. Similarly, in RQoL monitoring, combining voice data with clinical measurements could improve the management of respiratory conditions.

Balancing Model Performance with Interpretability

While high-performing deep-learning models are central to advancing vocal biomarkers, ensuring that these models are interpretable for clinical use is equally important. Healthcare providers need to understand how these models function to confidently integrate them into clinical workflows. Achieving this balance between accuracy and transparency is key for clinical adoption across all application areas.

Real-World Integration and Scalability

To unlock the full potential of vocal biomarkers, integrating them into widely accessible platforms like smartphones, telehealth systems, and wearables is essential. This will enable continuous, remote monitoring, particularly benefiting resource-limited settings. Tools for fatigue monitoring, T2D screening, and RQoL assessment could be embedded into

user-friendly apps, allowing real-time symptom/condition tracking and personalized health management. Ensuring scalability through such platforms will make these technologies accessible to both patients and healthcare providers, offering cost-effective, non-invasive solutions for widespread health monitoring.

Challenges in Implementing Voice AI in Healthcare

The integration of voice AI in healthcare, while promising, is fraught with several challenges that may limit its broader adoption and effectiveness. One of the most significant issues is the inherent biases that can arise from the data used to train the algorithms, which may not be representative of the diverse patient populations they are meant to serve[117]. This lack of diversity can skew algorithm performance and limit its applicability across different demographic groups. Additionally, ethical challenges emerge, particularly concerning the privacy and security of sensitive health data captured through voice recordings. Handling such data raises concerns about consent and the potential for misuse, which must be navigated carefully.

Furthermore, the field of voice AI suffers from a lack of reproducibility and standardization, with varying methodologies and metrics used across studies, making it difficult to compare results or establish best practices. The absence of clear guidelines exacerbates this issue, leading to inconsistencies in how voice AI tools are developed and validated. Recognizing these challenges, initiatives like Bridge2AI[118], led by Dr. Yael Bensoussan, have emphasized the importance of establishing standardized protocols and developing ethically sourced, diverse voice datasets[119,120]. These collaborative efforts aim to create a large-scale, multimodal voice database that adheres to rigorous standards, facilitating reproducibility and scalability in voice AI research. Transitioning these technologies from research settings into clinical practice presents its own set of hurdles, including regulatory approvals, integration into existing healthcare systems, and acceptance by healthcare professionals and patients. Addressing these challenges through standardized protocols and robust validation frameworks will ensure the efficacy, fairness, and ethical use of voice AI applications in healthcare.

Broader Implications and Future Research Directions

Future research can explore the following directions to maximize the impact of vocal biomarkers in healthcare:

External Validation and Replication Studies

Conducting external validation and replication studies is critical to establish the generalizability, robustness, and trustworthiness of vocal biomarkers. These studies should involve diverse populations in terms of age, gender, language, and socioeconomic distribution to ensure the reliability of findings across different settings.

New Data Collection in Varied Settings

Collecting new data in varied settings and among diverse demographic groups will enhance the generalizability and robustness of the vocal biomarkers. This approach will also help identify any potential biases or limitations in the current models.

Multimodal Approaches

The future of digital health lies in combining various data modalities to create a more comprehensive picture of an individual's health. Beyond voice, incorporating signals from wearable sensors, physiological data, and environmental factors can enhance the precision of health assessments. This multimodal approach provides deeper insights into both chronic and acute conditions, advancing applications such as precision health, hospital-at-home care, and digital clinical trials (Figure 23).

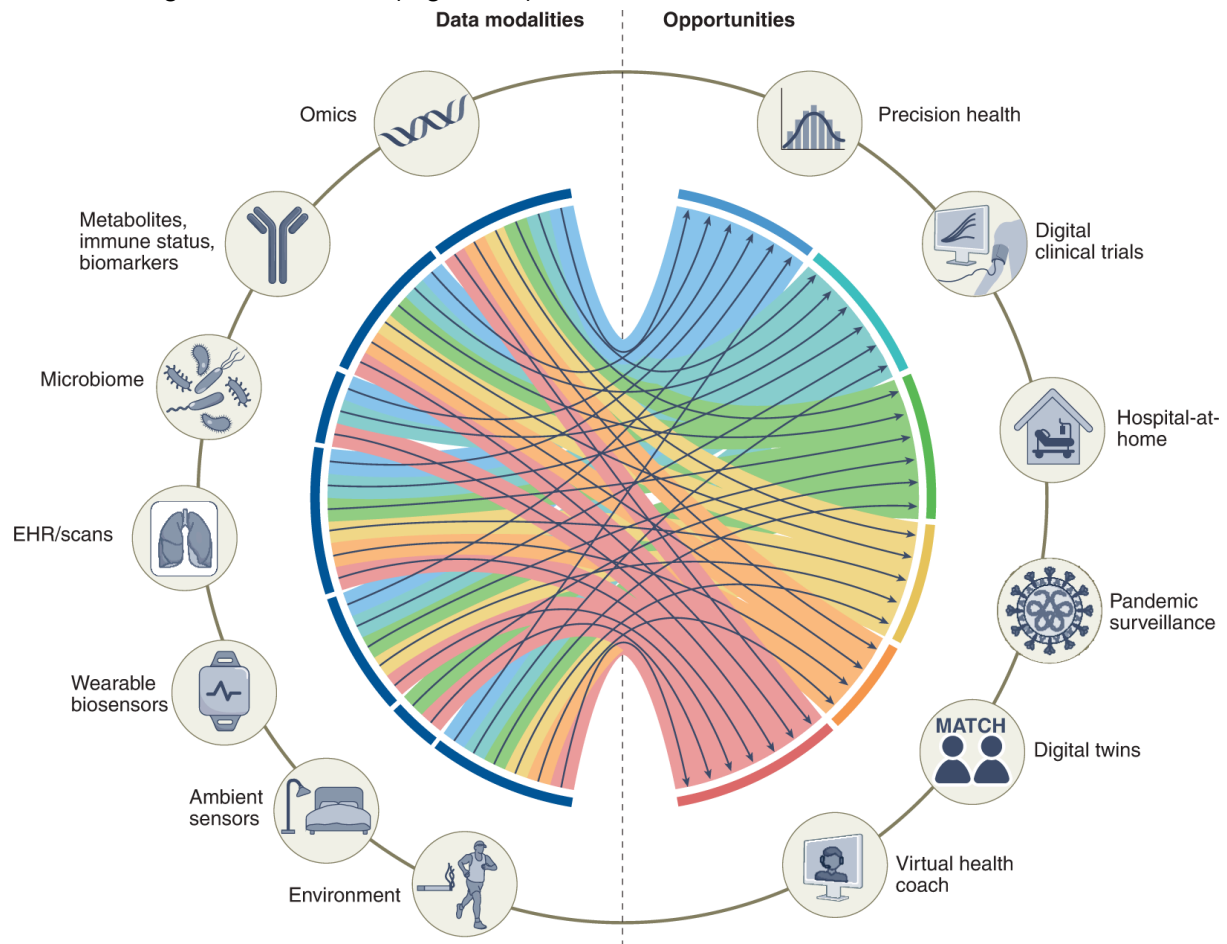


Figure 23. Data modalities and opportunities for multimodal biomedical AI (from Acosta, J.N., Falcone, G.J., Rajpurkar, P. et al. Multimodal biomedical AI. *Nat Med* 28, 1773–1784 (2022)).

Shifting from Traditional Questionnaires to Contextual Voice Recordings

The reliance on self-reported questionnaires for health monitoring limits the potential of voice-based analysis. Future research should explore innovative methods that move away from these traditional approaches. One promising alternative is the use of contextual voice recordings, where participants engage in guided conversations on relevant topics, such as disease management, daily experiences, emotions, and mental states. This method allows for a more natural and dynamic collection of vocal data, potentially revealing subtler and more meaningful biomarkers related to a person's health status. By capturing real-world vocal expressions, we can better harness the full potential of vocal biomarkers for real-time, non-invasive health monitoring.

Vocal Biomarkers Replication and Validation

Replicating and validating vocal biomarkers in different populations and settings is crucial to ensure their effectiveness and reliability. To improve reproducibility across diverse environments, it is essential to establish international standards and guidelines for the development and deployment of these biomarkers. Defining these standards will help build trust in vocal biomarker technologies and facilitate their integration into mainstream healthcare, ensuring consistent and reliable outcomes worldwide.

Personalized Health Monitoring

Developing personalized health monitoring systems that adapt to an individual's unique characteristics and health conditions can improve the effectiveness of digital health interventions. Tailoring these systems to account for personal variability will enhance patient outcomes.

Ethical and Privacy Considerations

Addressing ethical and privacy concerns related to the use of vocal biomarkers is essential. Ensuring that data collection, storage, and analysis comply with ethical standards and privacy regulations will foster trust and acceptance among users.

Collaboration with Healthcare Providers

Collaborating with healthcare providers to integrate vocal biomarkers into clinical workflows can enhance their practical utility. Engaging healthcare professionals in the development and implementation process will ensure that these tools meet clinical needs and standards.

Funding and Policy Support

Securing funding and policy support for research on vocal biomarkers is vital for advancing this field. Encouraging investment in digital health technologies and creating supportive policies can accelerate the development and adoption of these innovative tools.

Final reflections

As research on vocal biomarkers progresses, it becomes increasingly evident that voice AI has the potential to revolutionize healthcare by providing non-invasive, accessible means to monitor a wide range of health conditions. This potential is especially promising in the context of telehealth, where continuous, real-time patient monitoring can enhance care delivery, reduce the burden on healthcare systems, and offer solutions in situations where traditional methods may not be feasible or accessible.

However, while the promise of vocal biomarkers is significant, we are still in the early stages of this journey. There is a long path ahead with numerous challenges to refine these technologies, ensure their accuracy, and integrate them safely into clinical practice. The incorporation of AI and machine learning in health technology presents unique challenges, particularly in maintaining robustness, and accuracy, and addressing ethical considerations within clinical settings.

Ensuring the generalizability and robustness of vocal biomarkers across diverse populations is therefore critical. This will require external validation studies, the development of international standards and guidelines, and collaboration with healthcare providers to integrate these tools into clinical workflows. Establishing such standards, alongside fostering interdisciplinary partnerships, will be vital in building trust and promoting the widespread adoption of vocal biomarkers in healthcare.

To advance these technologies further, one key future direction will involve moving beyond traditional questionnaires, which often limit the potential of voice-based analysis. Contextual voice recordings, where individuals engage in guided conversations about their daily experiences, disease management, emotions, and mental states, could reveal more nuanced biomarkers, offering a richer understanding of health status. When combined with multimodal approaches that incorporate additional health data, these methods can provide a more precise and comprehensive understanding of chronic and acute conditions.

I believe this work lays a strong foundation for future research, contributing to the development of voice AI and its applications in healthcare. As we continue to innovate and address the challenges ahead, I am optimistic that these efforts will lead to more precise, reliable, and accessible use of vocal biomarkers. Ultimately, the goal is to improve patient outcomes and make healthcare more personalized and responsive to individual needs, transforming how we approach health and disease management.

Chapter 8

Contributions

This chapter outlines the significant contributions made during my doctoral studies, spanning various dimensions of research, development, and education. It details my active participation in international conferences, including both oral and poster presentations, and highlights my involvement in the TAILOR-funded exchange program[121], which enriched my research experience. Furthermore, it discusses how my research and technical developments in voice processing enabled the preparation, submission, and selection of a research grant to valorize my work into a Luxembourg Institute of Health (LIH) spin-off. Additionally, this chapter reflects on my contributions to related work and my role in the academic growth of master's students through co-supervision.

Vocalive Platform for Vocal Biomarker Development

Project background

I contributed significantly to the development of an innovative platform aimed at enhancing the adoption of vocal biomarkers in healthcare. My primary contribution involved building a robust pipeline to pre-process and process real-life voice recordings, thereby increasing the reliability and robustness of vocal biomarkers in real-world settings. This involved developing algorithms and methods to handle varying audio qualities, extract meaningful features, and optimize the data for machine learning applications. This work forms the foundation of Vocalive, a platform designed to leverage IP-protected technology to integrate vocal biomarkers into healthcare. The project has gained significant traction and has been supported by funding from the JUMP program facilitated by the National Research Fund (FNR) in Luxembourg[121,122]. This effort exemplifies technology transfer and research valorization, paving the way for Vocalive to become a spin-off company of the Luxembourg Institute of Health (LIH).

Overview of Vocalive

Vocalive is a pioneering service platform designed for the development and deployment of vocal biomarkers in healthcare. Using a proprietary voice processing algorithm, Vocalive enables the use of voice recordings collected in non-controlled environments ("in the wild"), a critical advancement over traditional methods that rely on highly standardized clinical settings. This approach enhances the practical application of vocal biomarkers, making them more adaptable for real-life patient monitoring, such as through telemonitoring solutions and smartphone applications.

Services Provided by Vocalive

Vocalive offers a range of services tailored to improve the AI readiness of voice audio data and support the development of custom vocal biomarkers:

Standard Automated Services: These include professional voice processing and audio quality control, designed to enhance the quality of input audio data for third-party biomarker algorithms.

Custom Services for Pharma and CROs: Vocalive provides comprehensive, end-to-end services for developing vocal biomarkers tailored to specific health outcomes. This includes

everything from data collection to the final analysis.

External Validation and Other Custom Services: Leveraging an existing database of voice and labeled medical data, Vocalive offers external validation services, custom data analyses, and tools for data collection and audio processing.

Vocal Biomarker Marketplace: In addition to service offerings, Vocalive plans to launch an agnostic marketplace for evidence-based vocal biomarkers. This marketplace will feature ready-to-use vocal biomarkers with detailed transparency cards that outline their intended use, performance specifics, applicable languages, and limitations, supported by peer-reviewed publications. This will facilitate easy access to validated biomarkers for various stakeholders, promoting broader adoption and integration into clinical practices.

Workflow and Technical Aspects

The Vocalive workflow includes:

Real-World Data Collection: Capturing voice recordings via apps on mobile devices, with storage in WAV format.

Audio Preprocessing: Harmonizing recordings of varying audio quality to prepare them for biomarker development, including feature extraction and machine learning modeling.

Audio Quality Control: Assessing and ensuring the quality of preprocessed audio, with detailed reports on the data quality and reasons for the exclusion of any recordings.

Health Data Integration: Collecting and preprocessing clinical data and patient-reported outcomes, which can be integrated with audio features for comprehensive model training.

Future Outlook and Impact

Vocalive is positioned to transform the landscape of vocal biomarker development by making it more accessible, efficient, and applicable in everyday health management. The platform's emphasis on real-world data application, combined with its comprehensive service offerings and forthcoming marketplace, underscores its potential to enhance preventive healthcare and personalized medicine strategies.

Participation in the TAILOR-Funded Exchange Program

During my PhD program, my research was greatly enriched by my participation in the TAILOR-Funded Exchange Program, which facilitated a research visit to the Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) in Germany. This collaboration enabled me to work closely with experts in AI and healthcare, particularly focusing on the development of advanced machine learning techniques, such as multitask learning, to enhance the detection and monitoring of depressive symptoms using vocal biomarkers.

Contribution to Depression Detection and Monitoring

Depression affects approximately 4.4% of the global population, necessitating early and accurate detection to reduce its long-term impact. Traditional methods, like self-reported questionnaires, often lack objectivity and accuracy, leading to a growing interest in voice-based biomarkers as an alternative. During my research at DFKI, I aimed to improve the accuracy of these biomarkers by exploring the relationship between depressive symptoms and other correlated symptoms, such as fatigue. By leveraging multitask learning, we sought to enhance the model's ability to generalize across different languages and demographics, using the large multilingual Colive Voice dataset. This approach aimed to refine the prediction of depressive symptoms by incorporating multiple related health indicators, thereby increasing the robustness and reliability of voice-based models.

Multitask Learning

Multitask learning is an approach where a single model is simultaneously trained on multiple related tasks[123]. This methodology leverages commonalities across tasks, leading to improved learning efficiency and prediction accuracy compared to training separate models for each task. It is particularly effective when the tasks share underlying features but have different outputs.

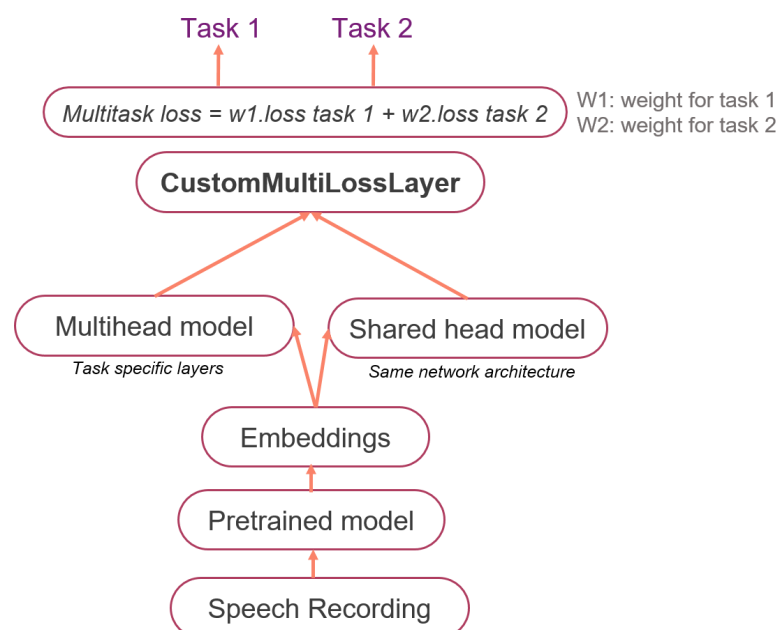


Figure 24. Multitask learning framework for speech analysis

In the context of speech analysis, this workflow starts with speech recordings processed through a pre-trained model to extract embeddings. These embeddings are then fed into either a multi-head model, which has separate heads for each task, or a shared head model, which uses a single head for all tasks (Figure 24). A custom multi-loss layer combines the losses from multiple tasks into a single scalar value, weighted to determine the relative importance of each task's loss. This setup allows for the simultaneous training of models on multiple voice-related tasks, leveraging shared information to enhance overall performance and efficiency.

Methodological Advancements

My research at DFKI involved developing neural network models trained on real-life voice recordings to predict depressive symptoms, initially achieving an accuracy of 65%. To improve performance, I integrated additional symptoms like fatigue and stratified the data by gender and language to minimize biases. Using advanced feature extraction techniques like BYOL-S and OpenSMILE, I conducted regression tasks on PHQ-9 and FSS scores to establish a ground truth. This work faced challenges, such as data imbalance and distribution issues, which we addressed by focusing on binary classification tasks and using machine learning algorithms better suited for limited sample sizes. We further explored the Net Reclassification Index (NRI) to assess how the integration of one classifier's output could enhance another's performance, thus advancing our understanding of model interactions and dependencies.

Collaborative Research and Knowledge Sharing

The collaborative environment at DFKI was instrumental in guiding my work, particularly in handling complex, high-dimensional voice data and addressing data imbalance in AI models. This experience not only improved my technical skills but also fostered a deeper understanding of AI trustworthiness and explainable algorithms, which is crucial when working with sensitive mental health data. The collaboration facilitated the exchange of best practices and set the foundation for future joint research efforts, including the development of a collaborative research paper to document our findings and methodologies.

Societal and Technical Impact

The research conducted during this exchange has significant societal and technical implications. By improving the accuracy and generalizability of voice-based depressive symptom detection models through multitask learning, we contribute to more reliable, non-invasive screening methods that can be widely implemented, particularly in regions with limited access to mental health professionals. Our work emphasizes the need for inclusive, language-independent models to ensure fair and accessible mental health diagnostics globally. Moreover, the technical advancements in incorporating additional symptoms and refining AI models contribute to the field's progress in developing trustworthy and explainable AI algorithms.

Future Directions

Building on this work, future efforts will focus on expanding the dataset to include more diverse language groups to enhance model generalizability, addressing data imbalance issues to improve predictive accuracy, and integrating additional symptoms to refine our multitask learning approach. Continued collaboration with DFKI and other TAILOR partners

will be key to further advancing the development of scalable, AI-driven healthcare solutions that can improve early diagnosis and patient outcomes in mental healthcare.

Contribution to related work

During my PhD, I have also contributed to related work through the co-supervision of three master's students, each for six months. My role involved providing them with comprehensive guidance and support throughout their research projects. This included holding weekly meetings to track their progress, offering methodological and technical assistance to ensure the successful execution of their work, and advising them on the preparation of their final reports and defenses. My involvement in their projects aimed to foster their academic development while ensuring they achieved the objectives set for their respective studies.

-Hugo Allemand:

Title: Identification of a digital vocal biomarker to predict Long COVID.

Abstract: COVID-19 is a complex disease that can lead to remaining symptoms after the infection phase. This condition is called Long COVID and can take various forms, last from several months to more than a year, and deeply affect the daily lives of people. The study aimed to develop a digital vocal biomarker that could predict whether a person will develop Long COVID, based on data collected from Predi-COVID study participants during the COVID-19 acute phase. We used data from 181 participants who were followed up regularly during 12 months. Based on the smartphone's operating system, we created several models and tested different sets of features using cross-validation. We obtained an accuracy of 66% for iOS users and 74% for Android users. These preliminary findings support our initial hypothesis but need to be completed with further analyses using more data.

-Léo Gerard:

Title: Voice-Based Health Mapping: Exploring Vocal Biomarkers for Health Conditions in the General Population.

Abstract: In this work, we used DDRTree, an unsupervised machine learning algorithm, to do a mapping of the health characteristics of the general population, and used it to compare the voice characteristics of groups of people with different health conditions such as COVID-19, diabetes, cancer, or respiratory diseases, to study the links between voice and health. The dataset used contained the answers to a questionnaire regarding various health conditions and symptoms and the audio features extracted from the audio recordings of the Colive Voice study, composed of 4569 participants. We trained two trees, one using male participants and one using female participants. For both trees, the horizontal axis was associated with the general health status of the participants, and the vertical axis was associated with their smoking habits. Moreover, for the one using exclusively female participants, we observed that cardiovascular diseases and their risk factors were also associated with their vertical axis. Our results established links between these conditions and the different extracted audio features and may be the first step in the elaboration of a new vocal biomarker estimating the health status of a patient.

-Maëlle Cornec:

Title: Vocal Biomarker for Fatigue Severity Monitoring.

Abstract: This study aims to develop voice-based digital biomarkers for detecting and monitoring fatigue severity, as measured by the Fatigue Severity Scale (FSS) score, using real-world data from the Colive Voice study, involving 4,232 participants. Given the complexity and variability of fatigue as a symptom across different health conditions, our

approach combines traditional audio feature extraction with advanced machine learning techniques. We employed the OpenSMILE toolkit to extract various acoustic features and analyzed their correlations with fatigue levels, revealing that certain voice features are significantly correlated with fatigue, with differences observed across gender and language groups. To enhance classification performance, we explored the application of transfer learning models (YAMNet, VGGish, and FRILL) to generate embeddings that capture deeper voice characteristics. Models trained on these embeddings demonstrated potential in classifying fatigue levels, with varying performance across different classifiers and datasets. Our findings suggest that using specific vocal features and transfer learning-based embeddings can effectively discriminate fatigue levels, particularly in female participants. The study highlights the potential of using voice as a non-invasive and scalable tool for remote fatigue monitoring, contributing to precision health strategies and advancing digital health solutions.

Conferences

Throughout my PhD, I have actively participated in various prestigious conferences, presenting my research to the international academic community. These engagements have provided opportunities for receiving feedback, networking with peers, and staying updated with the latest advancements in the field. Below is a summary of my conference participation:

Oral communications

SFD 2022 (*Congrès annuel de la Société Francophone du Diabète*), Nice-France: *Identification de biomarqueurs vocaux pour le dépistage et le suivi des personnes avec diabète: premiers résultats de l'étude Colive Voice.*

ML seminar 2024, Online-Luxembourg: Exploring Voice AI: Technologies, Algorithms, and Applications in Healthcare.

EASD 2024 (*European Association for the Study of Diabetes Annual Meeting*), Madrid-Spain: Towards Non-Invasive And Accessible Type 2 Diabetes Screening Through Voice AI Technology.

Poster presentations

ATTD 2022 (*Advanced Technologies and Treatments for Diabetes International Conference*), Online: Identification of vocal biomarkers for screening diabetes and monitoring the health of people with diabetes: preliminary results from the Colive Voice study.

ADA 2024 (*American Diabetes Association Scientific Sessions*), Orlando-USA: Towards Non-Invasive And Accessible Type 2 Diabetes Screening Through Voice AI Technology.

PhD days 2024, Luxembourg: Towards Non-Invasive And Accessible Type 2 Diabetes Screening Through Voice AI Technology.

Voice2AI 2024 (Awarded Best Poster for the topic "Voice Biomarkers Help Patient Outcomes and Access"), Tampa-USA: Towards Non-Invasive And Accessible Type 2 Diabetes Screening Through Voice AI Technology.

Interspeech 2024, Kos-Greece: Multimodal Fusion for Vocal Biomarkers Using Vector Cross-Attention.

References

1. Kraus VB. Biomarkers as drug development tools: discovery, validation, qualification and use. *Nat Rev Rheumatol*. 2018;14. doi:10.1038/s41584-018-0005-9
2. Fagherazzi G, Fischer A, Betsou F, Vaillant M, Ernens I, Masi S, et al. Protocol for a prospective, longitudinal cohort of people with COVID-19 and their household members to study factors associated with disease severity: the Predi-COVID study. *BMJ Open*. 2020;10: e041834.
3. Umapathy K, Ghoraani B, Krishnan S. Audio Signal Processing Using Time-Frequency Approaches: Coding, Classification, Fingerprinting, and Watermarking. *EURASIP J Adv Signal Process*. 2010;2010: 1–28.
4. Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*. 2020;5: 96–116.
5. A review of depression and suicide risk assessment using speech analysis. *Speech Commun*. 2015;71: 10–49.
6. Jeon H, Jung Y, Lee S, Jung Y. Area-Efficient Short-Time Fourier Transform Processor for Time–Frequency Analysis of Non-Stationary Signals. *NATO Adv Sci Inst Ser E Appl Sci*. 2020;10: 7208.
7. librosa: Audio and Music Signal Analysis in Python. [cited 10 Oct 2024]. Available: <https://proceedings.scipy.org/articles/Majora-7b98e3ed-003>
8. Opensmile. [cited 10 Oct 2024]. doi:10.1145/1873951.1874246
9. [Interspeech 2016 ComParE] Computational Paralinguistics Challenge - First Sub-Challenge Open. [cited 10 Oct 2024]. Available: <https://groups.google.com/g/ml-news/c/g3Qab0agy4k/m/wtQkE0e8BwAJ>
10. Eyben F, Scherer KR, Schuller BW, Sundberg J, André E, Busso C, et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. [cited 10 Oct 2024]. Available: <https://ieeexplore.ieee.org/document/7160715>
11. models/research/audioset/vggish at master · tensorflow/models. In: GitHub [Internet]. [cited 10 Oct 2024]. Available: <https://github.com/tensorflow/models/tree/master/research/audioset/vggish>
12. models/research/audioset/yamnet at master · tensorflow/models. In: GitHub [Internet]. [cited 10 Oct 2024]. Available: <https://github.com/tensorflow/models/tree/master/research/audioset/yamnet>
13. Cramer AL, Wu H-H, Salamon J, Bello JP. Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings. [cited 10 Oct 2024]. Available: <https://ieeexplore.ieee.org/document/8682475>
14. Niizumi D, Takeuchi D, Ohishi Y, Harada N, Kashino K. BYOL for Audio: Self-Supervised Learning for General-Purpose Audio Representation. [cited 10 Oct 2024]. Available: <http://dx.doi.org/10.1109/IJCNN52387.2021.9534474>

15. Elbanna G, Biryukov A, Scheidwasser-Clow N, Orlandic L, Mainar P, Kegler M, et al. Hybrid Handcrafted and Learnable Audio Representation for Analysis of Speech Under Cognitive and Physical Load. 2022 [cited 10 Oct 2024]. doi:10.21437/Interspeech.2022-10498
16. Baeovski A, Zhou H, Mohamed A, Auli M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. 2020. Available: <http://arxiv.org/abs/2006.11477>
17. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014. Available: <http://arxiv.org/abs/1409.1556>
18. Rana R, Singhal R. Chi-square Test and its Application in Hypothesis Testing. *Journal of the Practice of Cardiovascular Sciences*. 2015;1: 69.
19. := TQMP.ORG =: [cited 10 Oct 2024]. Available: <https://www.tqmp.org/RegularArticles/vol04-1/p013/>
20. := TQMP.ORG =: [cited 10 Oct 2024]. Available: <https://www.tqmp.org/RegularArticles/vol04-1/p013/>
21. := TQMP.ORG =: [cited 10 Oct 2024]. Available: <https://www.tqmp.org/RegularArticles/vol04-1/p013/>
22. Ostertagová E, Ostertag O. Methodology and Application of Oneway ANOVA. *American Journal of Mechanical Engineering*. 2013;1: 256–261.
23. Mansournia MA, Waters R, Nazemipour M, Bland M, Altman DG. Bland-Altman methods for comparing methods of measurement and response to criticisms. *Global Epidemiology*. 2021;3. doi:10.1016/j.gloepi.2020.100045
24. StandardScaler. In: scikit-learn [Internet]. [cited 10 Oct 2024]. Available: <https://scikit-learn.org/dev/modules/generated/sklearn.preprocessing.StandardScaler.html>
25. SelectKBest. In: scikit-learn [Internet]. [cited 10 Oct 2024]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html
26. Principal components analysis (PCA). *Comput Geosci*. 1993;19: 303–342.
27. Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*. 2021;2: 1–21.
28. 1. Supervised learning. In: scikit-learn [Internet]. [cited 10 Oct 2024]. Available: https://scikit-learn.org/stable/supervised_learning.html
29. Menard S. Logistic Regression: From Introductory to Advanced Concepts and Applications. SAGE; 2010.
30. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. [cited 10 Oct 2024]. Available: <https://ieeexplore.ieee.org/document/708428>
31. Breiman L. Random Forests. *Mach Learn*. 2001;45: 5–32.
32. Cunningham Pádraig. k-Nearest Neighbour Classifiers - A Tutorial. *ACM Computing Surveys (CSUR)*. 2021 [cited 10 Oct 2024]. doi:10.1145/3459665

33. Leon F, Floria S-A, Bădică C. Evaluating the effect of voting methods on ensemble-based classification. [cited 10 Oct 2024]. Available: <https://ieeexplore.ieee.org/document/8001122>
34. XGBoost. [cited 10 Oct 2024]. doi:10.1145/2939672.2939785
35. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521: 436–444.
36. Cross-Validation. 2019; 542–545.
37. StratifiedKFold. In: scikit-learn [Internet]. [cited 10 Oct 2024]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html
38. McKinney W. pandas: a Foundational Python Library for Data Analysis and Statistics. 2011 [cited 10 Oct 2024]. Available: <https://www.semanticscholar.org/paper/pandas%3A-a-Foundational-Python-Library-for-Data-and-McKinney/1a62eb61b2663f8135347171e30cb9dc0a8931b5>
39. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. *Nature*. 2020;585: 357–362.
40. Varoquaux Gaël, Duchesnay Édouard. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011 [cited 10 Oct 2024]. doi:10.5555/1953048.2078195
41. TensorFlow: A system for large-scale machine learning. [cited 10 Oct 2024]. Available: <http://research.google/pubs/tensorflow-a-system-for-large-scale-machine-learning/>
42. Joseph FJJ, Nonsiri S, Monsakul A. Keras and TensorFlow: A Hands-On Experience. *Advanced Deep Learning for Engineers and Scientists*. 2021; 85–111.
43. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. 2019. Available: <http://arxiv.org/abs/1912.01703>
44. Digital health. [cited 10 Oct 2024]. Available: https://www.who.int/europe/health-topics/digital-health#tab=tab_1
45. Digital health. [cited 10 Oct 2024]. Available: https://www.who.int/europe/health-topics/digital-health#tab=tab_1
46. Fagherazzi G, Goetzinger C, Rashid MA, Aguayo GA, Huiart L. Digital Health Strategies to Fight COVID-19 Worldwide: Challenges, Recommendations, and a Call for Papers. *J Med Internet Res*. 2020;22: e19284.
47. WHO guideline Recommendations on Digital Interventions for Health System Strengthening. 2019 [cited 10 Oct 2024]. Available: <https://pubmed.ncbi.nlm.nih.gov/31162915/>
48. Frank SR. Digital Health Care—The Convergence of Health Care and the Internet. *J Ambul Care Manage*. 2000;23: 8.
49. Eysenbach G. What is e-health? *J Med Internet Res*. 2001;3: e833.
50. Rosenlund M, Kinnunen U-M, Saranto K. The Use of Digital Health Services Among Patients and Citizens Living at Home: Scoping Review. *J Med Internet Res*. 2023;25: e44711.

51. Butcher CJT, Hussain W. Digital healthcare: the future. *Future Healthcare Journal*. 2022;9: 113.
52. Digital diabetes: Perspectives for diabetes prevention, management and research. *Diabetes Metab*. 2019;45: 322–329.
53. Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions—A call for innovative evidence generation approaches. *npj Digital Medicine*. 2020;3: 1–14.
54. FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) Resource. 2016 [cited 10 Oct 2024]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK326791/>
55. Zhang Z. Mechanics of human voice production and control. *J Acoust Soc Am*. 2016;140: 2614.
56. Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods*. 2018;151: 41–54.
57. Bio-acoustic features of depression: A review. *Biomed Signal Process Control*. 2023;85: 105020.
58. 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.
59. Hecker P, Steckhan N, Eyben F, Schuller BW, Arnrich B. Voice Analysis for Neurological Disorder Recognition—A Systematic Review and Perspective on Emerging Trends. *Front Digit Health*. 2022;4: 842301.
60. Gaikwad P, Venkatesan M. Speech Recognition-Based Prediction for Mental Health and Depression: A Review. *Proceedings of Congress on Control, Robotics, and Mechatronics*. 2024; 13–24.
61. Computerized analysis of speech and voice for Parkinson’s disease: A systematic review. *Comput Methods Programs Biomed*. 2022;226: 107133.
62. Voice Disorder. [cited 10 Oct 2024]. doi:10.1016/j.pmr.2006.06.004
63. Tansey CM, Louie M, Loeb M, Gold WL, Muller MP, de Jager J, et al. One-Year Outcomes and Health Care Utilization in Survivors of Severe Acute Respiratory Syndrome. *Arch Intern Med*. 2007;167: 1312–1320.
64. Qi R, Chen W, Liu S, Thompson PM, Zhang LJ, Xia F, et al. Psychological morbidities and fatigue in patients with confirmed COVID-19 during disease outbreak: prevalence and associated biopsychosocial risk factors. *medRxiv*. doi:10.1101/2020.05.08.20031666
65. Norton C, Czuber-Dochan W, Bassett P, Berliner S, Bredin F, Darvell M, et al. Assessing fatigue in inflammatory bowel disease: comparison of three fatigue scales. *Aliment Pharmacol Ther*. 2015;42. doi:10.1111/apt.13255
66. Abd-Elfattah HM, Abdelazeim FH, Elshennawy S. Physical and cognitive consequences of fatigue: A review. *Journal of Advanced Research*. 2015;6: 351.
67. Elsaïs A, Wyller VB, Loge JH, Kerty E. Fatigue in myasthenia gravis: is it more than muscular weakness? *BMC Neurol*. 2013;13. doi:10.1186/1471-2377-13-132

68. DeLuca J, Genova HM, Hillary FG, Wylie G. Neural correlates of cognitive fatigue in multiple sclerosis using functional MRI. *J Neurol Sci.* 2008;270. doi:10.1016/j.jns.2008.01.018
69. Kratz AL, Schilling S, Goesling J, Williams DA. The PROMIS FatigueFM Profile: a self-report measure of fatigue for use in fibromyalgia. *Qual Life Res.* 2016;25. doi:10.1007/s11136-016-1230-9
70. Morgul E, Bener A, Atak M, Akyel S, Aktaş S, Bhugra D, et al. COVID-19 pandemic and psychological fatigue in Turkey. *Int J Soc Psychiatry.* 2021;67. doi:10.1177/0020764020941889
71. Arnold P, Njemini R, Vantieghem S, Duchateau J, Mets T, Beyer I, et al. Peripheral muscle fatigue in hospitalised geriatric patients is associated with circulating markers of inflammation. *Exp Gerontol.* 2017;95. doi:10.1016/j.exger.2017.05.007
72. Mendelson M, Nel J, Blumberg L, Madhi SA, Dryden M, Stevens W, et al. Long-COVID: An evolving problem with an extensive impact. *S Afr Med J.* 2020;111. doi:10.7196/SAMJ.2020.v111i11.15433
73. Crook H, Raza S, Nowell J, Young M, Edison P. Long covid-mechanisms, risk factors, and management. *BMJ.* 2021;374. doi:10.1136/bmj.n1648
74. Carfi A, Bernabei R, Landi F. Persistent Symptoms in Patients After Acute COVID-19. *JAMA.* 2020;324: 603–605.
75. Halpin SJ, McIvor C, Whyatt G, Adams A, Harvey O, McLean L, et al. Postdischarge symptoms and rehabilitation needs in survivors of COVID-19 infection: A cross-sectional evaluation. *J Med Virol.* 2021;93. doi:10.1002/jmv.26368
76. Townsend L, Dyer AH, Jones K, Dunne J, Mooney A, Gaffney F, et al. Persistent fatigue following SARS-CoV-2 infection is common and independent of severity of initial infection. *PLoS One.* 2020;15. doi:10.1371/journal.pone.0240784
77. Fischer A, Zhang L, Elbéji A, Wilmes P, Oustric P, Staub T, et al. Long COVID Symptomatology After 12 Months and Its Impact on Quality of Life According to Initial Coronavirus Disease 2019 Disease Severity. *Open Forum Infectious Diseases.* 2022;9. doi:10.1093/ofid/ofac397
78. American Diabetes Association. Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care.* 2010;33: S62.
79. Galicia-Garcia U, Benito-Vicente A, Jebari S, Larrea-Sebal A, Siddiqi H, Uribe KB, et al. Pathophysiology of Type 2 Diabetes Mellitus. *Int J Mol Sci.* 2020;21. doi:10.3390/ijms21176275
80. Ogurtsova K, Guariguata L, Barengo NC, Ruiz PL, Sacre JW, Karuranga S, et al. IDF diabetes Atlas: Global estimates of undiagnosed diabetes in adults for 2021. *Diabetes Res Clin Pract.* 2022;183. doi:10.1016/j.diabres.2021.109118
81. Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants. *Lancet.* 2016;387. doi:10.1016/S0140-6736(16)00618-8
82. Lindström J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes Care.* 2003;26. doi:10.2337/diacare.26.3.725

83. Jones PW, Quirk FH, Baveystock CM, Littlejohns P. A Self-complete Measure of Health Status for Chronic Airflow Limitation: The St. George's Respiratory Questionnaire. *Am Rev Respir Dis*. 2012 [cited 11 Oct 2024]. doi:10.1164/ajrccm/145.6.1321
84. Chauvin A, Rupley L, Meyers K, Johnson K, Eason J. Outcomes in Cardiopulmonary Physical Therapy: Chronic Respiratory Disease Questionnaire (CRQ). *Cardiopulm Phys Ther J*. 2008;19: 61.
85. Hyland ME, Singh SJ, Sodergren SC, Morgan MPL. Development of a Shortened Version of the Breathing Problems Questionnaire Suitable for Use in a Pulmonary Rehabilitation Clinic: A Purpose-Specific, Disease-Specific Questionnaire. *Qual Life Res*. 1998;7: 227–233.
86. Le VQ11, un questionnaire de qualité de vie spécifique à la BPCO utilisable en clinique. *Rev Mal Respir*. 2010;27: 472–481.
87. Elbéji A, Zhang L, Higa E, Fischer A, Despotovic V, Nazarov PV, et al. Vocal biomarker predicts fatigue in people with COVID-19: results from the prospective Predi-COVID cohort study. *BMJ Open*. 2022;12: e062463.
88. A density-based algorithm for discovering clusters in large spatial databases with noise. [cited 11 Oct 2024]. doi:10.5555/3001460.3001507
89. Speech during sustained operations. *Speech Commun*. 1996;20: 55–70.
90. Greeley HP, Berg J, Friets E, Wilson J, Greenough G, Picone J, et al. Fatigue estimation using voice analysis. *Behav Res Methods*. 2007;39: 610–619.
91. Vocal Signs and Symptoms Related to COVID-19 and Risk Factors for their Persistence. *J Voice*. 2024;38: 189–194.
92. National Institute of General Medical Sciences. In: National Institute of General Medical Sciences (NIGMS) [Internet]. [cited 11 Oct 2024]. Available: <https://nigms.nih.gov/>
93. Post COVID-19 condition (Long COVID). [cited 11 Oct 2024]. Available: <https://www.who.int/europe/news-room/fact-sheets/item/post-covid-19-condition>
94. Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS). In: Mayo Clinic [Internet]. [cited 11 Oct 2024]. Available: <https://www.mayoclinic.org/diseases-conditions/chronic-fatigue-syndrome/symptoms-causes/syc-20360490>
95. Komaroff AL, Lipkin WI. ME/CFS and Long COVID share similar symptoms and biological abnormalities: road map to the literature. *Front Med*. 2023;10: 1187163.
96. Differences in Self-Rated, Perceived, and Acoustic Voice Qualities Between High- and Low-Fatigue Groups. *J Voice*. 2011;25: 544–552.
97. Hunter SK. Sex differences in human fatigability: mechanisms and insight to physiological responses. *Acta Physiol* . 2014;210: 768–789.
98. Lerdal A. Fatigue Severity Scale. *Encyclopedia of Quality of Life and Well-Being Research*. 2014; 2218–2221.
99. Swain MG. Fatigue in chronic disease. *Clin Sci* . 2000;99. Available: <https://pubmed.ncbi.nlm.nih.gov/10887052/>

100. Zimek D, Miklusova M, Mares J. Overview of the Current Pathophysiology of Fatigue in Multiple Sclerosis, Its Diagnosis and Treatment Options – Review Article. *Neuropsychiatr Dis Treat*. 2023;19: 2485.
101. Weber D, O'Brien K. Cancer and Cancer-Related Fatigue and the Interrelationships With Depression, Stress, and Inflammation. *J Evid Based Complementary Altern Med*. 2017;22: 502.
102. Fischer A, Elbeji A, Aguayo G, Fagherazzi G. Recommendations for Successful Implementation of the Use of Vocal Biomarkers for Remote Monitoring of COVID-19 and Long COVID in Clinical Practice and Research. *Interact J Med Res*. 2022;11. doi:10.2196/40655
103. Patel K, Horak H, Tiryaki E. Diabetic neuropathies. *Muscle Nerve*. 2021;63. doi:10.1002/mus.27014
104. Gölaç H, Atalik G, Türkcan AK, Yilmaz M. Disease related changes in vocal parameters of patients with type 2 diabetes mellitus. *Logoped Phoniatr Vocol*. 2022;47. doi:10.1080/14015439.2021.1917653
105. Hamdan AL, Jabbour J, Nassar J, Dahouk I, Azar ST. Vocal characteristics in patients with type 2 diabetes mellitus. *Eur Arch Otorhinolaryngol*. 2012;269. doi:10.1007/s00405-012-1933-7
106. Pinyopodjanard S, Suppakitjanusant P, Lomprew P, Kasemkosin N, Chailurkit L, Ongphiphadhanakul B. Instrumental Acoustic Voice Characteristics in Adults with Type 2 Diabetes. *J Voice*. 2021;35. doi:10.1016/j.jvoice.2019.07.003
107. Hamdan AL, Kurban Z, Azar ST. Prevalence of phonatory symptoms in patients with type 2 diabetes mellitus. *Acta Diabetol*. 2013;50. doi:10.1007/s00592-012-0392-3
108. Kaufman JM, Thommandram A, Fossat Y. Acoustic Analysis and Prediction of Type 2 Diabetes Mellitus Using Smartphone-Recorded Voice Segments. *Mayo Clinic Proceedings: Digital Health*. 2023;1: 534–544.
109. Digital voice-based biomarker for monitoring respiratory quality of life: findings from the colive voice study. *Biomed Signal Process Control*. 2024;96: 106555.
110. Leong P, Ruane LE, Phyland D, Koh J, MacDonald MI, Baxter M, et al. Inspiratory vocal cord closure in COPD. *Eur Respir J*. 2020;55. doi:10.1183/13993003.01466-2019
111. Website. Available: [https://journal.chestnet.org/article/S0012-3692\(21\)03371-7/fulltext](https://journal.chestnet.org/article/S0012-3692(21)03371-7/fulltext)
112. Quality of Life Predictors in Voice of Individuals With Chronic Obstructive Pulmonary Disease. *J Voice*. 2022 [cited 11 Oct 2024]. doi:10.1016/j.jvoice.2022.05.017
113. Nallanthighal VS, Härmä A, Strik H. Detection of COPD Exacerbation from Speech: Comparison of Acoustic Features and Deep Learning Based Speech Breathing Models. [cited 11 Oct 2024]. Available: <https://ieeexplore.ieee.org/document/9747785>
114. VOICE-BASED MONITORING OF COPD. *Chest*. 2021;160: A2173–A2174.
115. Vertigan AE, Kapela SL, Gibson PG. Laryngeal Dysfunction in Severe Asthma: A Cross-Sectional Observational Study. *J Allergy Clin Immunol Pract*. 2021;9. doi:10.1016/j.jaip.2020.09.034

116. 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.
117. Gorla S, Polle R, Fara S, Cummins N. Revealing confounding biases: A novel benchmarking approach for aggregate-level performance metrics in health assessments. *Interspeech 2024*. ISCA: ISCA; 2024. pp. 1440–1444.
118. Bridge2AI - Voice. [cited 15 Oct 2024]. Available: <https://www.b2ai-voice.org/>
119. Evangelista E, Kale R, McCutcheon D, Rameau A, Gelbard A, Powell M, et al. Current Practices in Voice Data Collection and Limitations to Voice AI Research: A National Survey. *Laryngoscope*. 2024;134: 1333–1339.
120. Bensoussan Y, Ghosh S, Rameau A, Boyer M, Bahr R, Watts S, et al. Developing Multi-Disorder Voice Protocols: A team science approach involving clinical expertise, bioethics, standards, and DEI. *Interspeech 2024*. ISCA: ISCA; 2024. pp. 1445–1449.
121. CEF Collaboration Exchange Fund. In: TAILOR [Internet]. EU project TAILOR; 24 Nov 2022 [cited 11 Oct 2024]. Available: <https://tailor-network.eu/cef/>
122. JUMP. In: FNR [Internet]. 28 Feb 2017 [cited 11 Oct 2024]. Available: <https://www.fnr.lu/funding-instruments/jump/>
123. Zhang Y, Yang Q. An overview of multi-task learning. *Natl Sci Rev*. 2017;5: 30–43.

Appendix of Original Papers

Publication Cover Sheet (Objective 1)

Title of the Publication:

Vocal biomarker predicts fatigue in people with COVID-19: results from the prospective Predi-COVID cohort study.

Authors:

Abir Elbéji, Guy Fagherazzi, Zhang, Higa, Aurélie Fischer, Vladimir Despotovic, Petr V. Nazarov, Gloria Aguayo

Data and Figure Contributions:

Abir Elbéji was responsible for the generation and analysis of all figures in the manuscript.

Contribution to Manuscript Writing:

Elbéji led the drafting of the manuscript, including substantial contributions to the Methods section, where the study design and analysis techniques were described. Additionally, Elbéji wrote the Results section, detailing the machine learning models and performance, and contributed to the Discussion and Conclusion sections.

Statement of Contributions:

- Elbéji and Fagherazzi had full access to all the data and ensured the integrity and accuracy of the data analysis.
- Fagherazzi, Zhang, and Fischer conceptualized and designed the study.
- Elbéji, Zhang, Higa, Fischer, Despotovic, Nazarov, and Aguayo contributed to data collection, analysis, and interpretation.
- The statistical analysis was carried out by Elbéji, Zhang, Higa, and Fischer.
- Elbéji drafted the initial manuscript, while all authors provided significant revisions.
- Fagherazzi secured the funding for the project.
- Fischer provided administrative and technical support.
- All authors critically reviewed the manuscript and approved the final version.

Abir Elbéjji¹, Lu Zhang², Eduardo Higa¹, Aurélie Fischer¹, Vladimir Despotovic², Petr V. Nazarov², Gloria A. Aguayo¹, Guy Fagherazzi¹

- ### Reprints and correspondence:

Website: <https://ddp.lih.lu/>

Abstract

Objective

To develop a vocal biomarker for fatigue monitoring in people with COVID-19.

Design Prospective cohort study.

Setting Predi-COVID data between May 2020 and May 2021.

Participants

A total of 1772 voice recordings was used to train an AI-based algorithm to predict fatigue, stratified by gender and smartphone's operating system (Android/iOS). The recordings were collected from 296 participants tracked for two weeks following SARS-CoV-2 infection.

primary and secondary outcome measures

Four machine learning algorithms (Logistic regression, k-nearest neighbors, support vector machine, and soft voting classifier) were used to train and derive the fatigue vocal biomarker. The models were evaluated based on the following metrics: Area Under the ROC curve (AUC), accuracy, F1-score, precision, and recall. The Brier score was also used to evaluate the models' calibrations.

Results

The final study population included 56% of women and had a mean (\pm SD) age of 40 (\pm 13) years. Women were more likely to report fatigue ($P<.001$). We developed four models for Android female, Android male, iOS female, and iOS male users with a weighted AUC of 86%, 82%, 79%, 85%, and a mean Brier Score of 0.15, 0.12, 0.17, 0.12, respectively. The vocal biomarker derived from the prediction models successfully discriminated COVID-19 participants with and without fatigue.

Conclusions

This study demonstrates the feasibility of identifying and remotely monitoring fatigue thanks to voice. Vocal biomarkers, digitally integrated into telemedicine technologies, are expected to improve the monitoring of people with COVID-19 or Long-COVID.

ClinicalTrials.gov Identifier: NCT04380987

Strengths and limitations

- This is the first study supporting the hypothesis that fatigue can be accurately monitored based on voice in people with COVID-19.
- The analyses were based on a multilingual database of standardized voice recordings collected in real-life from people with confirmed SARS-CoV-2 infection as determined by PCR.
- There is no similar dataset available yet in the literature to replicate our findings.
- The vocal biomarker is trained on a binary outcome (Fatigue, Yes/No) and does not reflect the entire spectrum of fatigue severity. Further work should be performed in that direction.

Introduction

Coronavirus disease 2019 (COVID-19) is a global outbreak. More than 199 million confirmed cases of COVID-19 have been detected worldwide as of 4 August 2021, with more than 4 million deaths reported by the World Health Organization¹. The worldwide population and healthcare systems have been greatly impacted by the COVID-19 pandemic. The pandemic has essentially put whole healthcare systems under pressure, requiring national or regional lockdowns². Finding solutions that allow healthcare providers to focus on the more important and urgent patients, was, and still is, critical.

This outbreak continues to impact people, with many patients suffering from a range of acute symptoms, such as fatigue. Fatigue is a common symptom in patients with COVID-19 that can impact their quality of life, treatment adherence, and can be associated with numerous complications³. Recent findings showed that fatigue is a major symptom of the frequently reported Long-COVID syndrome. After recovering from the acute disease caused by the SARS outbreak, up to 60% of patients reported chronic fatigue 12 months later⁴. This supports the need for long-term monitoring solutions for these patients.

In general, fatigue can be of two types: physical and mental⁵ experiencing lack of energy, inability to start and perform everyday activities, and lack of desire to do things. In the context of COVID-19, determinants of fatigue were categorized as both central and psychological factors, the latest might also be indirectly caused by pandemic-related fear and anxiety^{6,7}.

Fatigue affects men and women differently and has previously been shown to be reported differently in the two genders. Men and women have different anatomy and physiology, resulting in significant sex differences in fatigability⁸.

Telemedicine, artificial intelligence (AI), and big data predictive analytics are examples of digital health technologies that have the potential to minimize the damaging effects of COVID-19 by improving responses to public health problems at a population level⁹. Using telemonitoring technologies to enable self-surveillance and remote monitoring of symptoms might therefore help to improve and personalize COVID-19 care delivery¹⁰.

Voice is a promising source of digital data since it is rich, user-friendly, inexpensive to collect, and non-invasive, and can be used to develop vocal biomarkers that characterize disease states. Previous research was mostly conducted in the field of neurodegenerative diseases, such as Parkinson's disease¹¹ and Alzheimer's disease¹². There are also studies that confirm the relation of voice disorders to fatigue, e.g., in Chronic Fatigue Syndrome (CFS). Neuromuscular, neuropsychological and hormonal dysfunction associated with CFS can influence the phonation and articulation, and alter tension, viscosity and thickness of the tissue of the larynx, tongue and lips, leading to decreased voice quality¹³. Increased fatigue affects voice characteristics, such as pitch, word duration¹⁴ and timing of articulated sounds¹⁵. Vocal changes related to fatigue are more observed in consonant sounds that require a high average airflow¹⁶.

In the context of the COVID-19 pandemic, respiratory sounds (e.g coughs, breathing, and voice) are also used as sources of information to develop COVID-19 screening tools^{17,18,19}. However, no previous work has been devoted to investigating the association of voice with COVID-19 symptoms.

We hypothesized that there is an association between fatigue and voice in patients with COVID-19 and that it is possible to train an AI-based model to identify fatigue and subsequently generate a digital vocal biomarker for fatigue monitoring. We used data from the large hybrid prospective Predi-COVID cohort study to investigate this hypothesis.

Methods

Study design

This project uses data from the Predi-COVID study²⁰. Predi-COVID is a hybrid cohort study that started in May 2020 in Luxembourg and involved participants who should meet all of the following requirements: (1) a signed informed consent form; (2) participants with confirmed SARS-CoV-2 infection as determined by PCR at one of Luxembourg's certified laboratories; and (3) 18 years and older.

This study combines data from the national surveillance system, which is used for virtually all COVID-19 positive patients. Biological sampling, electronic patient-

reported outcomes, and smartphone voice recording were collected to identify vocal biomarkers of respiratory syndromes and fatigue in this study. More details about the Predi-COVID study can be found elsewhere²⁰.

Health Inspection collaborators made the initial phone contact with potential participants. Those who consented to participate were contacted by a qualified nurse from the Clinical and Epidemiological Investigation Center (CIEC - Luxembourg Institute of Health), who outlined the study and arranged home or hospital visits.

Patient and Public Involvement

The Predi-COVID initiative was an emergency response from national research institutions grouped under 'Research Luxembourg' to fight the COVID-19 pandemic in Luxembourg and contribute to the general effort in the crisis. Therefore, for timing and safety reasons, patients with COVID-19 were not directly included to participate in the study design. However, the first participants included in Predi-COVID provided feedback on general workflow, data collection, questionnaires, and sampling, which was taken into account in an amendment to the protocol²⁰.

Data collection

Participants were followed for up to a year using a smartphone app to collect voice data. To ensure a minimum quality level, participants were asked to record it in a quiet environment while maintaining a certain distance from the microphone, and an audio example of what was required was also provided.

All the participants of this study were invited to record two audio types. The first, Type 1 audio, required participants to read paragraph 1 of article 25 of the Declaration of Human Rights²¹, in their preferred language: French, German, English, or Portuguese; and the second, Type 2 audio, required them to hold the [a] vowel phonation without breathing for as long as they could (see Supplementary Online Material 1 for more details).

Predi-COVID collects data in conformity with the German Society of Epidemiology's best practices guidelines²². To draft the manuscript, we followed the TRIPOD criteria

for reporting AI-based model development and validation, as well as the corresponding checklist.

All Predi-COVID participants recruited between May 2020 and May 2021 who reported their fatigue status (“I feel well” as “No Fatigue” and “I am fatigued”/“I don’t feel well” as “Fatigue”) on the same day as the audio recordings during the 14 days of follow-up were included in this study²³. As a result, several audio recordings for a single participant were available for both audio types²⁴.

Audio characteristics and vocal biomarker training

The audio recordings were collected in two formats, 3gp format (Android devices) and m4a format (iOS devices). Based on the smartphone’s operating system and the user’s gender (male/female), we trained one model for each category. This stratification was performed to minimize data heterogeneity and deal with sex as a potential confounding bias.

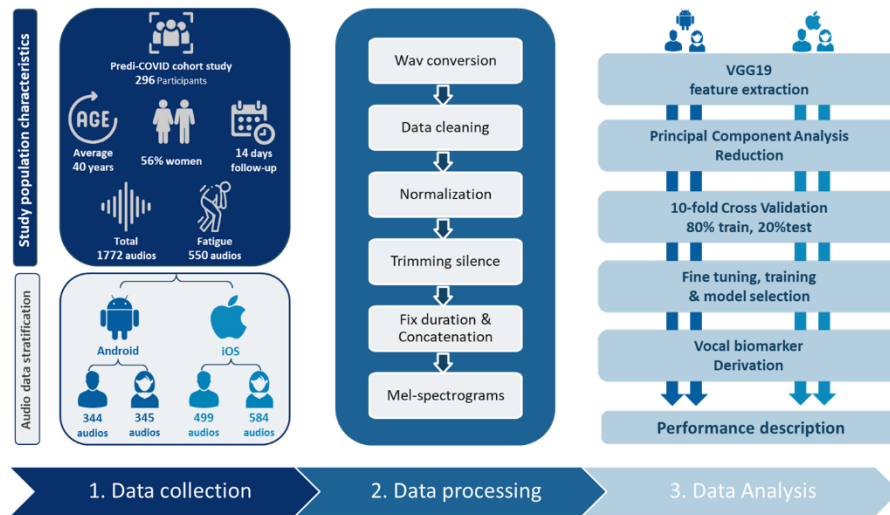


Figure 1. General Pipeline

- (1) Data collection from Predi-COVID and stratification. (2) Audio data processing and mel-spectrogram creation. (3) Data analysis for both male and female users of each device was performed in parallel.

Audio pre-processing

All of the raw audio recordings were pre-processed (Figure 1). They were initially converted to .wav files, with audios lasting less than 2 seconds being excluded. Then, an audio clustering (DBSCAN) on basic features (duration, average, sum, and standard deviation of signal power, and fundamental frequency) was performed to detect outliers that were manually checked while excluding poor quality audios with 1)

too noisy, 2) incorrect text reading, 3) type 1 and type 2 audios mixed, or 4) extended silence in the middle. Finally, peak normalization was used to boost the volume of quiet audio segments, and leading and trailing silences longer than 350 milliseconds were trimmed.

Feature extraction

We used transfer learning for the feature extraction process since it is adapted for small training databases²⁵. Transfer learning is a technique where a model is constructed and trained with a set containing a large amount of data and then transfer and apply this learning to our dataset on top of it. It has the advantage of reducing the amount of data required while shortening training time and improving performance when compared to models built from scratch²⁶.

Convolutional neural networks require a fixed input size, whereas audio instances in our dataset were of variable length. To deal with this issue, Zero-padding was used to set the duration of each audio file to 50 seconds (the maximum length in our database). To raise the amount of information fed to the classifiers, type 1 (text reading) and type 2 ([a] phonation) audios were concatenated and used as a single input to the learning models.

All the audio recordings were first resampled to 8kHz and then converted to Mel-spectrograms using the Librosa library in Python. The hop-length was 2048 samples, and the number of Mel coefficients was set to 196. The Mel spectrograms were passed through VGG19 convolutional neural network architecture provided by Keras, which was pre-trained on the ImageNet database²⁷. This approach, presented in Figure 2, may be considered as a feature extraction step, as it converts audio recordings to 512 feature maps, each of a size 6x6, leading to a total of 18432 features.

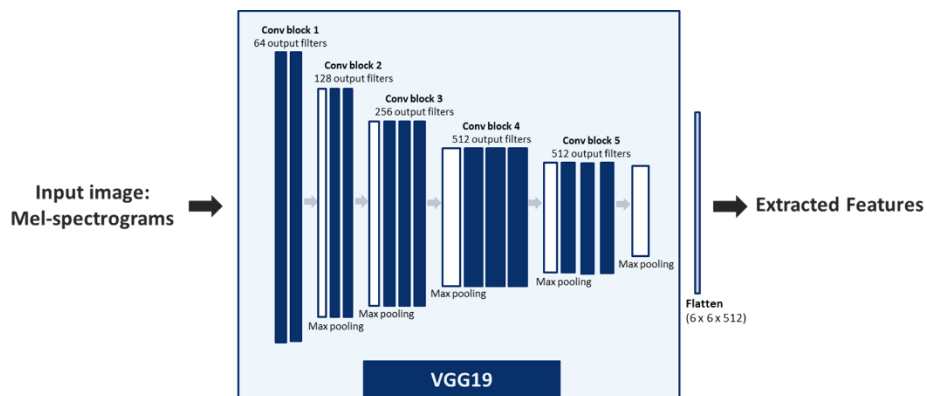


Figure 2. VGG19 Feature extraction

The Vgg19 pre-trained model's architecture for extracting features from mel-spectrograms. The extracted features are used as input to machine learning algorithms.

This large number of features is computationally expensive. Principal Component Analysis (PCA)²⁸ is therefore used for dimensionality reduction and to select the number of relevant components explaining the maximum of the variance in the data.

Statistical analysis

We divided our data into “Fatigue” and “No Fatigue” groups based on the participant’s reported answers for the inclusion and daily fatigue assessment of Predi-COVID. To characterize participants, descriptive statistics were used, which included means, standard deviations for quantitative variables, and counts and percentages for qualitative variables. The two population groups (3gp (Android users) and m4a (iOS users)) were compared using a student test for continuous variables, and a χ^2 test for categorical variables.

A 10-fold cross-validation procedure was conducted on the training cohort participants to evaluate four classification models (logistic regression (LR), k-nearest neighbors (KNN), support vector machine (SVM), and soft voting classifier (VC), scikit-learn implementation in Python) at different regularization levels via a grid search, with the following evaluation metrics: area under the ROC curve (AUC), accuracy, F1-score, precision, and recall. The Brier score was also used to evaluate the calibration of the selected models.

The predicted probability of being classified as fatigued from the best model was considered as our final vocal biomarker, which may be used as a quantitative metric to monitor fatigue.

Results

Study population characteristics

The final study population is composed of 296 participants of whom 165 were women (56%), with an average age of 40 years (SD = 13). To record both audio types, 109 (37%) participants utilized Android smartphones (3gp format), whereas 187 (63%) used iOS devices (m4a format). We found no difference in the distribution of age, gender, body mass index, smoking, antibiotic usage, and asthma, between the two types of devices ($P\text{-value} > .05$). The overall rate of comorbidities in this study was relatively low: there were 31 (10%) participants who used antibiotics and only 12 (4%) participants with asthma. More details are shown in Table 1.

Table 1: Study population characteristics

The clinical data in the table above describe the overall population of the study. The total number and its percentage are used to represent all categorical data. The table below summarizes general information for describing audio data.

All p-values comparing iOS (m4a) and Android users (3gp) were calculated using chi2 and Student's t-tests.

		All	m4a		3gp		P-values (m4a, 3gp)
			Female	Male	Female	Male	
Participants (N)	Total	296	107	80	51	58	-
Age (years)	mean (SD)	40.3 (12.6)	38.8 (13.4)	42.9 (12.7)	37.8 (11.6)	41.5 (11.3)	0.28
Body Mass Index (kg/m²)	mean (SD)	24.1 (4.7)	24.6 (5.5)	26.5 (4.1)	24.1 (3.8)	26.6 (4.17)	0.95
Antibiotic (%)	No	265 (90%)	93 (87%)	73 (91%)	44 (86%)	55 (95%)	0.87
	Yes	31 (10%)	14 (13%)	7 (9%)	7 (14%)	3 (5%)	
Asthma (%)	No	284 (96%)	104 (97%)	75 (94%)	47 (92%)	58 (100%)	0.82
	Yes	12 (4%)	3 (3%)	5 (6%)	4 (8%)	0 (0%)	
Smoking (%)	Never	199 (67%)	77 (72%)	51 (64%)	36 (71%)	35 (60%)	0.41
	Former smoker	53 (18%)	19 (18%)	20 (25%)	9 (18%)	13 (22%)	
	Current smoker	44 (15%)	11 (10%)	9 (11%)	6 (11%)	10 (18%)	
Audio recordings	Total	1772	584	499	345	344	<0.001
	No Fatigue	1222 (69%)	394 (67%)	370 (74%)	190 (55%)	268 (78%)	
	Fatigue	550 (31%)	190 (33%)	129 (26%)	155 (45%)	76 (22%)	
Mean (SD) and maximum of audio recording per participant in the 14-day follow-up period	mean (SD)	6 (5)	6 (5)	6 (5)	6 (5)	6 (5)	-
	max	16	14	16	15	14	

Participants reported their fatigue status on average 6 days during the first 14 days of follow-up, resulting in the analysis of 1772 audio recordings for each audio type (type

1 and type 2) when all inclusion criteria were met, including 550 audio recordings for participants with fatigue. In both audio sets, women reported experiencing fatigue at a higher rate than men ($P\text{-value} < .001$). Women constituted 155 (60%) of all fatigued Android users and 190 (67%) of all fatigued iOS users.

Prediction models

We reduced the extracted features from Mel-spectrograms to 250 top components with PCA, explaining 97% and 99% of the variance in the data for iOS and Android audio sets respectively. We then compared the performances of the machine learning algorithms to select the best models for the derivation of the vocal biomarkers.

The voting classifier was the best model selected for the development of the vocal biomarker for male iOS users, with an AUC of 85% and overall accuracy, precision, recall, and f1-score of 89%. The model selected for female iOS users was SVM with an overall precision of 79% and an AUC of 79%. For male Android users, the selected model is the voting classifier with precision, recall an f1-score of 84%, and a weighted AUC of 82%. For female Android users, the SVM was selected with an overall precision of 80% and an AUC of 86%. More details are shown in Table 2.

As shown in Figure 3, the calibrations of the selected models were good (Mean Brier Scores = 0.15, 0.12, 0.17, and 0.12 respectively for Android female users, Android male users, iOS female users, and iOS male users).

Derivation of the digital fatigue vocal biomarker

Based on the model selected for each audio set, we derived the trained vocal biomarkers which quantitatively represent the probability of being labeled as fatigued.

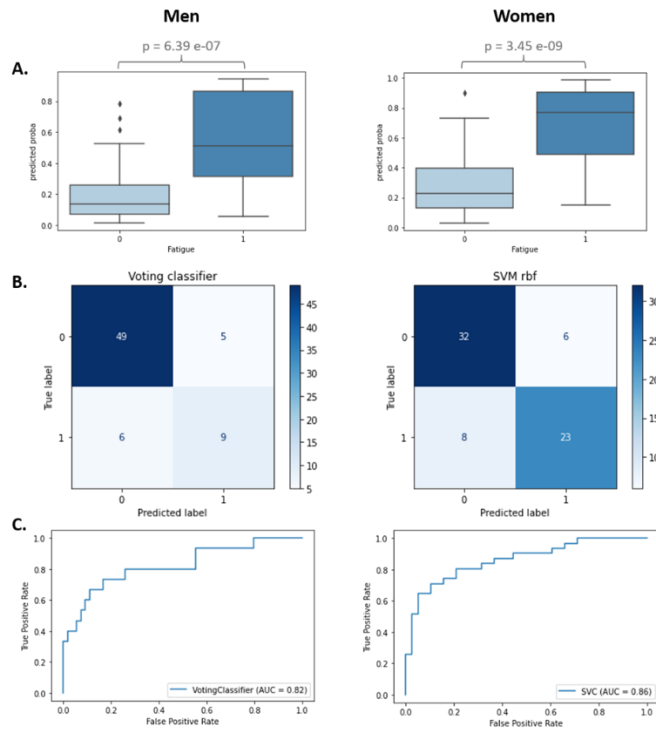


Figure 3a. Derivation of the digital fatigue vocal biomarker for Android users

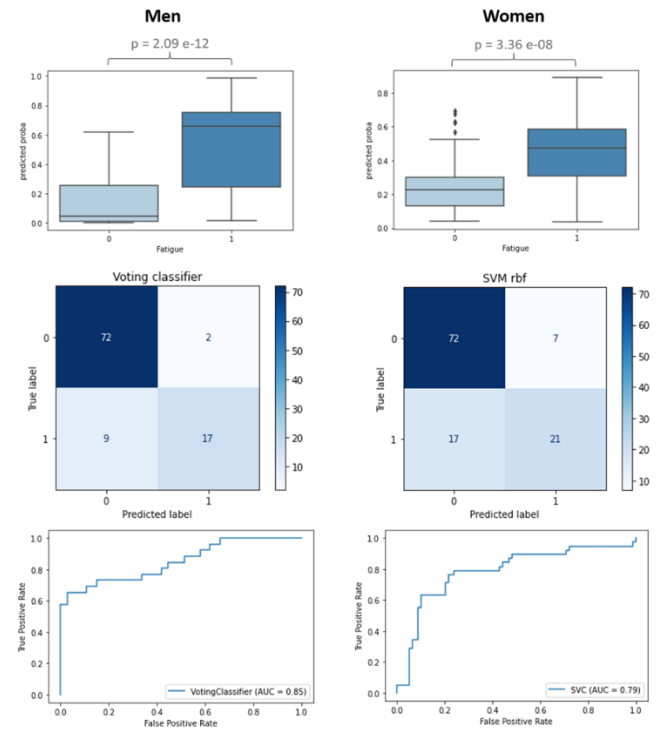


Figure 3b. Derivation of the digital fatigue vocal biomarker for iOS users

Table 2: Results of the prediction models

The selected models were selected using Recall_1 and weighted AUC and are highlighted in bold. Class 0: No fatigue, Class 1: Fatigue

Audio_format	Gender	ML model	Accuracy	Ov.Precision	Precision_0	Precision_1	Ov.Recall	Recall_0	Recall_1	Ov.f1score	f1-score_0	f1-score_1	Weighted AUC
3gp (Android)	Female	LR	0.77	0.77	0.81	0.73	0.77	0.76	0.77	0.77	0.78	0.75	0.85
		KNN	0.72	0.73	0.70	0.77	0.72	0.87	0.55	0.72	0.78	0.64	0.76
		SVM	0.80	0.80	0.80	0.79	0.80	0.84	0.74	0.80	0.82	0.77	0.86
		VC	0.78	0.78	0.81	0.75	0.78	0.79	0.77	0.78	0.80	0.76	0.86
	Male	LR	0.78	0.79	0.87	0.50	0.78	0.85	0.53	0.79	0.86	0.52	0.81
		KNN	0.83	0.83	0.83	0.80	0.83	0.98	0.27	0.79	0.90	0.40	0.84
		SVM	0.84	0.83	0.88	0.67	0.84	0.93	0.53	0.83	0.90	0.59	0.82
		VC	0.84	0.84	0.89	0.64	0.84	0.91	0.60	0.84	0.90	0.62	0.82
m4a (iOS)	Female	LR	0.72	0.72	0.80	0.56	0.72	0.77	0.61	0.72	0.79	0.58	0.75
		KNN	0.68	0.65	0.72	0.50	0.68	0.86	0.29	0.65	0.78	0.37	0.67
		SVM	0.79	0.79	0.81	0.75	0.79	0.91	0.55	0.79	0.86	0.64	0.79
		VC	0.77	0.76	0.80	0.69	0.77	0.89	0.53	0.76	0.84	0.60	0.78
	Male	LR	0.73	0.74	0.83	0.48	0.73	0.80	0.54	0.73	0.81	0.51	0.80
		KNN	0.89	0.89	0.89	0.89	0.89	0.97	0.65	0.88	0.93	0.76	0.81
		SVM	0.85	0.84	0.86	0.76	0.85	0.95	0.58	0.84	0.90	0.67	0.85
		VC	0.89	0.89	0.89	0.89	0.89	0.97	0.65	0.88	0.93	0.76	0.85

KNN: K-Nearest Neighbors, LR: Logistic Regression, Ov. : Overall, SVM: Support Vector Machine, VC: Voting Classifier

Discussion

In this study, we built an AI-based pipeline to develop a vocal biomarker for both genders and both types of smartphones (male/female, Android/iOS) that effectively recognize fatigued and non-fatigued participants with COVID-19.

We stratified the data to prevent data heterogeneity, which is considered contamination and makes it difficult to build a reliable and consistent classification model(s), resulting in poorer prediction performance. This contamination is caused by two factors: first, significant gender differences in fatigability, since it has previously been shown that men and women experience and report fatigue differently, and second, different microphone types incorporated in both smartphone devices used by the participants (iOS and Android), which have a direct impact on the quality of the recorded audios (machine learning algorithms separate the audio formats rather than the fatigue status if there is no constant microphone. (see Supplementary Online Material 2 for more details).

With the increased interest in remote voice analysis as a noninvasive and powerful telemedicine tool, various studies have been carried out, mostly in neurological disorders (eg, Parkinson's disease¹¹ and Alzheimer's disease²⁹) and mental health (eg, stress and depression³⁰). Recently, a significant research effort has evolved to employ respiratory sounds for COVID-19 and the main focus was on the use of cough^{17,31} and breathing³² to develop a COVID-19 screening tool. However, no previous work has been devoted to investigating the association of voice with COVID-19 symptoms, precisely fatigue.

Fatigue is one of the commonly reported symptoms of COVID-19 and Long-COVID syndrome³³, which can persist regardless of how severe COVID-19's acute stage is³⁴. A variety of cerebral, peripheral, and psychosocial factors^{35,7} play a role in the development of fatigue. It may also occur from chronic inflammation in the brain and at neuromuscular junctions. New evidence shows that patients with Long-COVID syndrome continue to have higher measures of blood clotting, thrombosis³⁶, which may also explain the persistence of fatigue. COVID-19 is associated with variations in airway resistance³⁷. This narrowing of the airway is manifested in the increase in

audible turbulence in both sighing and yawning, which is frequently associated with fatigue³⁸.

Human voice is produced by the flow of air from the lungs through the larynx, which causes the vocal fold vibrations, generating a pulsating airstream³⁹. The process is controlled by the laryngeal muscle activation⁴⁰ but involves the entire respiratory system to provide the air pressure necessary for phonation. Decreased pulmonary function in COVID-19 patients can cause reduced glottal airflow that is essential for normal voice production⁴¹. Furthermore, in case of increased fatigue, the voice production process may be additionally disturbed due to reduced laryngeal muscle tension, resulting in dysphonia that appears in up to 49% of COVID-19 patients⁴¹.

Study Limitations

This study has several limitations. First, although our data was stratified based on gender and smartphone devices, the mix of languages might also result in different voice features subsequently, in different model performances. There is presently no comparable dataset with similar audio recordings for further external validation of our findings. Thus, more data should be collected to improve the transferability of our vocal biomarker to other populations. Second, our data labeling was only based on a qualitative self-reported fatigue status. A fatigue severity scale would allow a quantitative assessment of fatigue severity in a uniform and unbiased way throughout all participants. Finally, time series voice analysis for each participant was not included in the study. More investigation, including time series analysis, would establish a personalized baseline for each participant, potentially enhancing the performance of our vocal biomarkers.

Conclusion

In this study, we demonstrated the association between fatigue and voice in people with COVID-19 and developed a fatigue vocal biomarker that can accurately predict the presence of fatigue. These findings suggest that vocal biomarkers, digitally incorporated into telemonitoring technologies, might be used to identify and remotely monitor this symptom in patients suffering from COVID-19 as well as other chronic diseases.

Acknowledgments

We thank all participants that accepted to be involved in the study, members that collaborated to the launch and monitoring of the Predi-COVID cohort, as well as its scientific committee, the IT team responsible for the development of the application, and the nurses in charge of recruitment, data collection, and management on the field.

Contributors

Elbéji and Fagherazzi had full access to all of the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis.

Fagherazzi, Zhang, and Fischer conceptualized and designed the study.

Elbéji, Zhang, Higa, Fischer, Despotovic, Nazarov, Aguayo, and Fagherazzi collected and analyzed data and contributed to the interpretation.

The statistical analysis was carried out by Elbéji, Zhang, Higa, and Fischer.

Elbéji drafted the initial manuscript. Elbéji, Zhang, Higa, Fischer, Despotovic, Nazarov, Aguayo, and Fagherazzi critically revised the manuscript for more important intellectual content. The funding was obtained by Fagherazzi. Fischer provided administrative, technical, and material support. The corresponding author certifies that all listed authors fulfill the authorship criteria and that no other authors that meet the criteria have been omitted.

Funding Statement

The Predi-COVID study is supported by the Luxembourg National Research Fund (FNR) (Predi-COVID, grant number 14716273), the André Losch Foundation, and the Luxembourg Institute of Health.

Competing interests: None declared.

Data Sharing Statement

Audio data, datasets and source code used in this study are publicly available.

Audio data available in Zenodo repository, [DOI: 10.5281/zenodo.5937844]

Datasets and source code available in Github, [https://github.com/LIHVOICE/Predi_COVID_Fatigue_Vocal_Biomarker].

Ethics Statement

The National Research Ethics Committee of Luxembourg (study number 202003/07) provided ethics approval to the study in April 2020.

References

- 1 WHO Coronavirus (COVID-19) Dashboard. <https://covid19.who.int> (accessed Aug 5, 2021).
- 2 Website. <https://www.oecd.org/coronavirus/policy-responses/the-territorial-impact-of-covid-19-managing-the-crisis-and-recovery-across-levels-of-government-a2c6abaf/>.
- 3 Qi R, Chen W, Liu S, *et al.* Psychological morbidities and fatigue in patients with confirmed COVID-19 during disease outbreak: prevalence and associated biopsychosocial risk factors. *medRxiv* 2020; : 2020.05.08.20031666.
- 4 Tansey CM, Louie M, Loeb M, *et al.* One-year outcomes and health care utilization in survivors of severe acute respiratory syndrome. *Arch Intern Med* 2007; **167**: 1312–20.
- 5 Neuropsychological and neurophysiological correlates of fatigue in post-acute patients with neurological manifestations of COVID-19: Insights into a challenging symptom. *J Neurol Sci* 2021; **420**: 117271.
- 6 Rudroff T, Fietsam AC, Deters JR, Bryant AD, Kamholz J. Post-COVID-19 Fatigue: Potential Contributing Factors. *Brain Sciences* 2020; **10**: 1012.
- 7 Morgul E, Bener A, Atak M, *et al.* COVID-19 pandemic and psychological fatigue in Turkey. *International Journal of Social Psychiatry*. 2021; **67**: 128–35.

- 8 Hunter SK. Sex differences in human fatigability: mechanisms and insight to physiological responses. *Acta Physiol* 2014; **210**: 768–89.
- 9 Gunasekeran DV, Tseng RMW, Tham Y-C, Wong TY. Applications of digital health for public health responses to COVID-19: a systematic scoping review of artificial intelligence, telehealth and related technologies. *NPJ Digital Medicine* 2021; **4**. DOI:10.1038/s41746-021-00412-9.
- 10 DeMerle K, Angus DC, Seymour CW. Precision Medicine for COVID-19: Phenotype Anarchy or Promise Realized? *JAMA* 2021; **325**: 2041–2.
- 11 Tracy JM, Özkanca Y, Atkins DC, Hosseini GR. Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson’s disease. *J Biomed Inform* 2020; **104**. DOI:10.1016/j.jbi.2019.103362.
- 12 Laguarda J, Subirana B. Longitudinal Speech Biomarkers for Automated Alzheimer’s Detection. *Frontiers in Computer Science*. 2021; **3**. DOI:10.3389/fcomp.2021.624694.
- 13 Cho S-W, Yin CS, Park Y-B, Park Y-J. Differences in self-rated, perceived, and acoustic voice qualities between high- and low-fatigue groups. *J Voice* 2011; **25**: 544–52.
- 14 Whitmore J, Fisher S. Speech during sustained operations. *Speech Communication*. 1996; **20**: 55–70.
- 15 Vollrath M. Automatic measurement of aspects of speech reflecting motor coordination. *Behavior Research Methods, Instruments, & Computers*. 1994; **26**: 35–40.
- 16 Greeley HP, Berg J, Friets E, *et al*. Fatigue estimation using voice analysis. *Behavior Research Methods*. 2007; **39**: 610–9.
- 17 Detection of COVID-19 from voice, cough and breathing patterns: Dataset and preliminary results. *Comput Biol Med* 2021; **138**: 104944.
- 18 Orlandic L, Teijeiro T, Atienza D. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data*.

- 2021; **8**. DOI:10.1038/s41597-021-00937-4.
- 19 Bartl-Pokorny KD, Pokorny FB, Batliner A, *et al.* The voice of COVID-19: Acoustic correlates of infection in sustained vowels. *J Acoust Soc Am* 2021; **149**: 4377.
 - 20 Fagherazzi G, Fischer A, Betsou F, *et al.* Protocol for a prospective, longitudinal cohort of people with COVID-19 and their household members to study factors associated with disease severity: the Predi-COVID study. *BMJ Open* 2020; **10**: e041834.
 - 21 United Nations. Universal Declaration of Human Rights | United Nations. <https://www.un.org/en/about-us/universal-declaration-of-human-rights> (accessed Nov 18, 2021).
 - 22 Hoffmann W, Latza U, Baumeister SE, *et al.* Guidelines and recommendations for ensuring Good Epidemiological Practice (GEP): a guideline developed by the German Society for Epidemiology. *Eur J Epidemiol* 2019; **34**: 301–17.
 - 23 [dataset]
LIHVOICE.Predi_COVID_Fatigue_Vocal_Biomarker/Android_audioset.csv at main · LIHVOICE/Predi_COVID_Fatigue_Vocal_Biomarker. GitHub. https://github.com/LIHVOICE/Predi_COVID_Fatigue_Vocal_Biomarker (accessed Jan 31, 2022).
 - 24 [dataset] Elbéji A, Zhang L, Higa E, *et al.* Audio recordings of COVID-19 positive individuals from the prospective Predi-COVID cohort study with their fatigue status. 2022; published online Feb 1. DOI:10.5281/zenodo.5937844.
 - 25 Barman R, Deshpande S, Agarwal S, Inamdar U, Devare M. Transfer Learning for Small Dataset. 2019; published online March 26. <http://dx.doi.org/> (accessed Nov 18, 2021).
 - 26 Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. *Journal of Big Data* 2016; **3**: 1–40.
 - 27 Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014; published online Sept 4. <http://arxiv.org/abs/1409.1556>

(accessed Aug 6, 2021).

- 28 Hasan BMS, Abdulazeez AM. A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. *Journal of Soft Computing and Data Mining* 2021; **2**: 20–30.
- 29 König A, Satt A, Sorin A, *et al.* Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's & dementia (Amsterdam, Netherlands)* 2015; **1**. DOI:10.1016/j.dadm.2014.11.012.
- 30 Zhang L, Duvvuri R, Chandra KKL, Nguyen T, Ghomi RH. Automated voice biomarkers for depression symptoms using an online cross-sectional data collection initiative. *Depress Anxiety* 2020; **37**. DOI:10.1002/da.23020.
- 31 Noninvasive Vocal Biomarker is Associated With Severe Acute Respiratory Syndrome Coronavirus 2 Infection. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes* 2021; **5**: 654–62.
- 32 COVID-19 Sounds App. <http://www.covid-19-sounds.org/> (accessed Nov 18, 2021).
- 33 Goërtz YMJ, Van Herck M, Delbressine JM, *et al.* Persistent symptoms 3 months after a SARS-CoV-2 infection: the post-COVID-19 syndrome? *ERJ Open Res* 2020; **6**. DOI:10.1183/23120541.00542-2020.
- 34 Townsend L, Dyer AH, Jones K, *et al.* Persistent fatigue following SARS-CoV-2 infection is common and independent of severity of initial infection. *PLoS One* 2020; **15**: e0240784.
- 35 Arnold P, Njemini R, Vantieghem S, *et al.* Peripheral muscle fatigue in hospitalised geriatric patients is associated with circulating markers of inflammation. *Experimental Gerontology*. 2017; **95**: 128–35.
- 36 Fogarty H, Townsend L, Morrin H, *et al.* Persistent endotheliopathy in the pathogenesis of long COVID syndrome. *J Thromb Haemost* 2021; **19**. DOI:10.1111/jth.15490.
- 37 Pan S-Y, Ding M, Huang J, Cai Y, Huang Y-Z. Airway resistance variation

correlates with prognosis of critically ill COVID-19 patients: A computational fluid dynamics study. *Comput Methods Programs Biomed* 2021; **208**: 106257.

- 38 Murry T. The voice and its disorders, 4th edition. By Margaret C. L. Greene, 446 pp, illus, J. B. Lippincott Co., Philadelphia, PA, 1980. \$47.50. 1981. DOI:10.1002/HED.2890030517.
- 39 Zörner S, Kaltenbacher M, Döllinger M. Investigation of prescribed movement in fluid–structure interaction simulation for the human phonation process. *Computers & Fluids*. 2013; **86**: 133–40.
- 40 Yin J, Zhang Z. Laryngeal muscular control of vocal fold posturing: Numerical modeling and experimental validation. *J Acoust Soc Am* 2016; **140**: EL280.
- 41 Dassie-Leite AP, Gueths TP, Ribeiro VV, Pereira EC, Martins P do N, Daniel CR. Vocal Signs and Symptoms Related to COVID-19 and Risk Factors for their Persistence. *J Voice* 2021; published online Aug 11. DOI:10.1016/j.jvoice.2021.07.013.

Legends

Table 1: Study population characteristics

Table 2: Results of the prediction models

Figure 1. General Pipeline

Figure 2. VGG19 Feature Extraction

Figure 3. Derivation of the digital fatigue vocal biomarker for Android and iOS users.

Publication Cover Sheet (Objective 2)

Title of the Publication:

A voice-based algorithm can predict type 2 diabetes status in USA adults: Findings from the Colive Voice study

Authors:

Abir Elbéji, Mégane Pizzimenti, Gloria Aguayo, Aurélie Fischer, Hanin Ayadi, Franck Mauvais-Jarvis, Jean-Pierre Riveline, Vladimir Despotovic, Guy Fagherazzi

Data and Figure Contributions:

Abir Elbéji was primarily responsible for data preparation and analysis, particularly for:

- Preparing the data for audio features extraction and machine learning models used in diabetes prediction.
- Conducting the data analysis for T2D prediction models.
- Led the preparation of the model performance results presented in the key figure panels.

Contribution to Manuscript Writing:

Elbéji drafted the manuscript with significant contributions from Fagherazzi and Despotovic. Elbéji was responsible for writing the Results section, focusing on the voice-based prediction algorithm and its performance across different demographic groups. Elbéji also contributed to interpreting the clinical and statistical findings in the Discussion.

Statement of Contributions:

- Elbéji and Fagherazzi conceptualized the study and were responsible for the overall study design and data analysis.

Pizzimenti, Fischer, and Ayadi coordinated data collection and preprocessing.

- Mauvais-Jarvis, Riveline, and Despotovic provided clinical expertise and contributed to the scientific content.

- All authors reviewed and revised the manuscript critically for important intellectual content.

- The statistical analysis was primarily conducted by Despotovic with assistance from Aguayo.

- Elbéji had full access to all the data and ensured the accuracy of the analysis.

A voice-based algorithm can predict type 2 diabetes status in USA adults: Findings from the Colive Voice study

Abir Elbéji¹, Mégane Pizzimenti¹, Gloria Aguayo¹, Aurélie Fischer¹, Hanin Ayadi¹, Franck Mauvais-Jarvis^{2,3}, Jean-Pierre Riveline^{4,5}, Vladimir Despotovic⁶, Guy Fagherazzi¹

Affiliations

1. Deep Digital Phenotyping Research Unit. Department of Precision Health. Luxembourg Institute of Health, 1 A-B rue Thomas Edison, L-1445 Strassen, Luxembourg
2. Section of Endocrinology and Metabolism, Deming Department of Medicine, Tulane University School of Medicine, New Orleans, LA, USA.
3. Southeast Louisiana, VA Medical Center, New Orleans, LA, USA
4. Institut Necker Enfants Malades, Inserm U1151, CNRS UMR 8253, Immediab Laboratory, Paris, France.
5. Department of Diabetology, Endocrinology and Nutrition, Assistance Publique - Hôpitaux de Paris, Lariboisière University Hospital, Paris, France. and INSERM UMR-S1151, CNRS UMR-S8253, Immediab Lab, Institut Necker-Enfants Malades, Université Paris Cité, Paris, France.
6. Bioinformatics Platform. Luxembourg Institute of Health, 1 A-B rue Thomas Edison, L-1445 Strassen, Luxembourg

Corresponding Author

Dr Guy Fagherazzi, PhD, ADR

Director of Department of Precision Health

Group Leader of the Deep Digital Phenotyping Research Unit

Luxembourg Institute of Health,

1 AB rue Thomas Edison, L-1445 STRASSEN, LUXEMBOURG

Phone: +352 6 21 58 95 54

Email: guy.fagherazzi@lih.lu

Website: <https://ddp.lih.lu/>

Short running title

Voice-based algorithm predicts type 2 diabetes in USA adults

Keywords

Type 2 diabetes; Screening; Prevention; Digital health

Abstract

Background

The pressing need to reduce undiagnosed type 2 diabetes (T2D) globally calls for innovative screening approaches. This study investigates the potential of using a voice-based algorithm to predict T2D status in adults, as the first step towards the development of a non-invasive and scalable screening method.

Methods

We analyzed pre-specified text recordings from 607 US participants from the Colive Voice study registered on ClinicalTrials.gov (NCT04848623). We used the hybrid BYOL-S/CvT embeddings to construct gender-specific algorithms to predict T2D status. These algorithms were evaluated using cross-validation, based on their accuracy, specificity, sensitivity, and Area Under the Curve (AUC). Performance of the best models was stratified by key factors such as age, BMI, and hypertension, and compared to the American Diabetes Association (ADA) score for T2D risk assessment using a Bland-Altman analysis.

Results

The voice-based algorithms demonstrated good overall predictive capacity (AUC=75% for males, 71% for females) and correctly predicted 71% of male and 66% of female T2D cases. Performances are improved in females aged 60 years or older (AUC=74%) and individuals with hypertension (AUC=75%). We observed an overall agreement above 93% with the ADA risk score.

Conclusion

Our findings suggest that voice-based algorithms could serve as a more accessible, cost-effective, and noninvasive screening tool for T2D. While the results are encouraging, further validation is needed, particularly while targeting early-stage T2D cases and more diverse populations.

Introduction

Diabetes mellitus (DM) is an endocrine system illness in which the body cannot regulate blood glucose levels. It is one of the most severe and common chronic diseases of our time, as it was responsible for 6.7 million deaths in 2021[1]. In 2022, about 1 in 10 people in the world is living with DM, and the number is expected to grow from 537 million adults, up to 643 million by 2030 and 783 million by 2045, as the result of population aging, economic development, urbanization, unhealthy eating habits, and sedentary lifestyle[1]. In the USA, according to the 2022 National Diabetes Statistics Report from the CDC[1,2], 37.3 million people, or 11.3% of the population, have diabetes. This total includes 28.7 million diagnosed cases and an estimated 8.5 million people who are living with undiagnosed diabetes.

One of the most urgent public health challenges in DM is reducing the number of undiagnosed cases worldwide. Currently, almost one in every two people with type 2 diabetes (T2D) is undiagnosed worldwide, and as a result, cannot begin treatment or preventive measures to avoid or delay complications[3]. It was demonstrated that undiagnosed DM is associated with a higher death risk when compared to normoglycemic individuals[4], as one-third of T2D patients do not present symptoms until complications appear[5]. From a health economics perspective, it has been previously reported that any undiagnosed diabetes case costs \$4,250 per year in the USA[6], generating preventable healthcare expenditures.

Nowadays, screening campaigns rely on invasive blood glucose analysis that costs around 825 billion dollars per year[7], which might be difficult to deploy at a large scale or to implement in countries or settings with limited resources and/or infrastructures. Alternative methods include scores to identify individuals at risk of developing diabetes during the next 5 to 10 years. The FINDRISC score[7,8] is widely used, although it is based on a

questionnaire with limited detection capacities (AUC around 76 %) and can be prone to errors or desirability biases.

In the United States (USA), The American Diabetes Association (ADA) diabetes risk test[9] was developed as a screening tool to classify high-risk subjects in the community and to raise awareness of modifiable risk factors and healthy lifestyles (5). The ADA diabetes risk test scoring includes seven questions (total score of 0–11) regarding age, gender, gestational diabetes mellitus (GDM), family history of diabetes, high blood pressure, physical activity, and obesity (based on body mass index (BMI) via a weight-height chart). Those having scores of 5 and more are considered to be at high risk of having diabetes.

With the advancement of digital technologies and artificial intelligence, significant effort is being directed towards detecting diabetes through noninvasive methods. These methods range from human facial block color analysis using sparse representation classifiers[10], hair analysis through elemental composition[11–13], specialized eye exams aimed at detecting diabetic retinopathy[14], to voice analysis, which stands as one of the most promising technologies in healthcare applications. This includes early diagnosis of neurodegenerative diseases[15] and assisting in screening and monitoring symptoms of conditions like COVID-19[16] through the analysis of subtle speech pattern alterations and vocal biomarkers.

Previous works have suggested that people with diabetes have different voice features than people without diabetes. People with T2D with poor glycemic control or with neuropathy are also more likely to have phonatory symptoms compared to controls[17], namely a higher average score for vocal grading, straining[18], and hoarseness[19] that are affecting patients' quality of life. From an acoustic perspective, it has been shown that voice parameters like jitter, shimmer, smoothed amplitude perturbation quotient, noise-to-harmonic ratio, relative average perturbation, mean fundamental frequencies, maximum phonation

time, and amplitude perturbation quotient show significant differences in their values between T2D patients and people without diabetes[20]^[21]. However, previous studies relied on relatively small sample sizes, a lack of diversity in the participant profiles, and a lack of validation with audio recordings captured in real-world settings.

Building on this groundwork, our study distinguishes itself by leveraging data from the Colive Voice program to develop and assess the performance of a voice-based AI algorithm for T2D status detection in the adult population in the USA. This initiative not only serves as a first step toward using voice analysis as a first-line T2D screening strategy but also offers insights into the complex nature of T2D and its interaction with voice characteristics. Accordingly, we place special emphasis on considering a wide array of demographic and health-related parameters. This holistic approach is crucial as these factors can significantly affect voice characteristics and, consequently, their potential as indicators for disease states.

Methods

Study population

In 2021, the Luxembourg Institute of Health initiated a worldwide, multilingual research program named Colive Voice. Its ongoing project serves as a screening platform for vocal biomarkers, for screening or monitoring various chronic diseases and frequent health symptoms. To ensure diversity, Colive Voice collects voice recordings from participants above the age of 15 years, regardless of their health status and conditions, in English, French, German, and Spanish globally. Each participant contributes with standardized vocal tasks which are then annotated with clinical and demographic data.

Ethics Statement

Colive Voice is registered on ClinicalTrials.gov (NCT04848623) and was approved by the National Research Ethics Committee of Luxembourg (study number 202103/01) in March 2021. All participants provided informed consent to take part in the study.

Collected Data

Colive Voice participants are invited to complete a comprehensive questionnaire to gather a diverse range of information: demographic characteristics, lifestyle habits, anthropometric data, symptoms, drug use, and history of chronic diseases. Regarding diabetes, Colive Voice gathers data on the diagnosis, type of diabetes, duration since diagnosis as well as treatment categories, and HbA1c levels. For the present work, we included English-speaking participants from the USA and we analyzed each gender separately. Participants were invited to record a standardized reading task using the 25th article from the Human Rights Declaration (Fig 1). All the collected raw audio data was processed and quality-checked to ensure consistency throughout the study. There was no missing data in this study, ensuring a robust and complete dataset for analysis.

Voice feature extraction

OpenSmile

OpenSmile[22] is an open-source toolkit, popularly used for generating handcrafted low-level descriptors (LLD) from audio inputs. These descriptors encapsulate key characteristics of audio signals over time such as pitch, intensity, and spectral properties. OpenSmile computes functionals on these LLD contours, capturing statistical attributes like peaks, means, and ranges to provide a higher-level overview of the audio signal. Among the feature sets that OpenSmile offers, the ComParE set stands out. Comprising 6373 static features, ComParE is notable for its comprehensive nature, offering a rich and extensive array of data points. This vast collection of features facilitates the detection of complex patterns in the audio data, offering an in-depth understanding of the audio source.

BYOL-S/CvT

The hybrid model, BYOL-S/CvT[23], is a new method that detects cognitive and physical load in speech. It uses both data-driven features from the self-supervised BYOL-S model trained on Audioset and handcrafted features from OpenSmile. This mix improves the model's performance and helps it learn speech patterns better than traditional methods. The BYOL-S/CvT model is also efficient and fast, needing only a single step during the decision-making stage, and produces 2048-dimensional embeddings.

Data analysis

In this study, the authors adhered to the TRIPOD criteria (the Transparent Reporting of a Multivariable Prediction Model of Individual Prognosis Or Diagnosis) standards for the reporting of AI-based algorithm development and validation and used the corresponding checklist to guide the drafting of the manuscript.

To mitigate gender bias and manage imbalanced data challenges in our machine learning algorithm training, we first stratified the dataset based on gender. Following this, we used a simple random sampling technique to generate balanced group sizes, ensuring a more equitable and effective training process. Individuals without endocrine diseases, including diabetes, were selected randomly from the general USA population to create a control group that matched the size of the group of participants with T2D.

To enhance our algorithm's performance, we first normalized the extracted features and embeddings using a standard scaler, which helps ensure consistent variance across all features. As high-dimensional inputs could lead to overfitting and poor generalization in machine learning algorithms, we used Principal Component Analysis (PCA) to reduce the

dimensionality of the BYOL-S/CvT embeddings. For OpenSmile features, we used feature selection using the SelectKBest function from scikit-learn.

Once the normalization, reduction, and feature selection processes were complete, the resulting features were fed into three different classifiers: Logistic Regression (LR), Support Vector Machine with a radial basis function kernel (SVM RBF), and Multi-Layer Perceptron classifiers (MLP).

To evaluate the performance and compare the classifiers, we used stratified 5-fold cross-validation, ensuring no data leakage via the Pipeline functionality from scikit-learn. This pipeline handled scaling and PCA reduction for the BYOL-S embeddings, as well as scaling and feature selection for OpenSmile features. We measured the algorithms' performances using accuracy, specificity, sensitivity, and AUC metrics.

For optimal results, we fine-tuned the number of PCA components and the algorithms' hyperparameters using the grid search function from scikit-learn. We then used the best feature-classifier combination to select the most performant algorithm for each gender.

Influence of cofactors and their impact on algorithms' performance metrics

In order to highlight how different cofactors influence the efficacy of our predictive algorithms, we conducted a performance stratification analysis. This analysis was segmented by age (below and over 60 years), BMI (below and over 25-29.9 kg/m²). Additionally, we examined conditions including hypertension, migraine, diagnosed depression, smoking, stress, and fatigue (measured by the Fatigue Severity Scale[24]), designating each condition's status as either 'present' and 'absent' or 'severe' and 'mild'. To reinforce confidence in our performance metrics and facilitate comparisons, we employed a bootstrapping technique. This involved generating multiple subsamples for each combination of comorbidity and its

status. The bootstrapping process, repeated for each comorbidity, involves sampling with replacement from the original dataset and subsequently recalculating the metrics for each subsample.

With the objective of developing a screening tool in mind, the assessment of specificity and sensitivity metrics was prioritized, but AUC was also reported. High sensitivity guarantees that true cases are not missed, while high specificity reduces false alarms, optimizing resource use and building user trust. Performance metrics were computed independently for each bootstrap iteration within the respective groups. To evaluate the statistical significance of performance differences between categories, we employed the Mann-Whitney U test. Finally, to account for multiple comparisons, we adjusted the p-values obtained from these statistical tests using a Bonferroni correction.

Sensitivity analysis

As a sensitivity analysis, we conducted a Bland-Altman analysis between the voice-based algorithms and the ADA risk score, which serves as a gold standard for assessing T2D risk in the USA population[9]. Due to data limitations, physical activity levels and family history of diabetes were not available in Colive Voice and were set to zero for all participants by default. In this context, the modified ADA risk score ranges from zero, denoting no T2D risk, to seven, indicating a high risk.

Results

Population characteristics

We analyzed 323 females and 284 males based on T2D status. The majority were identified as white: 73.3% of females with T2D, 76.5% of females without T2D, 77.5% of males with T2D, and 71.8% of males without T2D.

Significant differences were identified across the groups, including age, BMI (t-test p-value < 0.001), and prevalence of hypertension (chi2 p-value < 0.001). Those with T2D, in both genders, had higher average ages and BMIs than those without T2D. Specifically, females with T2D had an average age of 49.5 years and a BMI of 35.8 kg/m², compared to 40.0 years and 28.0 kg/m² in those without T2D. Male participants with T2D had an average age of 47.6 years and BMI of 32.8 kg/m², whereas those without averaged 41.6 years and 26.6 kg/m².

Hypertension was more prevalent among the T2D group. Among females with T2D, 50% reported hypertension, compared to 11.18% in the group without T2D. For males, a similar trend was observed, with 58.5% of those with T2D having hypertension, compared to 12.7% without the condition.

Depression diagnosis history also was more prevalent in those with T2D (chi2 p-value < 0.001), especially in females: 61.7% with T2D reported depression, compared to 45.3% without T2D. Among males, the rates were 48.6% for those with T2D and 31.7% for those without T2D. Other health conditions and scores are included in Table 1.

Table 1: Study population characteristics

	Female group			Male group		
T2D status	Without T2D	With T2D	P-value	Without T2D	With T2D	P-value
Participants (N)	161	162	-	142	142	-
Age (year)	40.0 (13.5)	49.5 (12.1)	<0.001	41.6 (14.0)	47.6 (13.4)	<0.001
Body Mass Index (kg/m²)	28.0 (7.3)	35.8 (8.9)	<0.001	26.6 (5.5)	32.8 (8.5)	<0.001
Ethnicity: White	118 (73.3%)	124 (76.5%)	0.28	110 (77.5%)	102 (71.8%)	0.59
Ethnicity: Black	20 (12.4%)	21 (13.0%)		10 (7.0%)	12 (8.5%)	
Ethnicity: Other	23 (14.3%)	17 (10.5%)		22 (15.5%)	28 (19.7%)	
Fatigue Severity Scale	32.3 (13.4)	40.3 (12.3)	<0.001	31.3 (12.8)	40.3 (12.3)	<0.001
Perceived stress (% yes)	38 (23.6%)	49 (30.3%)	0.48	29 (20.4%)	38 (26.7%)	0.16
Smoking (% yes)	28 (17.4%)	19 (11.7%)	0.22	32 (22.5%)	34 (23.9%)	0.24
Migraine (% yes)	33 (20.5%)	43 (26.5%)	0.25	16 (11.3%)	19 (13.4%)	0.72
Thyroidic disease (% yes)	0 (0%)	37 (22.8%)	<0.001	0 (0%)	10 (0.7%)	<0.01
Hypertension (% yes)	18 (11.2%)	81 (50.0%)	<0.001	18 (12.7%)	83 (58.5%)	<0.001
Diagnosed depression (% yes)	73 (45.3%)	100 (61.7%)	<0.01	45 (31.7%)	69 (48.6%)	<0.01
HbA1c (%)	-	7.14 (1.8)	-	-	7.20 (1.7)	-
Diabetes treatment (% yes)	-	126 (77.8%)	-	-	114 (80.3%)	-
Diabetes duration (year)	-	8.9 (7.3)	-	-	9.1 (7.6)	-

The table presents clinical data describing the overall population of the study. Categorical data are represented by total numbers and percentages, with the calculated p-values derived from chi-square tests. Continuous data are represented by mean and standard deviation, with p-values calculated using the Student's t-test.

Algorithms' performances

In both genders, MLP classifiers trained with BYOL-S/CvT embeddings significantly outperformed those trained solely on OpenSMILE features in both males and females (Table 2).

Table 2: Results of the prediction models.

	Features	Dimensionality reduction	Classifier	Accuracy	Specificity	Sensitivity	AUC
Female group	Opensmile ComParE 2016 (6373)	200 selected features	LR	0.60 (0.03)	0.60 (0.03)	0.62 (0.07)	0.62 (0.02)
			MLP Classifier	0.63 (0.02)	0.61 (0.02)	0.74 (0.02)	0.66 (0.02)
			SVM RBF	0.57 (0.02)	0.57 (0.02)	0.63 (0.03)	0.61 (0.01)
	Byol-S embeddings (2048)	PCA, n_components= n_samples	LR	0.67 (0.04)	0.68 (0.04)	0.65 (0.11)	0.70 (0.06)
			MLP Classifier	0.67 (0.04)	0.66 (0.04)	0.67 (0.11)	0.71 (0.07)
			SVM RBF	0.66 (0.04)	0.65 (0.07)	0.67 (0.11)	0.71 (0.05)
Male group	Opensmile ComParE 2016 (6373)	100 selected features	LR	0.56 (0.02)	0.55 (0.01)	0.58 (0.05)	0.61 (0.05)
			MLP Classifier	0.61 (0.05)	0.61 (0.06)	0.63 (0.06)	0.64 (0.05)
			SVM RBF	0.57 (0.05)	0.57 (0.06)	0.54 (0.05)	0.57 (0.05)
	Byol-S embeddings (2048)	PCA, n_components= 100	LR	0.69 (0.04)	0.66 (0.07)	0.72 (0.03)	0.73 (0.06)
			MLP Classifier	0.71 (0.02)	0.70 (0.02)	0.73 (0.03)	0.75 (0.05)
			SVM RBF	0.70 (0.04)	0.64 (0.05)	0.76 (0.03)	0.78 (0.05)

Table 2 presents the mean and standard deviation (in parentheses) of the performance metrics across cross-validation folds. The selected algorithm for each gender group is highlighted in bold. Logistic Regression (LR), Multi-layer Perceptron (MLP), Support Vector Machine Radial basis function kernel (SVM RBF).

For the prediction of T2D in females, the classifier achieved a sensitivity of 0.67 ± 0.11 , specificity of 0.66 ± 0.04 , an AUC of 0.71 ± 0.07 and a Brier score of 0.31. For the prediction of T2D in males, the reported performance metrics were a sensitivity of 0.73 ± 0.03 , specificity of

0.70±0.02, an AUC of 0.75±0.05 and a Brier score of 0.22 (Fig 2). The predicted probability of having T2D is then used for the sensitivity analysis with ADA risk score.

Performance stratification

The specificity and sensitivity metrics showed variability across various subgroups.

When stratifying by key demographics, notable differences were observed for females across age categories, with females aged 60 and above exhibiting higher specificity (0.74±0.12), sensitivity (0.74±0.07), and AUC (0.74±0.07) compared to females aged below 60 for both specificity and sensitivity (0.65±0.04), and for AUC (0.65±0.03) (Table 3).

Table 3: Performance stratification of voice-based T2D status detection algorithms.

			Females			Males		
			Specificity	Sensitivity	AUC	Specificity	Sensitivity	AUC
Demographics	Age	<60 y	0.65 (0.04)	0.65 (0.04)	0.65 (0.03)	0.70 (0.04)	0.74 (0.04)	0.72 (0.03)
		≥ 60y	0.74 (0.12)	0.74 (0.07)	0.74 (0.07)	0.70 (0.11)	0.70 (0.10)	0.70 (0.07)
	Body Mass Index	<25 kg/m ²	0.68 (0.06)	0.58 (0.12)	0.63 (0.07)	0.70 (0.06)	0.78 (0.09)	0.74 (0.05)
		≥ 25 kg/m ²	0.65 (0.05)	0.68 (0.04)	0.67 (0.03)	0.69 (0.05)	0.72 (0.04)	0.71 (0.03)
Comorbidities	Hypertension	Present	0.76 (0.11)	0.75 (0.05)	0.75 (0.06)	0.72 (0.11)	0.76 (0.05)	0.74 (0.06)
		Absent	0.65 (0.04)	0.61 (0.05)	0.63 (0.03)	0.69 (0.04)	0.70 (0.05)	0.70 (0.03)
	Migraine	Present	0.86 (0.07)	0.75 (0.07)	0.80 (0.05)	0.67 (0.12)	0.71 (0.11)	0.69 (0.09)
		Absent	0.62 (0.04)	0.65 (0.04)	0.65 (0.04)	0.70 (0.04)	0.74 (0.04)	0.72 (0.03)
Lifestyle factors and symptoms	Smoking	Present	0.60 (0.09)	0.53 (0.12)	0.57 (0.07)	0.74 (0.09)	0.76 (0.07)	0.75 (0.06)
		Absent	0.67 (0.04)	0.69 (0.04)	0.68 (0.03)	0.69 (0.04)	0.72 (0.04)	0.71 (0.03)
	Depressive symptoms	Severe	0.75 (0.05)	0.71 (0.05)	0.73 (0.03)	0.71 (0.07)	0.71 (0.06)	0.71 (0.04)
		Mild	0.58 (0.05)	0.61 (0.06)	0.60 (0.04)	0.69 (0.05)	0.75 (0.05)	0.72 (0.03)
	Stress	Present	0.76 (0.07)	0.62 (0.07)	0.69 (0.05)	0.69 (0.09)	0.77 (0.07)	0.72 (0.06)
		Absent	0.63 (0.04)	0.70 (0.04)	0.66 (0.03)	0.70 (0.04)	0.72 (0.04)	0.71 (0.03)
	Fatigue	Severe	0.68 (0.06)	0.68 (0.05)	0.68 (0.04)	0.71 (0.06)	0.73 (0.05)	0.72 (0.04)

		Mild	0.65 (0.05)	0.66 (0.06)	0.65 (0.04)	0.69 (0.05)	0.73 (0.06)	0.71 (0.04)
--	--	-------------	-------------	-------------	-------------	-------------	-------------	-------------

This table provides an overview of various metrics, differentiated by gender across different demographic factors, comorbidities, and lifestyle factors. The statistical significance of performance differences between categories was evaluated using the Mann-Whitney U test, with all results being statistically significant ($p < 0.001$).

Conversely, no noticeable disparities were observed among males.

When considering comorbidities, hypertension emerged as a significant enhancer of the algorithm's performance in both genders. The presence of hypertension enhanced the sensitivity (0.75 ± 0.05 for females and 0.76 ± 0.05 for males), highlighting the algorithm's efficiency in detecting T2D in individuals with hypertension. On the other hand, for females, migraine considerably increases specificity to 0.86 ± 0.07 and sensitivity to 0.75 ± 0.07 , while for males with migraine, both specificity (0.67 ± 0.12) sensitivity (0.71 ± 0.11) is lower. This suggests that migraine has a more pronounced impact on the accuracy of T2D detection in women than in men.

Lifestyle factors and symptoms also influence performance. The presence of depressive symptoms significantly impacts the algorithm's performance in women, increasing both specificity (0.75 ± 0.05) and sensitivity (0.71 ± 0.05). Conversely, for men, the impact of depressive symptoms are less prominent, with a slight decrease in sensitivity (from 0.75 ± 0.05 to 0.71 ± 0.06) yet a stable specificity of 0.71 ± 0.07 . This demonstrates enhanced accuracy in detecting T2D in women with depression. Smoking and stress revealed gender-specific impacts; smoking led to higher sensitivity in males (0.76 ± 0.07) compared to a decreased sensitivity in females (0.53 ± 0.12). Similarly, stress resulted in increased sensitivity for men (0.77 ± 0.07) but decreased for women (0.62 ± 0.07). Fatigue showed a uniform impact on specificity in both genders yet an increase in sensitivity in females with severe fatigue (0.68 ± 0.05) compared to a stable sensitivity for males of (0.73 ± 0.05).

Overall, the data indicates that the algorithm's specificity and sensitivity are influenced by demographic factors, comorbidities, and lifestyle factors, with notable differences observed

between genders. These findings underscore the importance of considering these variables in the development and refinement of diagnostic tools, ensuring more accurate and gender-specific healthcare strategies in managing and diagnosing T2D.

Agreement with ADA risk score

In the Bland-Altman analysis, the mean difference indicates the average bias between the algorithm's scores and the ADA risk scores. This analysis indicates that the algorithm has a mean difference of 0.57 for females and -0.15 for males compared to the ADA risk score, with over 93% agreement within acceptable limits for both genders, showing consistent agreement across genders (see Supplemental Figure).

Furthermore, we calculated the AUC for the ADA score and found comparable results to the voice-based algorithm's performance: AUC for the ADA risk score was 0.72 for females and 0.71 for males, compared to the algorithm's AUC of 0.71 (0.07) for females and 0.75 (0.05) for males. These findings indicate that our voice-based algorithm performs similarly to the established ADA risk score, further supporting its potential as a reliable screening tool for T2D.

Discussion

In this study, using a large sample from the USA population, we developed voice-based algorithms to detect T2D status. Our goal was to explore the possibility of using a rapid, user-friendly voice recording as a T2D status predictor. We observed that the performance of the predictive algorithms was maximal when trained using the hybrid BYOL-S/CvT embeddings, achieving AUC scores of 0.75 and 0.71 for the male and female groups, respectively. Besides demonstrating overall fair to good performances, we also examined the influence of cofactors on voice-based T2D status prediction, which allowed us to identify key subgroups of the population with enhanced performances. In a sensitivity analysis, we

have confirmed a strong agreement with the currently used questionnaire-based ADA risk score, a gold standard for T2D risk assessment in the USA.

Undiagnosed T2D or delayed diagnosis can accelerate the occurrence of serious diabetes-related complications, including cardiovascular diseases, neuropathy, retinopathy, and nephropathy[25]. One potential under-investigated effect of T2D is its impact on voice, which may be due to the disease's influence on respiratory and neuromuscular functions[19,20,26]. It was already shown that pulmonary function is reduced in people with T2D compared to those with no diabetes[27]. For speech production, an individual needs a sufficient air intake, which then travels through the trachea and larynx, causing vocal fold vibrations. Articulating these vibrations into speech requires various small muscles in the neck and throat, connected by a large nerve network. Diabetes is commonly linked to peripheral neuropathy, but it can also impact other systems[28]. This includes potential nerve damage in the throat and neck region, which is vital for speech production. Research has suggested that diabetes can lead to voice changes, especially in those with poor glucose control, causing symptoms like hoarseness and strain[18,28]. These patients often have reduced maximum phonation times, indicating neuromuscular and respiratory alterations[18,20,28]. Building upon this, our study, with its larger sample size, offered a comprehensive exploration of the vocal and physiological complications associated with T2D. By assessing cofactors, we also highlighted how they influence voice patterns, providing valuable insights for future diagnostic strategies.

Key demographic indicators, mainly age, were central in T2D status prediction using voice, especially for women. This aligns with existing research that emphasizes the importance of this variable as a critical determinant of diabetes risk[29,30]. We observed that older females (≥ 60) exhibited higher specificity, sensitivity, and AUC compared to younger ones (< 60), but no difference was observed in males. An adult woman's average fundamental frequency range is 165 to 255 Hz, while a man's is 85 to 155 Hz[31]. In females, hormonal changes related to menopause can affect vocal cords and larynx and, consequently, cause a drop in the fundamental frequency of the voice[32]. These hormonal variations may interact with the

metabolic disruptions caused by diabetes, leading to observable changes in voice pitch. On the other hand, males, not subject to the same degree of hormonal fluctuations, may exhibit less noticeable alterations in fundamental frequency.

Additionally, hypertension emerged as a key influencer in T2D status detection using voice, improving the predictive performance for both genders by up to 6%. While hypertension is known to be associated with diabetes development[33], it is not commonly incorporated into standard T2D risk assessment tools and its correlation with voice changes remains relatively unexplored[33,34].

Another distinguishing feature of our study is the gender-specific analysis of voice-based algorithms. We found that while certain determinants of T2D status were consistently influential across genders, others displayed gender-specific variations. Discrepancies observed in the impact of conditions such as migraines and on the voice-based T2D status detection algorithm's performance might be traced back to inherent gender-based physiological differences. Women are also more likely to experience migraine than men, with more frequent and severe attacks[35]. Besides, migraine and diabetes have already been shown to be associated with women[36], and our study confirms that this association can be captured by changes in female voices[37]. The varying impact of smoking on the algorithm's performances between genders may reflect gender-specific vocal changes caused by smoking[37,38]. Depression affects voices differently between men and women, suggesting that depression is linked to a higher risk of diabetes in women, but not in men[39,40]. This gender-specific association might explain the observed disparities in voice changes. The physiological and psychological stresses associated with depression may induce subtle voice changes that vary between genders, potentially due to hormonal or neurological differences. This variation might be more pronounced in women due to the combined impact of hormonal disruptions related to both diabetes and depression. Stress and fatigue, both of which can affect voice quality[41,42], seem to influence the algorithm's performance in a gender-specific manner. These factors, known to play roles in glucose metabolism and insulin resistance[43], likely contribute to the voice patterns identified by the algorithm as

indicative of T2D risk.

Such a sensitivity analysis is rarely performed in the field of vocal or digital biomarkers, as authors frequently report overall performances only. Our approach underscores that integrating the analysis of the influence of key demographic and health parameters is essential before developing any reliable voice-based screening tool. This helps to understand the potential physiological influence of these factors on either voice features or the health outcome of interest. Identifying the key sub-groups in the population is crucial to determining where the performance of these tools could be optimal.

Strengths and limitations

This work has several strengths. First, we used the most comprehensive sample of USA participants with standardized voice ecological recordings, collected in a real-life setting, compared to existing datasets. Additionally, we performed the analysis separately, stratified for males and females, to account for major gender differences in voice characteristics and to mitigate gender bias. Voice features can vary significantly between males and females due to physiological and hormonal differences, which can affect the accuracy and performance of the algorithm if not accounted for. By developing separate models for each gender, we were able to fine-tune the algorithms for the specific characteristics of males and females, improving overall predictive performance and ensuring fairness and generalizability. Besides displaying overall good performances, we also performed additional analyses to identify important subgroups where the voice-based algorithms would perform even better. Our comparative analysis of cofactors emphasized the complex nature of T2D and its interaction with voice characteristics, providing some levels of interpretability and explainability to the algorithms. Importantly, we have been able to benchmark the voice-based algorithms against an existing screening strategy in the USA, and we demonstrated a strong agreement with the ADA risk score. This concordance reinforces the potential use of voice-based analysis as a viable first-line screening tool for T2D.

There is also scope for further refinement before such algorithms can be considered ready for implementation as a screening tool and several limitations have to be acknowledged in our study. First, due to data constraints in ADA score calculation, missing values for parameters, namely physical activity and family history of diabetes were assigned a value of zero for all participants by default. While this approach might introduce less variability in the ADA scores, the potential for misclassification arises. However, the impact of this limitation is somewhat limited since the ADA score is primarily driven by age and BMI, which are available in our study. Even though they represent different constructs, we have still observed a strong agreement between the voice-based algorithms and the ADA risk score. Another limitation is that our study relied on a sample of English speakers only, with diverse T2D durations. To robustly establish and reinforce the performance of a future screening tool in predicting T2D, a more diverse and large dataset is needed, while specifically targeting early-stage T2D and prediabetes cases. Additionally, conducting longitudinal studies will help to better understand how changes in voice characteristics correlate with the development and progression of T2D. This approach will provide insights into the main clinical diabetes-related parameters, such as glycemic control and diabetes-related complications, and help establish causal relationships. Furthermore, it is also important to generalize this research across different populations, with diverse backgrounds and languages. Expanding datasets will allow a deeper examination of nuanced factors, comorbidities, and their interactions affecting voice-based screening tools in predicting T2D.

Conclusion and Perspectives

This work demonstrates the potential of using voice analysis in a diabetes context. A voice recording could soon potentially be used as a scalable, non-invasive first-line diabetes screening strategy. Future research should focus on targeting individuals with early-stage T2D and prediabetes and expanding our findings to other populations in prospective studies. Given the high societal costs of undiagnosed diabetes in the USA, our findings open new

perspectives to improve secondary prevention, reduce the impact of diabetes and prevent severe complications and premature diabetes-related mortality.

Acknowledgments

Colive Voice study is funded by the Luxembourg Institute of Health. The funders played no role in the study design, data collection, analysis, and interpretation of data, or the writing of this manuscript.

We would like to thank all participants who contributed to the Colive Voice study, as well as our partners for their help in recruiting new participants. Special thanks go to Aurélie Fischer, Philippe Kayser, Luigi De Giovanni, Michael Schnell, and Aurore Dobosz for their substantial contribution to the Colive Voice study.

Authorship confirmation/contribution statement

AE and GF conceived the study. AE and GF wrote the first draft. AE, MP, HA, VD and GF had full access to the data in the study, verified the data, and had full responsibility for the decision to submit and publish. GA, AF, VD and GF contributed to protocol development and study design. AE and VD performed the analyses. All authors read the manuscript, critically revised it for academic content and approved the final version. GF is the guarantor of this work and, as such, has full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Competing interests

Guy Fagherazzi has provided advisory services and/or has received research grants and/or speaker honoraria from MSD, MSDAvenir, Eli Lilly, Roche Diabetes Care, AstraZeneca, Danone Research, Diabeloop, Sanofi, Bristol Myers Squibb, L'Oréal R&D, Abbvie Pharmaceutical, Pfizer, Vitalaire and Akuity Care. There is no conflict of interest to disclose.

Funding statement

Colive Voice study is funded by the Luxembourg Institute of Health. This work was further supported by the French-speaking Diabetes Society, the Luxembourg Diabetes Society and the Luxembourg Diabetes Association.

Data and code availability

Audio data and source codes used in this study are publicly available in a Github repository.

<https://github.com/LIHVOICE/Voice-and-diabetes-VOCADIAB>

References

1. Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res Clin Pract.* 2022;183: 109119.
2. National Diabetes Statistics Report 2022. [cited 4 Sep 2023]. Available: <https://repository.gheli.harvard.edu/repository/11854/>
3. Ogurtsova K, Guariguata L, Barengo NC, Ruiz PL-D, Sacre JW, Karuranga S, et al. IDF diabetes Atlas: Global estimates of undiagnosed diabetes in adults for 2021. *Diabetes Res Clin Pract.* 2022;183: 109118.
4. Wild SH, Smith FB, Lee AJ, Fowkes FG. Criteria for previously undiagnosed diabetes and risk of mortality: 15-year follow-up of the Edinburgh Artery Study cohort. *Diabet Med.* 2005;22. doi:10.1111/j.1464-5491.2004.01433.x
5. Standards of medical care for patients with diabetes mellitus. *Diabetes Care.* 2003;26 Suppl 1. doi:10.2337/diacare.26.2007.s33
6. Dall TM, Yang W, Gillespie K, Mocarski M, Byrne E, Cintina I, et al. The Economic Burden of Elevated Blood Glucose Levels in 2017: Diagnosed and Undiagnosed Diabetes, Gestational Diabetes Mellitus, and Prediabetes. *Diabetes Care.* 2019;42: 1661.
7. Zhou B, Lu Y, Hajifathalian K, Bentham J, Di Cesare M, Danaei G, et al. Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4·4 million participants. *Lancet.* 2016;387: 1513–1530.
8. Lim HM, Chia YC, Koay ZL. Performance of the Finnish Diabetes Risk Score (FINDRISC) and Modified Asian FINDRISC (ModAsian FINDRISC) for screening of undiagnosed type 2 diabetes mellitus and dysglycaemia in primary care. *Prim Care Diabetes.* 2020;14. doi:10.1016/j.pcd.2020.02.008
9. Lindström J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes Care.* 2003;26. doi:10.2337/diacare.26.3.725
10. Zhang B, Vijaya kumar BV, Zhang D. Noninvasive Diabetes Mellitus Detection Using Facial Block Color With a Sparse Representation Classifier. [cited 4 Sep 2023]. Available: <https://ieeexplore.ieee.org/document/6675828>
11. IEEE Xplore - Temporarily Unavailable. [cited 4 Sep 2023]. Available: <https://ieeexplore.ieee.org/document/6675828>
12. Saleh SAK, Fatani SH, Adly HM, Abdulkhaliq AA, Al-Amodi HS. Variations of Hair Trace Element Contents in Diabetic Females. *Journal of Biosciences and Medicines.* 2017;5: 49–56.
13. Jaime Miranda J, Taype-Rondan A, Tapia JC, Gastanadui-Gonzalez MG, Roman-Carpio R. HAIR FOLLICLE CHARACTERISTICS AS EARLY MARKER OF TYPE 2 DIABETES. *Med Hypotheses.* 2016;95: 39.
14. Diabetic Retinopathy: Present and Past. *Procedia Comput Sci.* 2018;132: 1432–1440.

15. Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson's disease. *J Biomed Inform.* 2020;104: 103362.
16. Fagherazzi G, Zhang L, Elbéji A, Higa E, Despotovic V, Ollert M, et al. A voice-based biomarker for monitoring symptom resolution in adults with COVID-19: Findings from the prospective Predi-COVID cohort study. *PLOS Digital Health.* 2022;1: e0000112.
17. Gölaç H, Atalik G, Türkcan AK, Yılmaz M. Disease related changes in vocal parameters of patients with type 2 diabetes mellitus. *Logoped Phoniatr Vocol.* 2022;47. doi:10.1080/14015439.2021.1917653
18. Hamdan AL, Jabbour J, Nassar J, Dahouk I, Azar ST. Vocal characteristics in patients with type 2 diabetes mellitus. *Eur Arch Otorhinolaryngol.* 2012;269. doi:10.1007/s00405-012-1933-7
19. Hamdan AL, Kurban Z, Azar ST. Prevalence of phonatory symptoms in patients with type 2 diabetes mellitus. *Acta Diabetol.* 2013;50. doi:10.1007/s00592-012-0392-3
20. Instrumental Acoustic Voice Characteristics in Adults with Type 2 Diabetes. *J Voice.* 2021;35: 116–121.
21. Stogowska E, Kamiński KA, Ziółko B, Kowalska I. Voice changes in reproductive disorders, thyroid disorders and diabetes: a review. *Endocrine Connections.* 2022;11. doi:10.1530/EC-21-0505
22. Florian Eyben Technische Universität München, München, Germany, Martin Wöllmer Technische Universität München, München, Germany, Björn Schuller Technische Universität München, München, Germany. *Opensmile.* [cited 4 Sep 2023]. doi:10.1145/1873951.1874246
23. Elbanna G, Biryukov A, Scheidwasser-Clow N, Orlandic L, Mainar P, Kegler M, et al. Hybrid Handcrafted and Learnable Audio Representation for Analysis of Speech Under Cognitive and Physical Load. *ArXiv.* 2022. Available: <https://arxiv.org/pdf/2203.16637.pdf>
24. Krupp LB, LaRocca NG, Muir-Nash J, Steinberg AD. The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus. *Arch Neurol.* 1989;46. doi:10.1001/archneur.1989.00520460115022
25. Wu Y, Ding Y, Tanaka Y, Zhang W. Risk Factors Contributing to Type 2 Diabetes and Recent Advances in the Treatment and Prevention. *Int J Med Sci.* 2014;11: 1185.
26. Blood Glucose Estimation From Voice: First Review of Successes and Challenges. *J Voice.* 2022;36: 737.e1–737.e10.
27. Davis TME, Drinkwater JJ, Davis WA. Pulmonary Function Trajectories Over 6 Years and Their Determinants in Type 2 Diabetes: The Fremantle Diabetes Study Phase II. *Diabetes Care.* 2024 [cited 17 Jan 2024]. doi:10.2337/dc23-1726
28. Patel K, Horak H, Tiriyaki E. Diabetic neuropathies. *Muscle Nerve.* 2021;63: 22–30.
29. Ganz ML, Wintfeld N, Li Q, Alas V, Langer J, Hammer M. The association of body mass index with the risk of type 2 diabetes: a case–control study nested in an electronic health records system in the United States. *Diabetol Metab Syndr.* 2014;6: 50.
30. Yan Z, Cai M, Han X, Chen Q, Lu H. The Interaction Between Age and Risk Factors for Diabetes and Prediabetes: A Community-Based Cross-Sectional Study. *Diabetes Metab*

Syndr Obes. 2023;16: 85.

31. Fitch JL, Holbrook A. Modal vocal fundamental frequency of young adults. *Arch Otolaryngol.* 1970;92. doi:10.1001/archotol.1970.04310040067012
32. Lã FMB, Ardura D. What Voice-Related Metrics Change With Menopause? A Systematic Review and Meta-Analysis Study. *J Voice.* 2022;36. doi:10.1016/j.jvoice.2020.06.012
33. Kim M-J, Lim N-K, Choi S-J, Park H-Y. Hypertension is an independent risk factor for type 2 diabetes: the Korean genome and epidemiology study. *Hypertens Res.* 2015;38: 783–789.
34. Sakai M. Case study on analysis of vocal frequency to estimate blood pressure. [cited 4 Sep 2023]. Available: <https://ieeexplore.ieee.org/document/7257173>
35. Allais G, Chiarle G, Sinigaglia S, Airola G, Schiapparelli P, Benedetto C. Gender-related differences in migraine. *Neurol Sci.* 2020;41: 429.
36. Fagherazzi G, El Fatouhi D, Fournier A, Gusto G, Mancini FR, Balkau B, et al. Associations Between Migraine and Type 2 Diabetes in Women: Findings From the E3N Cohort Study. *JAMA Neurol.* 2019;76. doi:10.1001/jamaneurol.2018.3960
37. Schwedt TJ, Peplinski J, Garcia-Filion P, Berisha V. Altered speech with migraine attacks: A prospective, longitudinal study of episodic migraine without aura. *Cephalalgia.* 2019;39: 722.
38. Towards the Objective Speech Assessment of Smoking Status based on Voice Features: A Review of the Literature. *J Voice.* 2023;37: 300.e11–300.e20.
39. Pan A, Lucas M, Sun Q, van Dam RM, Franco OH, Manson JE, et al. Bidirectional Association between Depression and Type 2 Diabetes in Women. *Arch Intern Med.* 2010;170: 1884.
40. Demmer RT, Gelb S, Suglia SF, Keyes KM, Aiello AE, Colombo PC, et al. Sex Differences in the Association between Depression, Anxiety, and Type 2 Diabetes Mellitus. *Psychosom Med.* 2015;77: 467.
41. Greeley HP, Berg J, Friets E, Wilson J, Greenough G, Picone J, et al. Fatigue estimation using voice analysis. *Behav Res Methods.* 2007;39: 610–619.
42. Van Puyvelde M, Neyt X, McGlone F, Pattyn N. Voice Stress Analysis: A New Framework for Voice and Effort in Human Performance. *Front Psychol.* 2018;9: 414457.
43. Yaribeygi H, Maleki M, Butler AE, Jamialahmadi T, Sahebkar A. Molecular mechanisms linking stress and insulin resistance. *EXCLI J.* 2022;21: 317.

Figure Legends

Fig 1: General workflow.

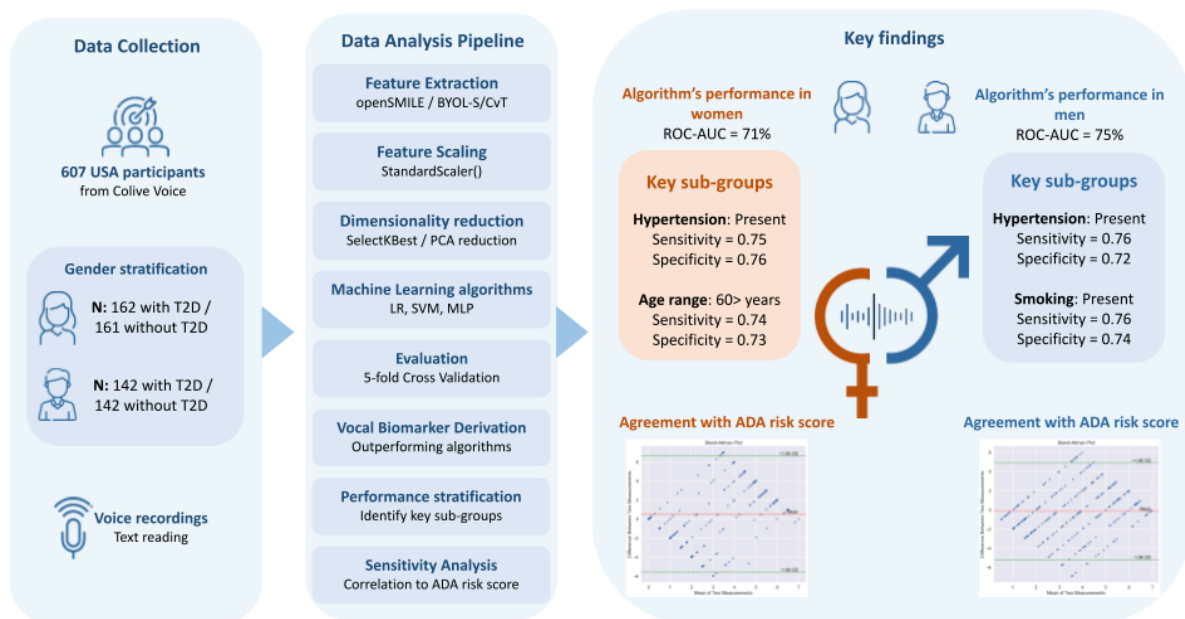
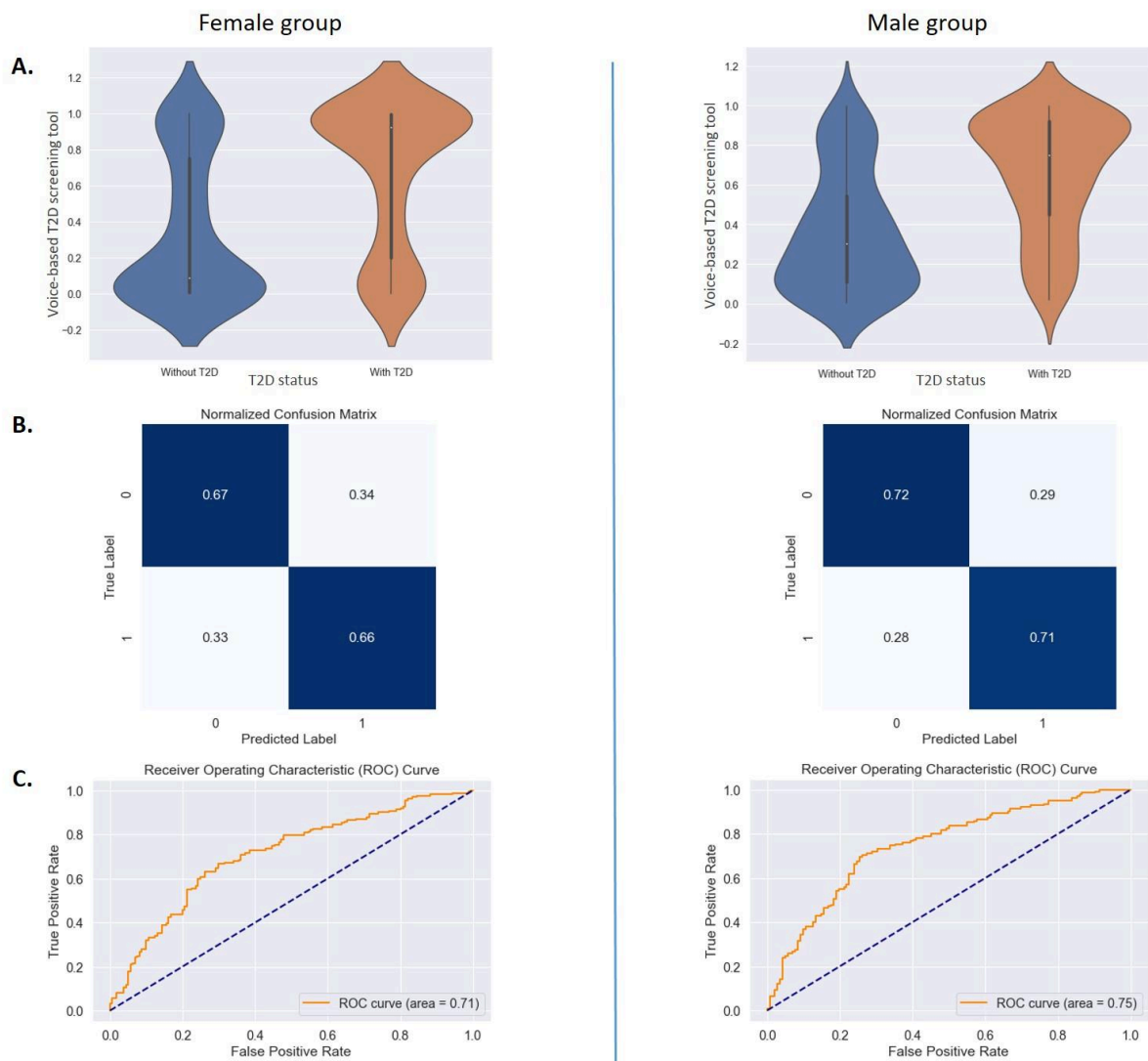


Fig 2: Voice-based T2D status detection algorithms' overall performance.



A: Predicted probability distribution by T2D status

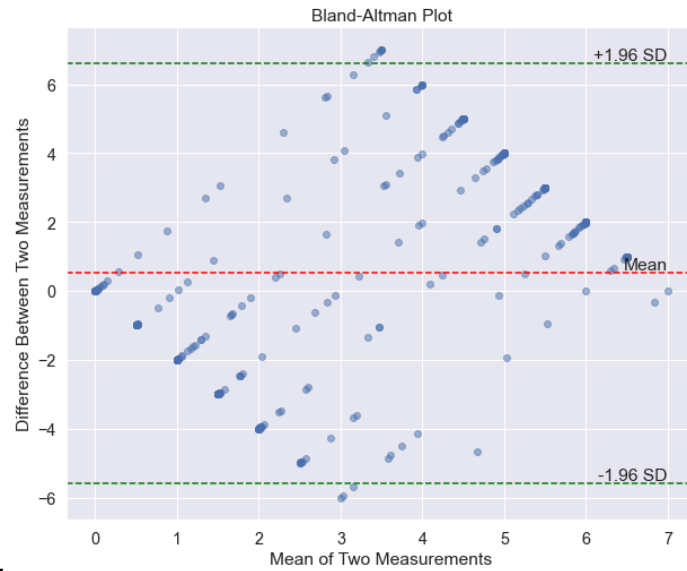
B: Confusion matrix of the selected models

C: AUC-ROC curve of the selected mode

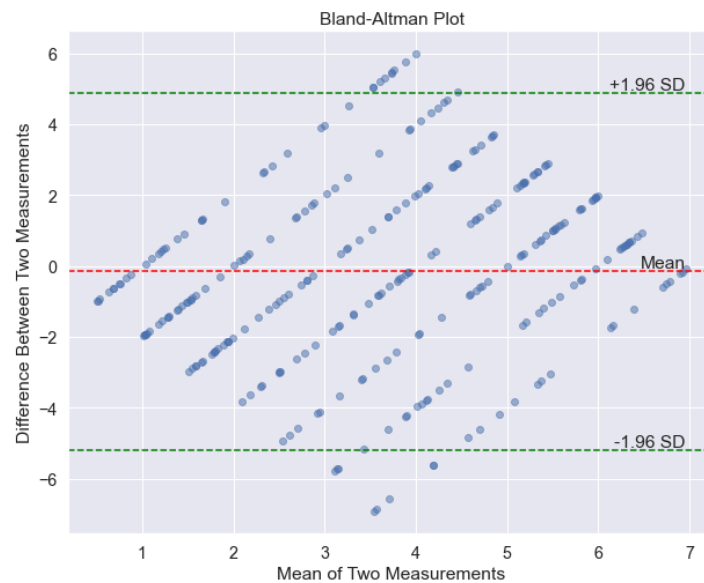
Supplementary Figure: Bland-Altman plot showing the agreement between the voice-based algorithms' predicted probability and the ADA risk score for both gender groups (A: Female group, B: Male group).

Note: The predicted probability was scaled by a factor of 7 for harmonization.

A.



B.



Publication Cover Sheet (Objective 3)

Title of the Publication:

Digital voice-based biomarker for monitoring respiratory quality of life: Findings from the Colive Voice study

Authors:

Vladimir Despotovic, **Abir Elbéji**, Kevser Fünfgeld, Mégane Pizzimenti, Hanin Ayadi, Petr V. Nazarov, Guy Fagherazzi

Data and Figure Contributions:

Abir Elbéji collaborated closely with Vladimir Despotovic in generating all data, results, and figures for the manuscript.

Contribution to Manuscript Writing:

Elbéji co-wrote substantial portions of the manuscript, specifically sections related to the methodological framework, machine learning models, data processing, and result interpretation. Elbéji collaborated closely with Despotovic on the original draft and contributed significantly to revisions. All authors contributed to reviewing and approving the final manuscript.

Statement of Contributions:

- Vladimir Despotovic: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Data curation, Conceptualization.
- Abir Elbéji: Writing – review & editing, Validation, Software, Methodology, Data curation.
- Kevser Fünfgeld: Writing – review & editing, Validation, Project administration, Conceptualization.
- Mégane Pizzimenti: Writing – review & editing, Validation, Project administration, Data curation.
- Hanin Ayadi: Writing – review & editing, Software, Methodology, Data curation.
- Petr V. Nazarov: Writing – review & editing, Supervision, Methodology, Conceptualization.
- Guy Fagherazzi: Writing – review & editing, Validation, Supervision, Methodology, Funding acquisition, Conceptualization

Digital Voice-Based Biomarker for Monitoring Respiratory Quality of Life: Findings from the Colive Voice Study

Vladimir Despotovic^a, Abir Elbéji^b, Kevser Fünfgeld^b, Mégane Pizzimenti^b, Hanin Ayadi^b, Petr V. Nazarov^{a,c}, Guy Fagherazzi^b

^a Bioinformatics Platform, Data Integration and Analysis Unit, Luxembourg Institute of Health, Strassen, Luxembourg

^b Deep Digital Phenotyping Research Unit, Department of Precision Health, Luxembourg Institute of Health, Strassen, Luxembourg

^c Multi-Omics Data Science, Department of Cancer Research, Luxembourg Institute of Health, Strassen, Luxembourg

Corresponding Author

Correspondence to [Vladimir Despotovic](#).

Abstract

Regular monitoring of the respiratory quality of life (RQoL) is essential in respiratory healthcare, facilitating prompt diagnosis and tailored treatment for chronic respiratory diseases. Voice alterations resulting from respiratory conditions create unique audio signatures that can potentially be utilized for disease screening or monitoring. Analyzing data from 1908 participants from the Colive Voice study, which collects standardized voice recordings alongside comprehensive demographic, epidemiological, and patient-reported outcome data, we evaluated various strategies to estimate RQoL from voice, including handcrafted acoustic features, standard acoustic feature sets, and advanced deep audio embeddings derived from pretrained convolutional neural networks. We compared models using clinical features alone, voice features alone, and a combination of both. The multimodal model combining clinical and voice features demonstrated the best performance, achieving an accuracy of 70.34% and an AUC of 0.77. A model utilizing voice features alone reached an accuracy of 65.57% and an area under the curve (AUC) of 0.70, improving over the clinical features by 1.5%. Our digital voice-based biomarker is capable of accurately predicting RQoL, either as an alternative to or in conjunction with clinical measures, and could be used to facilitate rapid screening and remote monitoring of respiratory health status.

Introduction

Monitoring chronic respiratory diseases or other conditions that affect breathing is a foundation of respiratory healthcare. Telemonitoring solutions can help in reducing the workload of clinicians, decrease hospital admissions and shorten clinician response time, thus enabling more timely intervention. Remote monitoring is of utmost importance for identifying clinically relevant deterioration in the Respiratory Quality of Life (RQoL) and may be used as a prognostic tool for chronic respiratory conditions, such as Chronic Obstructive Respiratory Disease (COPD) or asthma. A recent study proves that a decrease in RQoL by 4 points over a period of one year, measured by the St George's Respiratory Questionnaire (SGRQ)¹, was associated with increased hospitalization and mortality. Besides SGRQ, other questionnaires have been also developed for estimating RQoL, including Chronic Respiratory Disease Questionnaire (CRDQ)², Breathing Problems Questionnaire (BPQ)³, and VQ11⁴, just to name a few. Although questionnaires are considered essential in epidemiological studies, they are subjective, prone to biases and time-consuming; therefore, investigating alternative methods, such as analyzing voice characteristics, may provide valuable, scalable, easy-to-use solutions into assessing RQoL, requiring no invasive or cumbersome equipment, only a smartphone to record the voice.

The voice is a result of the airstream initiated in the lungs and respiratory airways, and passed through the larynx, causing the vibration of vocal folds, and furthermore through the oral and nasal cavity, where the sound is shaped and articulated. Respiratory diseases can alter the voice production process, resulting in distinctive changes in voice. Previous studies have shown that inspiratory closure of vocal folds, which causes refractory breathlessness, occurs frequently in COPD⁵. Changes in breathing and voice are highly correlated with altered lung function in patients with COPD⁶, most likely affected by respiratory and muscle damage⁷. Acoustic features extracted from the speech are clearly distinctive during COPD exacerbation and stable periods⁸, and are even distinguishable up to 7 days before the onset of symptoms⁶. Therefore, they could be used as an early warning system for COPD exacerbation.

Decreased voice-related quality of life, persistent cough and laryngeal dysfunction are also associated with up to 88% of patients with severe asthma⁹. Abnormal movements of vocal folds are caused by muscle tension in the vocal folds and larynx⁹. Vocal signatures extracted from voice recordings can be used to identify asthma worsening as a substitute to measures of lung function¹⁰.

There are multiple advantages of monitoring respiratory diseases using voice recordings. The technology is non-invasive, cost-efficient and practical, requiring only smartphones to capture the voice; thus, could be used from patients' homes for real-life remote monitoring in-between clinical visits or as a screening tool. Vocal biomarkers extracted from smartphone voice recordings were already used to identify pulmonary hypertension¹¹, and to monitor the recovery process of patients with influenza¹². A number of studies for screening of COVID-19 from voice and cough smartphone recordings has recently appeared, either for the detection of COVID-19^{13–16}, or for discriminating between the symptomatic and asymptomatic cases¹⁷.

Contrary to the previous research works which were mostly focused on the identification and/or monitoring of respiratory diseases from voice, in this paper we investigated whether RQoL can be assessed from voice features. Instead of targeting a single respiratory disease, we analyze RQoL in a general population containing participants with multiple respiratory conditions (e.g. asthma, COPD) as well as participants with no history of respiratory diseases, by stratifying them according to VQ11 scores, and comparing voice signatures extracted from sustained vowel phonation recordings. As an objective measure, vocal biomarkers can increase the reliability of screening based only on subjective self-reports. To support this hypothesis, we used data from the international, multilingual Colive Voice initiative to show that voice can be utilized as a universal biomarker for monitoring chronic respiratory conditions, either alone, or in addition to clinical parameters extracted from self-administered questionnaires. To our knowledge, this is the first study that proposes a multimodal approach combining voice features with clinical data.

Results

Study design

Colive Voice (<https://www.colivevoice.org>) is an international digital health study established and led by the Luxembourg Institute of Health which aims at identifying vocal biomarkers for remote monitoring and screening of various chronic diseases and frequent health symptoms. The multilingual audio databank is collected in four languages (English, French, German and Spanish) and contains recordings of multiple vocal tasks, including sustained vowel phonation, coughing, breathing, reading and counting. Voice recordings are associated with annotated clinical and demographic data, providing an in-depth patient characterization with validated disease-specific questionnaires on symptoms, treatments and quality of life. Colive Voice has been hosted online since June 2021 and is open for participation to anyone, under the condition that: 1) they sign the consent form and 2) they are at least 15 years old.

Assessment of respiratory quality of life (RQoL)

Part of the study is dedicated to the investigation of RQoL in the general population with voice recordings collected via a large-scale crowdsourced campaign, and accompanied with annotations of RQoL via self-administered VQ11 questionnaire, as well as clinical and demographic data.

Unlike SGRQ and BPQ, which are extensive (76 items in SGRQ and 33 items in BPQ) with complex scoring, making them unsuitable for repeated evaluations in clinical practice as well as a regular use in real life, VQ11 is a brief questionnaire with only 11 items distributed across functional components (3 items), psychological components (4 items) and social components (4 items). Although much simpler and faster to record, VQ11 shows a high correlation with SGRQ⁴. Each item in VQ11 is represented by five categories (not at all, a little, moderately, much, extremely) which reflect the participant's feeling about the statement associated with a particular item, and can be represented by a value from 1 to 5. The total score is obtained by summing all individual items, leading to a score between 11 and 55 with a lower value indicating better RQoL⁴. We stratify the participants in the study into two categories using the cut-off VQ11 score of 22: 1) Impaired RQoL (VQ11 \geq 22), and 2) normal RQoL (VQ11<22)^{18,19}.

Since the number of participants with impaired RQoL was significantly lower than the normal RQoL, we select a balanced subset matched by age and gender composed of 1908 sustained vowel recordings in total, equally distributed between two groups.

Voice recording task

We use in this paper sustained phonation of a single vowel /a/ produced at a comfortable pitch and loudness as long as possible, since it provides valuable information about the pulmonary function, and in addition, it is less susceptible to language bias, which may be present in the multi-lingual data collection. Reduced pulmonary function leads to decreased airflow necessary to support phonation²⁰, which in turn is reflected in reduced RQoL.

Evaluation of RQoL from socio-demographic/clinical data

Before evaluating the relevance of vocal biomarkers for estimating RQoL, we set up a baseline experiment where only socio-demographic data (Body Mass Index (BMI), smoking habits) and clinical data (day and night coughing, chest pain, sore throat, as well as associated diseases such as asthma and COPD) from the participants' self-reports were used for prediction of RQoL status. Categorical variables were encoded as one-hot representations, leading to 23 features in total. Performance is averaged over 5 folds (Table 1), with the best area under the receiver operating characteristic curve (AUC) of 0.70, and accuracy of 64.1% obtained using the Logistic Regression (LR) classifier. We also presented the feature importance based on the mean impurity decrease for the Random Forest (RF) model in Figure 1, revealing that BMI is the most important socio-demographic variable, followed by clinical symptoms related to day and night coughing.

Evaluation of RQoL from voice recordings

We investigated whether voice-related information could be used as a digital biomarker for RQoL. To that end, we extracted a set of handcrafted audio features (Suppl. Table 1), as well as two widely used general audio feature sets (eGeMAPS and ComParE). In addition to this, four state-of-the-art deep audio embeddings are evaluated (VGGish, YAMNet, OpenL3, BYOL-A) which proved to be highly competitive across multiple audio tasks. The features were either fed directly to the classifier, or in the case of deep audio embeddings, underwent Principal Component Analysis (PCA) to reduce the dimensionality of feature vectors. The results for the assessment of RQoL from voice were provided in Table 2, with the best performance reaching AUC equal to 0.7 and accuracy of 65.57% using BYOL-A

deep audio embeddings. BYOL-A substantially outperforms all other feature extraction techniques by over 2%.

To highlight the characteristics of sustained vowel phonation labeled with normal and impaired RQoL, we showed in Figure 2 spectrograms of two participants matched by age and gender (males, 67 years old): one with normal RQoL without the history of pulmonary diseases, but with a diagnosed COVID-19 more than 3 weeks before the recording was made; and one with extremely impaired RQoL (VQ11 score: 46) diagnosed with asthma-COPD overlap syndrome. Even though the normal RQoL example is actually a boundary case (VQ11 score: 21, cut-off value 22), the differences in spectrograms are clearly visible. While the normal RQoL recording is represented by uninterrupted phonation, with clearly distinctive harmonics (Figure 2a), impaired RQoL recording is characterized by strangled voice with multiple stoppages and voice breaks, and increased energy areas in higher frequency bands, which are most likely caused by aperiodic noise produced at a glottal constriction (Figure 2b). Furthermore, the absence of higher harmonics above 1kHz can be observed throughout the spectrogram, and as phonation progresses, even the adjacent lower harmonics become smeared and more difficult to distinguish. However, for the impaired RQoL voice recordings with VQ11 score closer to cut-off value, the differences are not so distinct.

Evaluation of RQoL from fused socio-demographic/clinical data and voice recordings

Finally, by fusing socio-demographic/clinical with voice features, we can quantify how much voice features can boost the performance of the socio-demographic and clinical data, uncovering the full potential of the multimodal data fusion. The results for the assessment of RQoL from multimodal features are provided in Table 3, whereas the comparison of the best-performing machine learning model for the socio-demographic/clinical features only, voice features only and multimodal features obtained after their fusion is presented in Figure 3. By using intermediate fusion (feature level fusion) we show that clinical data extracted from questionnaires and voice features obtained as the higher-level representations extracted from raw audio signals are complementary, leading to a substantial performance boost (accuracy equal to 70.34% and AUC equal to 0.77 using the combination of BYOL-A audio embeddings and socio-demographic/clinical features). Note that specificity is, in general, higher than sensitivity for all models, i.e. the models are still better at predicting normal than impaired RQoL. This is also visible from the confusion matrix of the best-performing model (fused BYOL-A and socio-demographic/clinical features, trained with LR classifier) shown in Figure 4a, where it is clear that the number of false negatives is substantially larger than the number of false positives. Using the Brier score as a measure of calibration, the same multimodal model achieves the lowest average Brier score over all folds equal to 0.19 with a nearly linear calibration curve, as shown in Figure 4b. Figure 4c displays the ROC curve of the best-performing model.

Finally, since our objective was to quantify how much the vocal biomarkers increase the reliability of screening based only on subjective self-reports, net reclassification improvement (NRI) was used to estimate the improvement in performance after fusing vocal biomarkers with socio-demographic/clinical predictors. Table 3 reveals that vocal biomarkers indeed improve the predictive capability of demographic and clinical variables for all acoustic features, with the biggest improvement measured by NRI of 0.19 for eGeMAPS features modeled with SVM.

Discussion

In this large international study, we have developed a digital voice-based biomarker for monitoring of RQoL using a combination of standard self-reported clinical information and voice-related features. We have shown that voice brings complementary information to improve the performances of the predictive model and increase the reliability of screening based only on subjective self-reports, reaching a full potential when both clinical and voice modalities are used conjointly in a multimodal setup.

RQoL has been evaluated from socio-demographic and clinical factors in various respiratory diseases, but mainly focusing on a single disorder, such as COPD^{21 22}, asthma²³, idiopathic pulmonary fibrosis²⁴, or COVID-19²⁵. There were also attempts to investigate the effect of several respiratory diseases simultaneously on RQoL by using a multicase-control design, where the use cases were COPD, asthma, allergic and non-allergic rhinitis²⁶. However, limited efforts were made to evaluate RQoL in the general population. A large five-year cohort study in Malawi was carried out to investigate the high prevalence of reduced lung function in Sub-Saharan Africa and its association with RQoL in the general population²⁷.

To establish a baseline for evaluation of RQoL from vocal biomarkers in the general population, we first estimated RQoL based on a number of socio-demographic (BMI, smoking habits) and clinical variables (day and night coughing, chest pain, sore throat, asthma, COPD). The feature importance analysis revealed that BMI is the most important socio-demographic variable. This confirms previous findings that BMI is significantly correlated with RQoL in COPD²⁸ and asthma²⁹, suggesting furthermore that RQoL of obese patients improves after weight reduction²⁸.

We further investigated whether digital biomarkers extracted from voice can act as a substitute for standard clinical measures estimated from questionnaires. Contrary to questionnaires which are mostly done during on-site clinical visits and can be tedious, voice recordings allow quick and easy-to-use data collection at patients' homes; thus, substantially facilitating remote patient monitoring³⁰. Our vocal biomarkers outperformed socio-demographic/clinical predictive factors by approximately 1.5% in terms of accuracy, confirming their potential to be a surrogate for clinical measures. The best-performing features are BYOL-A, which are general-purpose audio representations extracted with a model pretrained on a large amount of out-of-domain audio data in a self-supervised manner, i.e. requiring no annotations³¹. After freezing the convolutional layers, only the classification head is fine-tuned with the sustained vowel phonation collected within the Colive Voice study. This allows training the deep neural network models even with limited available voice data, and furthermore enables deploying for real-time inference, in applications that require low latency. However, deep audio embeddings such as BYOL-A suffer from limited interpretability, which might be an issue in a clinical application. Therefore, trade-off between performance and interpretability has to be considered when selecting the audio features.

Finally, fusing clinical and voice features in a multimodal setup allows focusing on different aspects of RQoL, localizing a broad range of information extracted from different modalities, and enabling more robust prediction models. The fusion of audio features with textual (word embeddings) and vision features (facial action units) has already been shown to improve the performance of unimodal approaches for the detection of clinical depression^{32 33}. A deep multimodal fusion model that learns indicators of Alzheimer's disease from audio and text modalities, as well as disfluency features, increases the predictive power of audio features³⁴. Fusion of speech, handwriting and gait data enables accurate evaluation of neurological state in different stages of Parkinson's disease³⁵. To the

best of our knowledge, there were no previous attempts to combine voice features with clinical data for application in healthcare. By using intermediate feature level fusion we proved that voice features and clinical variables extracted from self-administered questionnaires are indeed complementary, leading to improved performance in comparison to both unimodal approaches by almost 5% in terms of accuracy, and up to 7% in terms of AUC. The intermediate fusion has an advantage in flexibility of extracting marginal representations appropriate for each modality, and arguably reflects more closely the relationships between the modalities³⁶. To avoid producing high-dimensional joint feature representations, PCA was used to reduce the dimensionality of feature vectors coming from different modalities to the same length.

To further evaluate not only the ability of the model to accurately predict the class labels, but also the associated probability, the Brier score was used. The well-calibrated model is neither underconfident, nor overconfident, i.e. the true frequency of the positive label (impaired RQoL score in our case) against its predicted probability is approximately linear. This is confirmed by a solid average Brier score, and a calibration curve that does not deviate substantially from the perfectly calibrated model.

A major strength of this study is the fact that the dataset is acquired via a mobile app at participants' homes, i.e. in uncontrolled conditions close to real-world circumstances. This confirms the feasibility of using a digital voice-based biomarker to provide quantitative measurements of RQoL, and enable regular remote monitoring in real life without relying on costly, invasive or cumbersome equipment; thus, facilitating personalized and more timely treatment, according to the patient's needs and general health status. It is a step towards the development of scalable, non-invasive, easy-to-use and low-cost solutions for remote monitoring and rapid screening of respiratory health status.

However, a crowdsourced data collection poses multiple challenges and could be also observed as a limitation. There is a risk of acquiring low-quality answers from the self-administered questionnaires and introducing noise in the data, making it more difficult to infer the ground truth labels. We mitigated this risk by using a well-known, clinically validated questionnaire to assess RQoL. Recording voice using multiple devices, different qualities of microphones, and various recording conditions make data collection additionally challenging, resulting in different quality of audio recordings. For this purpose, we developed a proprietary data processing pipeline that harmonizes recordings and performs quality checks, but we cannot entirely exclude the possibility of having some low-quality recordings in our dataset.

To summarize, in this paper we developed a digital voice-based biomarker for monitoring RQoL in the general population. Our results confirm that vocal biomarkers can be a viable surrogate for standard clinical measures estimated from questionnaires, but also that the ultimate capacity is unlocked in a multimodal setup when clinical and voice data are used together. The best performance was obtained with a feature-level fusion of BYOL-A deep audio embeddings and socio-demographic/clinical variables, reaching an accuracy of over 70% and AUC of 0.77, a performance boost of over 2% in comparison to handcrafted acoustic features.

Methods

Data analysis pipeline

A full workflow of RQoL monitoring from data acquisition to the prediction of RQoL is shown in Figure 5. RQoL is estimated based on the VQ11 score and classified into impaired RQoL ($VQ11 \geq 22$) or normal RQoL ($VQ11 < 22$).

Data acquisition and preprocessing

Participants were recruited via an online crowdsourced campaign or through partnerships with various patient associations, academic institutions, hospitals, or other research initiatives (including Les Sentinelles and the ComPaRe study, AP-HP). The full list of partners is available on the Colive Voice website. Participants were invited to use an app (<https://app.colivevoice.org/>) accessible from participants' devices equipped with microphones (smartphone, tablet or laptop). Socio-demographic data include BMI and smoking habits, while clinical data contains information about day and night coughing, chest pain, sore throat, as well as associated diseases such as asthma and COPD from the participants' self-reported questionnaires. Categorical variables were encoded as one-hot representations, leading to 23 features in total.

Participants were advised to make voice recordings in a quiet environment without the external noise in order to preserve high-quality recordings. However, given that data is collected in uncontrolled conditions and to account for the challenges related to the use of different devices, microphones, and recording conditions for data collection, audio preprocessing using a proprietary pipeline was performed to harmonize and prepare the recordings and prepare them for the subsequent steps.

Statistical analysis

We utilized an independent two-tailed t-test to compare the means of groups with normal and impaired RQoL for continuous variables. For categorical variables, we applied a chi-square test. A p-value less than 0.05 indicates a statistically significant difference. Only variables that were statistically significant were used as socio-demographic and clinical features in further processing. Table 4 provides the study population characteristics.

Feature extraction and fusion

We first extracted a set of 72 handcrafted audio features, that contain time domain, spectral, cepstral, prosodic, and nonlinear dynamics features (Suppl. Table 1). Audio features were extracted using Surfboard³⁷, a Python library for feature extraction with application to the medical domain, as well as Parselmouth³⁸, a Python interface to Praat. We selected the audio features that are shown to be relevant for vocal biomarker research across multiple diseases.

In addition to this, we used standard audio feature sets, i.e. extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)³⁹ and ComParE, extracted using the openSMILE⁴⁰. The eGeMAPS is a minimalistic set of acoustic parameters for paralinguistic or clinical speech analysis which is composed of 88 energy/amplitude, frequency, spectral and temporal features, as well as statistical functionals applied to them (arithmetic mean, standard deviation, percentile). ComParE is a brute force audio feature set that contains 65 low-level acoustic descriptors and various statistical functionals applied to them, leading to a total of 6737 audio features.

Finally, we experiment with 4 different types of deep audio embeddings, i.e. VGGish⁴¹, YAMNet, OpenL3⁴², and BYOL-A³¹, which are state-of-the-art general audio features pretrained on large audio collections that are successfully used for a number of downstream tasks. Characteristics of different audio embeddings are provided in Suppl. Table 2.

VGGish is a pretrained convolutional neural network (CNN) mostly inspired by the VGG network used in computer vision. The network is adapted to accept 96x64 bin log-mel spectrograms at its input and extracts 128-dimensional embeddings from 960 ms segments of an audio signal. YAMNet employs the

Mobilenet v1 depthwise separable convolution architecture used with the same input as VGGish, but outputs 1024-dimensional embeddings for each 960 ms audio segment. Both VGGish and YAMNet are pretrained on the large-scale Audio Set dataset for audio event classification which contains more than 2 million of 10 s YouTube clips of sounds classified into 632 audio events. To summarize features across different audio segments and output the equal size feature vectors from recordings of different lengths average pooling was used.

OpenL3 uses CNN-based L3-Net for self-supervised learning via audio-visual correspondence, to learn whether a particular video frame corresponds to an audio frame; thus, requiring no annotations. The model is pretrained on two subsets of Audio Set, i.e. music and environmental subset, containing 296K and 195K clips respectively, and uses either 128 or 256 band Mel-spectrograms at the input, while the output audio embedding is 512 or 6144-dimensional vector for each 1s audio segment. We use a model pretrained on an environmental subset, with 256 band Mel-spectrograms and 6144-dimensional embeddings.

BYOL-A uses the Bootstrap Your Own Latent (BYOL) method for self-supervised learning of general-purpose image representations, adapted to work with audio. Normalized 96x64 bin log-mel spectrograms are used as an input, and two augmented versions of the input are created by shifting pitch and stretching time, which are further fed into two parallel networks (online and target network). The online network predicts the output representation of the target network, which is then iteratively updated as the exponential moving average of the parameters of the online network. The model is pretrained on the Audio Set dataset and produces 512, 1024 or 2048-dimensional general-purpose audio embeddings. We use 2048-dimensional embeddings.

In addition to voice features, we extracted the demographic/clinical data relevant for RQoL from the subjective self-reports. We used socio-demographic variables that were found statistically significant (body mass index (BMI), smoking habits), symptoms (day and night coughing, chest pain, sore throat), and associated diseases that can affect RQoL (asthma, COPD), as shown in Table 4. All categorical variables were encoded as one-hot representations, except for ensemble-based models (Random Forest, Extreme Gradient Boosting), where single feature representation was kept. One-hot encodings produce sparse feature vectors, which are not suitable for tree-based models, since splitting on such features produces a small gain, and is typically ignored in favor of continuous variables. Features were standardized before feeding them to classification models to put them on the same scale, i.e. all features have zero mean and unit standard deviation.

Given that the size of the audio feature vectors is substantially larger than the size of demographic/clinical features (up to 250 times larger for ComParE features), PCA was applied to audio embeddings prior to data fusion, to reduce their dimensionality to the first 23 principal components that explain most of the variance, to put the features from different modalities to equal dimension.

RQoL prediction

Features extracted in the previous section were fed into several classifiers: Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Multilayer Perceptron (MLP).

LR with L2 regularization is used to handle overfitting, as well as SVM with radial basis function kernel, where the model hyperparameters, i.e. the regularization parameter C and the kernel

coefficient γ are optimized using a grid search. Two ensemble models include RF and XGBoost. RF was composed of 500 fully grown trees (the optimal number of trees was determined after hyperparameter tuning), expanded until all leaves were pure or contained less than 2 samples, with the Gini index as the criterion for splitting the node, and the number of features at each split equal to the square root of the total number of features. All models are implemented using the scikit-learn 1.1.3 Python library.

XGBoost is a flexible and distributed gradient boosting algorithm, that allows for custom loss functions, as well as regularization techniques to mitigate the overfitting. We use XGBoost with 500 trees, L2 regularization and log loss objective function. XGBoost is implemented using the xgboost 1.5.0 Python library.

MLP was composed of two hidden layers with 256 neurons each and a ReLU activation function, followed by dropout layers for preventing overfitting with a dropout rate equal to 0.3, and an output layer with a sigmoid activation function is utilized in this paper. We used Adam optimizer, binary cross entropy loss function, batch size equal to 32, while the optimal learning rate (0.0001) and the number of epochs (30) are determined via grid search. Note that Adam has an adaptive per-parameter learning rate, which is computed using the initial learning rate as an upper limit. MLP is implemented using Tensorflow 2.9.1.

Evaluation

For evaluation of the model performance, we use accuracy, sensitivity, specificity, area under the receiver operating characteristic curve (AUC ROC), Brier score and NRI. Accuracy is the ratio of the number of correctly classified observations and the total number of observations. Sensitivity (true positive rate, recall) is the proportion of participants detected with impaired RQoL (true positives) among those who have impaired RQoL (true positives + false negatives), and shows the model's ability to correctly identify cases. Specificity (true negative rate) is the proportion of participants detected with normal RQoL (true negatives) among those who have normal RQoL (true negatives + false positives), and refers to the model's ability to correctly identify healthy controls. ROC curve plots sensitivity against false negative rate (1-specificity) at different classification thresholds, while AUC is an aggregated performance measure, which summarizes the ROC curve, with a value of 0.5 denoting random guess, and 1 denoting perfect classification.

To assess model calibration, i.e. the consistency between the predicted probability and the observations, the Brier score was used, which is the mean squared deviation of the predicted probability from the actual target. It is a value between 0 and 1, with a lower value indicating a better model.

Given the size of the dataset, to get reliable and robust performance estimates and preserve the class distribution across folds, we used stratified 5-fold cross-validation⁴³. Data is split into five folds, four merged and used for training, and the remaining one for testing. The process is repeated 5 times, so that each fold was used exactly once for testing, and the performance is then averaged over all folds.

Finally, since our aim was to quantify how much voice-related information can improve the reliability of RQoL screening on top of standard clinical features, we used NRI to estimate the improvement in performance due to adding vocal biomarkers to a set of socio-demographic and clinical predictors. The value can range from -2 to 2, with a bigger value indicating a larger improvement.

Data availability

The dataset generated and analysed during the current study is not publicly available due to the sensitivity of the data. Voice can be used as biometric data to uniquely identify an individual, therefore constitutes sensitive personal data. Anonymised voice data will be made available for academic research upon reasonable request directed to the principal investigator Guy Fagherazzi, guy.fagherazzi@lih.lu, and signing a Data Transfer Agreement.

Code availability

The underlying code for this study is not publicly available for proprietary reasons.

Acknowledgements

Colive Voice study is funded by the Luxembourg Institute of Health. The funder played no role in the study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

We would like to thank all participants that contributed to Colive Voice study, as well as our partners for their help in recruiting new participants. Special thanks go to Aurélie Fischer, Philippe Kayser, Luigi De Giovanni, Michael Schnell and Aurore Dobosz for their substantial contribution to the Colive Voice study.

Ethics declarations

Colive Voice has been approved by the National Research Ethics Committee in Luxembourg (N° 202103/01) in March 2021. Informed written consent was obtained electronically via the Colive Voice application from all participants in the study. The Colive Voice study protocol is also registered on ClinicalTrials.gov (NCT04848623).

Competing interests

All authors declare no financial or non-financial competing interests.

Author contributions

VD, GF, AE, HA contributed to preprocessing, analysis and/or interpretation of data. GF contributed to the conception and design of the study. GF, MP contributed to the data acquisition. VD drafted the article. GF, AE, KF, MP and PN critically reviewed the article. All authors read and approved the final manuscript.

References

1. Jones, P. W., Quirk, F. H., Baveystock, C. M. & Littlejohns, P. A self-complete measure of health status for chronic airflow limitation. The St. George's Respiratory Questionnaire. *Am. Rev. Respir.*

- Dis.* **145**, 1321–1327 (1992).
2. Chauvin, A., Rupley, L., Meyers, K., Johnson, K. & Eason, J. Research Corner Outcomes in Cardiopulmonary Physical Therapy: Chronic Respiratory Disease Questionnaire (CRQ): *Cardiopulm. Phys. Ther. J.* **19**, 61–67 (2008).
 3. Hyland, M. E., Bott, J., Singh, S. & Kenyon, C. A. Domains, constructs and the development of the breathing problems questionnaire. *Qual. Life Res. Int. J. Qual. Life Asp. Treat. Care Rehabil.* **3**, 245–256 (1994).
 4. Ninot, G., Soyeze, F. & Préfaut, C. A short questionnaire for the assessment of quality of life in patients with chronic obstructive pulmonary disease: psychometric properties of VQ11. *Health Qual. Life Outcomes* **11**, 179 (2013).
 5. Leong, P. *et al.* Inspiratory vocal cord closure in COPD. *Eur. Respir. J.* **55**, (2020).
 6. Khan, M. M. D. A., Naval, P. P., Kulshreshtha, R., Venneti, S. & Singh, A. VOICE-BASED MONITORING OF COPD. *CHEST* **160**, A2173–A2174 (2021).
 7. da Silva, G. dos A. P., Feltrin, T. D., Pichini, F. dos S., Cielo, C. A. & Pasqualoto, A. S. Quality of Life Predictors in Voice of Individuals With Chronic Obstructive Pulmonary Disease. *J. Voice* (2022) doi:10.1016/j.jvoice.2022.05.017.
 8. Nallanthighal, V. S., 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)* 9097–9101 (2022). doi:10.1109/ICASSP43922.2022.9747785.
 9. Vertigan, A. E., Kapela, S. L. & Gibson, P. G. Laryngeal Dysfunction in Severe Asthma: A Cross-Sectional Observational Study. *J. Allergy Clin. Immunol. Pract.* **9**, 897–905 (2021).
 10. Alam, Md. Z. *et al.* Predicting Pulmonary Function From the Analysis of Voice: A Machine Learning Approach. *Front. Digit. Health* **4**, (2022).
 11. Sara, J. D. S. *et al.* Non-invasive vocal biomarker is associated with pulmonary hypertension. *PLoS ONE* **15**, e0231441 (2020).

12. Tracey, B. *et al.* Voice Biomarkers of Recovery From Acute Respiratory Illness. *IEEE J. Biomed. Health Inform.* **26**, 2787–2795 (2022).
13. Han, J. *et al.* Sounds of COVID-19: exploring realistic performance of audio-based digital testing. *Npj Digit. Med.* **5**, 1–9 (2022).
14. Pah, N. D., Indrawati, V. & Kumar, D. K. Voice Features of Sustained Phoneme as COVID-19 Biomarker. *IEEE J. Transl. Eng. Health Med.* **10**, 1–9 (2022).
15. Al Ismail, M., Deshmukh, S. & Singh, R. Detection of Covid-19 Through the Analysis of Vocal Fold Oscillations. in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 1035–1039 (2021). doi:10.1109/ICASSP39728.2021.9414201.
16. Despotovic, V., Ismael, M., Cornil, M., Call, R. M. & Fagherazzi, G. Detection of COVID-19 from voice, cough and breathing patterns: Dataset and preliminary results. *Comput. Biol. Med.* **138**, 104944 (2021).
17. Fagherazzi, G. *et al.* A voice-based biomarker for monitoring symptom resolution in adults with COVID-19: Findings from the prospective Predi-COVID cohort study. *PLOS Digit. Health* **1**, e0000112 (2022).
18. Anane, I., Guezguez, F., Knaz, H. & Ben Saad, H. How to Stage Airflow Limitation in Stable Chronic Obstructive Pulmonary Disease Male Patients? *Am. J. Mens Health* **14**, 1557988320922630 (2020).
19. Zysman, M. *et al.* COPD burden on sexual well-being. *Respir. Res.* **21**, 311 (2020).
20. Yasien, D. G., Hassan, E. S. & Mohamed, H. A. Phonatory function and characteristics of voice in recovering COVID-19 survivors. *Eur. Arch. Otorhinolaryngol.* **279**, 4485–4490 (2022).
21. Pati, S. *et al.* An assessment of health-related quality of life among patients with chronic obstructive pulmonary diseases attending a tertiary care hospital in Bhubaneswar City, India. *J. Fam. Med. Prim. Care* **7**, 1047–1053 (2018).
22. Bove, D. G., Lavesen, M. & Lindegaard, B. Characteristics and health related quality of life in a population with advanced chronic obstructive pulmonary disease, a cross-sectional study. *BMC*

Palliat. Care **19**, 84 (2020).

23. Gonzalez-Barcala, F.-J., de la Fuente-Cid, R., Tafalla, M., Nuevo, J. & Caamaño-Isorna, F. Factors associated with health-related quality of life in adults with asthma. A cross-sectional study. *Multidiscip. Respir. Med.* **7**, 32 (2012).
24. Cox, I. A. *et al.* Health-related quality of life of patients with idiopathic pulmonary fibrosis: a systematic review and meta-analysis. *Eur. Respir. Rev.* **29**, (2020).
25. Meys, R. *et al.* Generic and Respiratory-Specific Quality of Life in Non-Hospitalized Patients with COVID-19. *J. Clin. Med.* **9**, 3993 (2020).
26. Cappa, V. *et al.* Health-related quality of life varies in different respiratory disorders: a multi-case control population based study. *BMC Pulm. Med.* **19**, 32 (2019).
27. Njoroge, M. W. *et al.* Changing lung function and associated health-related quality-of-life: A five-year cohort study of Malawian adults. *eClinicalMedicine* **41**, (2021).
28. Huber, M. B. *et al.* The relationship between body mass index and health-related quality of life in COPD: real-world evidence based on claims and survey data. *Respir. Res.* **21**, 291 (2020).
29. Sergeeva, G. & Emelyanov, A. Body mass index and quality of life in patients with asthma. *Eur. Respir. J.* **38**, (2011).
30. Fischer, A., Elbeji, A., Aguayo, G. & Fagherazzi, G. Recommendations for Successful Implementation of the Use of Vocal Biomarkers for Remote Monitoring of COVID-19 and Long COVID in Clinical Practice and Research. *Interact. J. Med. Res.* **11**, e40655 (2022).
31. Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N. & Kashino, K. BYOL for Audio: Exploring Pre-Trained General-Purpose Audio Representations. *IEEEACM Trans. Audio Speech Lang. Process.* **31**, 137–151 (2023).
32. Muzammel, M., Salam, H. & Othmani, A. End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis. *Comput. Methods Programs Biomed.* **211**, 106433 (2021).
33. Rohanian, M., Hough, J. & Purver, M. Detecting Depression with Word-Level Multimodal Fusion.

- in *Interspeech 2019* 1443–1447 (ISCA, 2019). doi:10.21437/Interspeech.2019-2283.
34. Rohanian, M., Hough, J. & Purver, M. Multi-Modal Fusion with Gating Using Audio, Lexical and Disfluency Features for Alzheimer’s Dementia Recognition from Spontaneous Speech. *Interspeech 2020* 2187–2191 (2020) doi:10.21437/Interspeech.2020-2721.
 35. Vásquez-Correa, J. C. *et al.* Multimodal Assessment of Parkinson’s Disease: A Deep Learning Approach. *IEEE J. Biomed. Health Inform.* **23**, 1618–1630 (2019).
 36. Stahlschmidt, S. R., Ulfenborg, B. & Synnergren, J. Multimodal deep learning for biomedical data fusion: a review. *Brief. Bioinform.* **23**, bbab569 (2022).
 37. Lenain, R., Weston, J., Shivkumar, A. & Fristed, E. Surfboard: Audio Feature Extraction for Modern Machine Learning. Preprint at <https://doi.org/10.48550/arXiv.2005.08848> (2020).
 38. Jadoul, Y., Thompson, B. & de Boer, B. Introducing Parselmouth: A Python interface to Praat. *J. Phon.* **71**, 1–15 (2018).
 39. Eyben, F. *et al.* The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Trans. Affect. Comput.* **7**, 190–202 (2016).
 40. Eyben, F., Weninger, F., Gross, F. & Schuller, B. Recent developments in openSMILE, the munich open-source multimedia feature extractor. in *Proceedings of the 21st ACM international conference on Multimedia* 835–838 (Association for Computing Machinery, 2013). doi:10.1145/2502081.2502224.
 41. Hershey, S. *et al.* CNN Architectures for Large-Scale Audio Classification. in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017).
 42. Cramer, J., Wu, H.-H., Salamon, J. & Bello, J. P. Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings. in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 3852–3856 (2019). doi:10.1109/ICASSP.2019.8682475.
 43. de Hond, A. A. H. *et al.* Perspectives on validation of clinical predictive algorithms. *Npj Digit. Med.* **6**, 1–3 (2023).

Table 1 RQoL assessment based on demographic and clinical features

ML model	Accuracy [%]	Sensitivity [%]	Specificity [%]	AUC	Brier score
LR	64.1 (1.71)	59.64 (3.98)	68.55 (4.86)	0.70 (0.03)	0.22 (0.01)
SVM	62.79 (2.11)	54.4 (6.48)	71.17 (4.74)	0.67 (0.03)	0.23 (0.01)
RF	61.43 (2.02)	53.25 (4.8)	69.59 (5.57)	0.66 (0.03)	0.24 (0.01)
XGBoost	62.26 (2.41)	53.25 (4.7)	71.27 (5.44)	0.66 (0.03)	0.24 (0.01)
MLP	63.84 (2.17)	56.39 (3.94)	71.27 (4.37)	0.70 (0.03)	0.22 (0.01)

Table 2 RQoL assessment based on handcrafted voice features, standard acoustic feature sets, and deep audio embeddings

ML model	Features	Accuracy [%]	Sensitivity [%]	Specificity [%]	AUC	Brier score
LR	Handcrafted	60.48 (3.42)	56.3 (4.75)	64.68 (4.64)	0.67 (0.03)	0.23 (0.01)
	eGeMAPS	59.54 (2.37)	53.25 (2.08)	65.82 (3.34)	0.64 (0.03)	0.23 (0.01)
	ComParE	62.58 (1.84)	54.62 (3.17)	70.55 (1.49)	0.67 (0.02)	0.23 (0)
	VGGish	61.69 (1.82)	59.86 (3.51)	63.52 (1.15)	0.66 (0.02)	0.23 (0.01)
	YAMNet	61.53 (2.63)	53.67 (2.74)	69.39 (3.32)	0.67 (0.01)	0.23 (0.01)
	OpenL3	62.16 (2.86)	56.71 (4.04)	67.61 (2.25)	0.67 (0.03)	0.23 (0.01)
	BYOL-A	65.57 (1.66)	59.96 (3.46)	71.17 (1.92)	0.70 (0.02)	0.22 (0)
SVM	Handcrafted	62.11 (3.84)	55.15 (4.65)	69.07 (4.56)	0.67 (0.04)	0.23 (0.01)
	eGeMAPS	58.91 (1.91)	45.91 (2.97)	71.90 (3.95)	0.63 (0.02)	0.23 (0)
	ComParE	63.37 (1.49)	47.70 (3.58)	79.04 (1.47)	0.66 (0.03)	0.23 (0.01)
	VGGish	60.74 (1.38)	52.62 (2.13)	68.87 (1.79)	0.66 (0.02)	0.23 (0)
	YAMNet	62.06 (1.69)	48.12 (2.55)	76.00 (2.30)	0.67 (0.01)	0.23 (0)
	OpenL3	63.58 (2.10)	57.23 (2.46)	69.92 (2.11)	0.67 (0.03)	0.23 (0.01)
	BYOL-A	63.99 (1.58)	52.2 (3.10)	75.79 (2.37)	0.69 (0.02)	0.22 (0.01)
RF	Handcrafted	60.64 (2.95)	58.49 (4.38)	62.79 (4.17)	0.65 (0.03)	0.23 (0.01)
	eGeMAPS	58.65 (2.17)	55.45 (2.96)	61.85 (2.03)	0.62 (0.02)	0.24 (0)
	ComParE	61.16 (1.92)	55.98 (3.06)	66.35 (2.29)	0.64 (0.03)	0.23 (0.01)
	VGGish	60.38 (1.52)	57.23 (1.77)	63.52 (1.99)	0.64 (0.02)	0.23 (0)
	YAMNet	61.79 (1.45)	57.34 (3.49)	66.25 (2.11)	0.66 (0.02)	0.23 (0)
	OpenL3	62.26 (1.66)	55.77 (3.26)	68.76 (0.75)	0.66 (0.03)	0.28 (0.01)
	BYOL-A	62.37 (2.50)	58.18 (3.75)	66.56 (1.89)	0.67 (0.03)	0.23 (0.01)
XGBoost	Handcrafted	58.07 (2.75)	57.23 (3.05)	58.91 (4.98)	0.62 (0.02)	0.31 (0.01)
	eGeMAPS	57.91 (1.18)	56.50 (1.32)	59.33 (2.75)	0.61 (0.01)	0.32 (0.01)
	ComParE	58.12 (2.28)	54.52 (4.90)	61.74 (2.39)	0.62 (0.03)	0.32 (0.02)
	VGGish	56.87 (1.12)	56.40 (1.49)	57.34 (1.58)	0.60 (0.02)	0.33 (0.01)
	YAMNet	58.76 (2.98)	57.97 (4.36)	59.54 (3.17)	0.63 (0.02)	0.32 (0.02)
	OpenL3	57.71 (2.70)	54.72 (2.78)	60.7 (4.32)	0.63 (0.02)	0.32 (0.01)
	BYOL-A	58.75 (2.72)	56.40 (3.15)	61.11 (3.86)	0.63 (0.02)	0.32 (0.01)
MLP	Handcrafted	61.58 (3.10)	58.08 (5.43)	65.09 (3.01)	0.66 (0.03)	0.23 (0.01)
	eGeMAPS	60.95 (1.56)	52.62 (1.50)	69.29 (2.88)	0.64 (0.02)	0.24 (0.01)
	ComParE	58.81 (2.30)	54.94 (7.62)	62.68 (5.99)	0.63 (0.03)	0.25 (0.01)
	VGGish	60.22 (2.63)	57.03 (5.72)	63.42 (3.63)	0.65 (0.03)	0.23 (0.01)
	YAMNet	62.26 (0.55)	54.72 (2.23)	69.81 (3.00)	0.67 (0.01)	0.23 (0)
	OpenL3	60.69 (1.72)	56.92 (5.83)	64.47 (2.83)	0.64 (0.02)	0.24 (0.01)
	BYOL-A	62.53 (3.27)	54.20 (5.21)	70.86 (5.05)	0.67 (0.03)	0.23 (0.01)

Table 3 RQoL assessment based on fused socio-demographic/clinical and voice features

ML model	Features	Accuracy [%]	Sensitivity [%]	Specificity [%]	AUC	Brier score	NRI
LR	Handcrafted	67.77 (2.07)	63.42 (2.47)	72.12 (2.06)	0.74 (0.01)	0.20 (0.01)	0.15
	eGeMAPS	67.14 (2.24)	61.53 (2.98)	72.74 (3.27)	0.73 (0.02)	0.21 (0.01)	0.16
	ComParE	68.92 (0.85)	64.26 (2.17)	73.59 (0.78)	0.75 (0.02)	0.20 (0.01)	0.14
	VGGish	68.92 (2.18)	65.62 (3.55)	72.22 (1.56)	0.75 (0.02)	0.20 (0.01)	0.14
	YAMNet	69.34 (1.65)	64.05 (2.27)	74.63 (2.60)	0.76 (0.02)	0.20 (0.01)	0.17
	OpenL3	69.76 (2.82)	65.83 (3.15)	73.69 (2.53)	0.76 (0.02)	0.20 (0.01)	0.16
	BYOL-A	70.34 (1.82)	66.78 (4.60)	73.90 (1.44)	0.77 (0.02)	0.19 (0.01)	0.10
SVM	Handcrafted	67.87 (1.24)	60.38 (1.80)	75.37 (2.40)	0.74 (0.01)	0.20 (0)	0.12
	eGeMAPS	67.24 (1.38)	56.08 (1.75)	78.41 (1.67)	0.73 (0.01)	0.21 (0.01)	0.19
	ComParE	66.98 (1.79)	58.70 (2.49)	75.26 (2.07)	0.72 (0.03)	0.21 (0.01)	0.13
	VGGish	68.66 (1.41)	61.01 (2.57)	76.31 (1.97)	0.75 (0.02)	0.20 (0.01)	0.16
	YAMNet	68.76 (1.94)	57.65 (2.75)	79.88 (2.65)	0.76 (0.02)	0.20 (0.01)	0.16
	OpenL3	62.95 (1.42)	61.64 (2.34)	64.26 (2.84)	0.69 (0.02)	0.25 (0.01)	0
	BYOL-A	69.18 (1.36)	61.75 (2.90)	76.63 (2.19)	0.76 (0.01)	0.20 (0)	0.13
RF	Handcrafted	66.41 (1.51)	64.89 (3.39)	67.92 (2.53)	0.74 (0.02)	0.21 (0)	0.12
	eGeMAPS	65.09 (2.04)	61.53 (2.61)	68.66 (1.76)	0.72 (0.02)	0.21 (0)	0.13
	ComParE	67.98 (2.18)	64.58 (4.37)	71.39 (2.22)	0.73 (0.02)	0.21 (0)	0.14
	VGGish	66.88 (2.03)	64.89 (2.92)	68.87 (2.36)	0.73 (0.02)	0.21 (0)	0.13
	YAMNet	67.98 (1.74)	65.94 (3.02)	70.02 (0.86)	0.74 (0.02)	0.21 (0)	0.13
	OpenL3	68.08 (1.92)	65.31 (3.22)	70.86 (1.22)	0.74 (0.02)	0.24 (0.02)	0.13
	BYOL-A	67.35 (1.51)	65.00 (3.82)	69.71 (0.93)	0.75 (0.01)	0.21 (0)	0.10
XGBoost	Handcrafted	65.04 (1.34)	64.15 (2.58)	65.94 (3.14)	0.71 (0.01)	0.27 (0.01)	0.14
	eGeMAPS	63.84 (2.63)	62.68 (2.51)	64.99 (4.39)	0.70 (0.01)	0.27 (0.01)	0.12
	ComParE	66.14 (1.37)	62.79 (3.11)	69.50 (1.92)	0.70 (0.02)	0.27 (0.02)	0.16
	VGGish	65.36 (2.08)	64.47 (2.26)	66.25 (3.08)	0.70 (0.02)	0.27 (0.02)	0.17
	YAMNet	65.31 (2.89)	64.36 (3.22)	66.25 (2.87)	0.71 (0.02)	0.27 (0.02)	0.13
	OpenL3	64.21 (1.86)	62.47 (2.24)	65.94 (2.58)	0.71 (0.02)	0.27 (0.02)	0.13
	BYOL-A	66.04 (2.41)	63.10 (2.39)	68.97 (3.26)	0.72 (0.02)	0.26 (0.01)	0.14
MLP	Handcrafted	67.92 (1.57)	64.89 (2.78)	70.96 (1.00)	0.74 (0.02)	0.21 (0.01)	0.13
	eGeMAPS	66.46 (2.75)	61.33 (4.1)	71.59 (2.18)	0.73 (0.03)	0.21 (0.01)	0.12
	ComParE	63.52 (2.96)	59.66 (6.18)	67.40 (5.17)	0.68 (0.03)	0.23 (0.01)	0.09
	VGGish	67.46 (2.69)	63.74 (4.77)	71.17 (0.91)	0.74 (0.02)	0.21 (0.01)	0.14
	YAMNet	68.97 (1.70)	64.58 (4.61)	73.37 (2.14)	0.76 (0.01)	0.20 (0.01)	0.07
	OpenL3	63.58 (2.09)	59.23 (5.56)	67.93 (2.44)	0.69 (0.02)	0.23 (0.01)	0.06
	BYOL-A	68.45 (2.23)	63.21 (3.21)	73.70 (3.3)	0.75 (0.02)	0.20 (0.01)	0.13

Table 4 Socio-demographic and clinical data

		Total			Normal Respiratory Quality of life (VQ11<22)			Impaired Respiratory Quality of Life (VQ11≥22)			p value
Participants		1908			954 (50%)			954 (50%)			NA
Mean VQ11 score		21.6 (8.2)			15 (3)			28.3 (6.1)			<0.0001
Gender	F	M	O	F	M	O	F	M	O	1	
	1280 (67.1%)	608 (31.9%)	20 (1%)	640 (67.1%)	304 (31.9%)	10 (1%)	640 (67.1%)	304 (31.9%)	10 (1%)		
Age		42.4 (14.2)			42.4 (14.1)			42.5 (14.2)			0.948
BMI [kg/m²]	Underweight	66 (3.5%)			35 (3.7%)			31 (3.2%)			<0.0001
	Normal weight	792 (41.5%)			490 (51.3%)			302 (31.7%)			
	Overweight	466 (24.4%)			224 (23.5%)			242 (25.4%)			
	Obesity	601 (30.6%)			205 (21.5%)			379 (39.7%)			
Smoking status	Not at all	1533 (80.4%)			806 (84.5%)			727 (76.2%)			<0.0001
	Less than daily	98 (5.1%)			50 (5.2%)			48 (5%)			
	Daily	277 (14.5%)			98 (10.3%)			179 (18.8%)			
Day coughing	No	1181 (61.9%)			704 (73.8%)			477 (50%)			<0.0001
	Transient	597 (31.3%)			235 (24.6%)			362 (38%)			
	Frequent	130 (6.8%)			15 (1.6%)			115 (12%)			
Night coughing	No	1414 (74.1%)			802 (84%)			612 (64.2%)			<0.0001
	Transient	396 (20.8%)			137 (14.4%)			259 (27.1%)			
	Frequent	98 (5.1%)			15 (1.6%)			83 (8.7%)			
Chest pain	Yes	191 (10%)			43 (4.5%)			148 (15.5%)			<0.0001
Sore throat	Yes	190 (10%)			71 (7.4%)			119 (12.5%)			0.0002
Asthma	Yes	306 (16%)			118 (12.4%)			188 (19.7%)			<0.0001
COPD	Yes	73 (3.8%)			18 (1.9%)			55 (5.8%)			<0.0001

F - Female; M - Male; O - Other; NA - Not Applicable; BMI - Body Mass Index; p<0.05 is considered statistically significant.

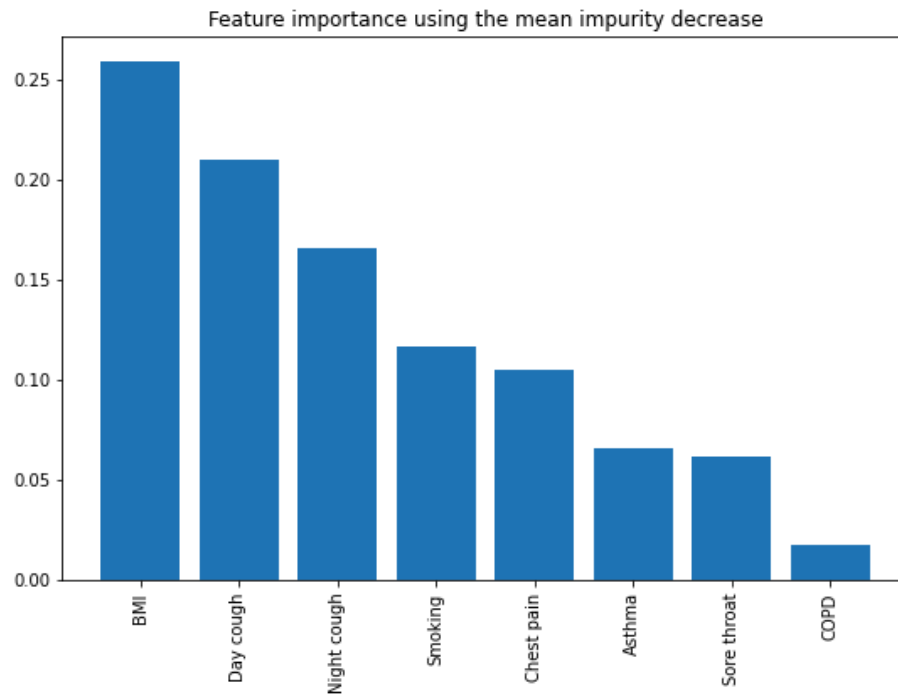


Figure 1 Feature importance for socio-demographic and clinical features based on the mean impurity decrease

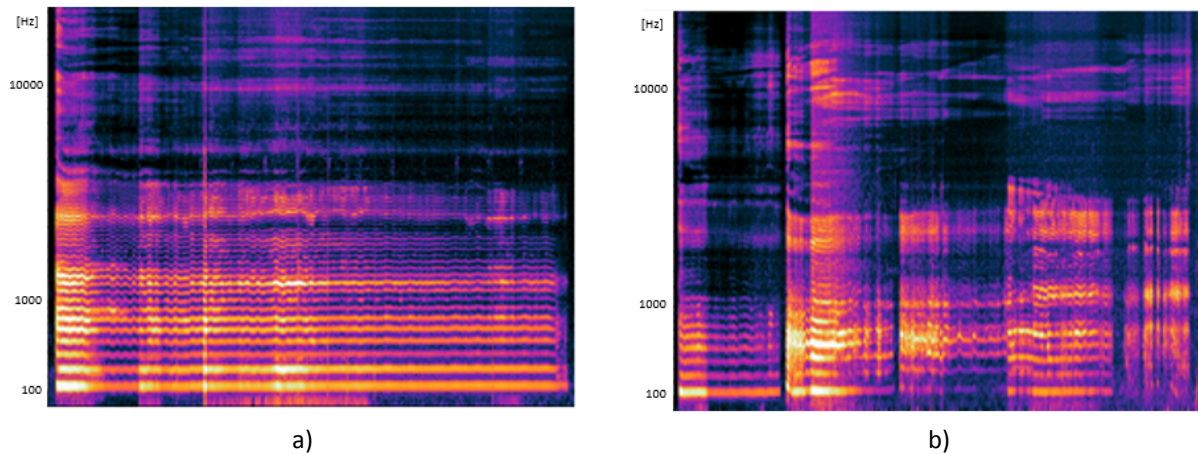


Figure 2 Spectrograms of sustained vowel phonation of participants matched by age and gender (male, age 67) with **a)** normal RQoL (VQ11 score: 21); and **b)** impaired RQoL (VQ11 score: 46). Normal RQoL spectrogram is represented by uninterrupted phonation, with clearly distinctive harmonics. Impaired RQoL spectrogram is characterized by strangled voice with multiple stoppages and voice breaks, and increased energy areas in higher frequency bands. The absence of higher harmonics above 1kHz can be observed, and as phonation progresses, the adjacent lower harmonics become smeared and more difficult to distinguish.

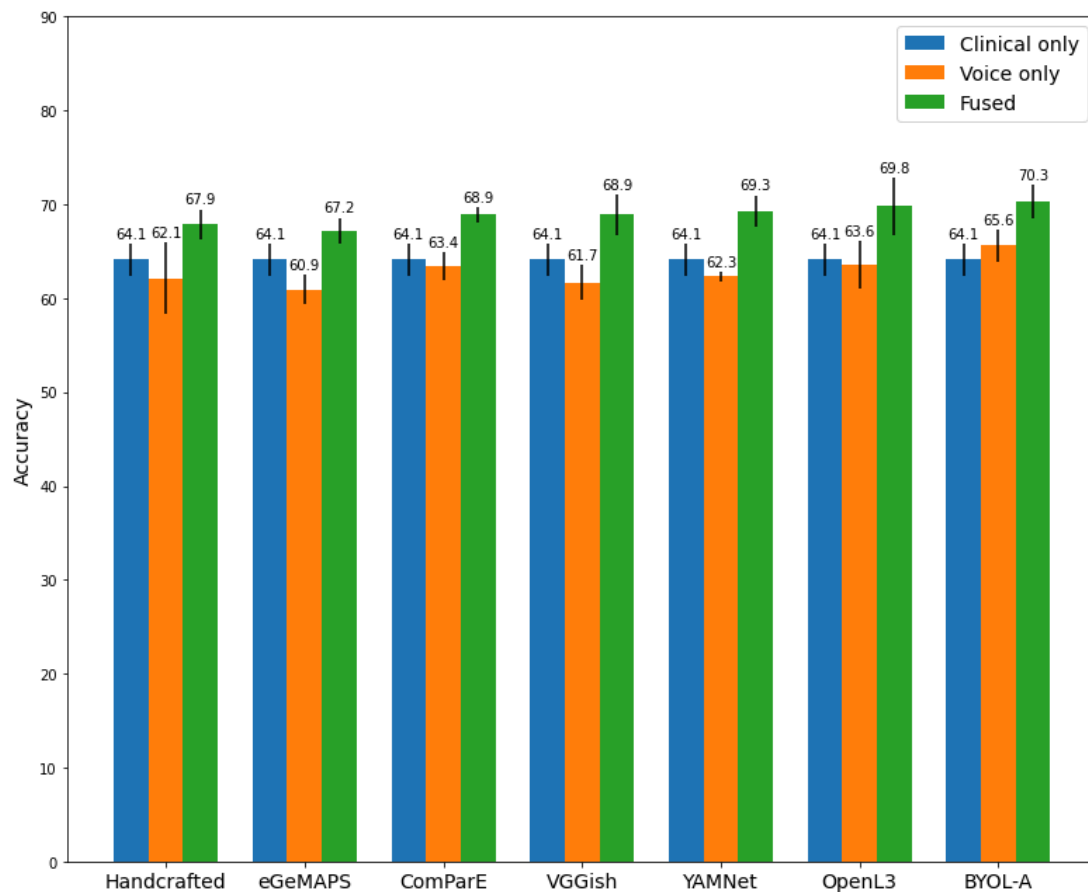
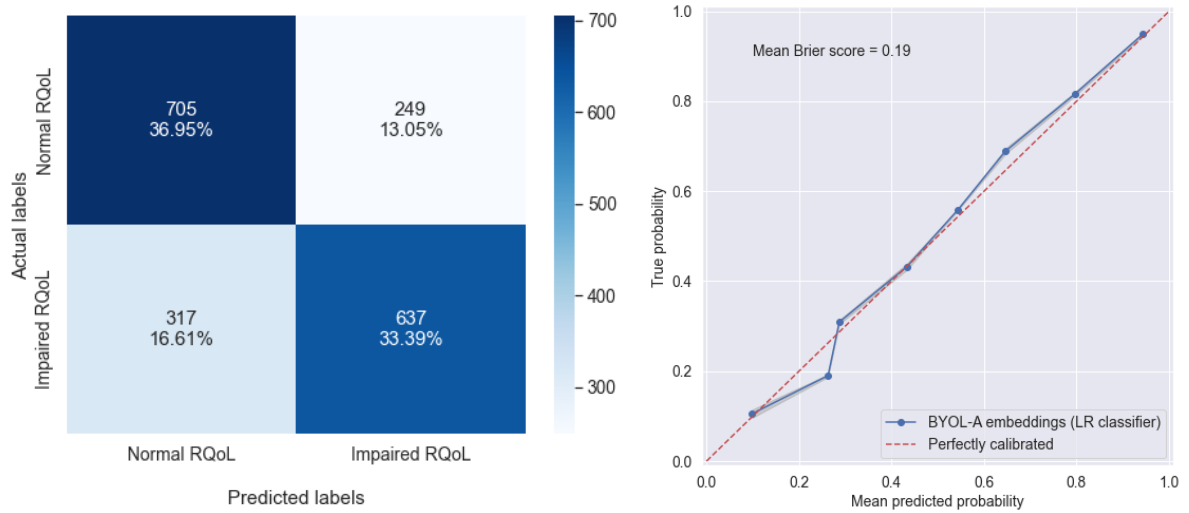
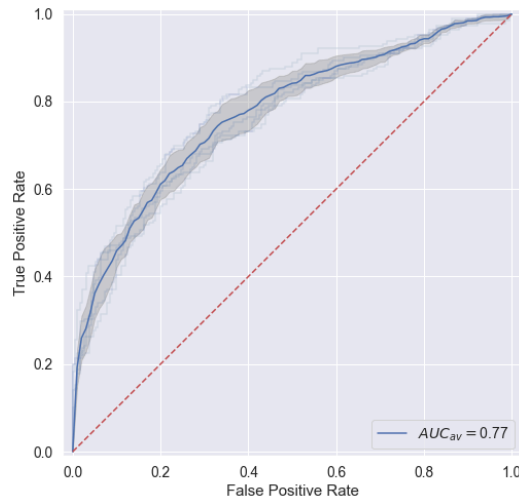


Figure 3 Accuracy with the best-performing machine learning model for socio-demographic/clinical features only, voice features only and fused clinical and voice (multimodal) features. Models with both clinical and voice data (“Fused”) systematically outperformed models where clinical variables only or voice features only were used. Error bars represent the standard deviation.



a)

b)



c)

Figure 4 a) Confusion matrix; b) Probability calibration curve; and c) ROC curve of the best-performing model (fused BYOL-A deep audio embeddings and socio-demographic/clinical features, trained with logistic regression classifier). Light blue lines denote the ROC curves across 5 cross-validation folds, whereas a thick blue line represents the average ROC curve. Standard deviation is highlighted with the shaded area.

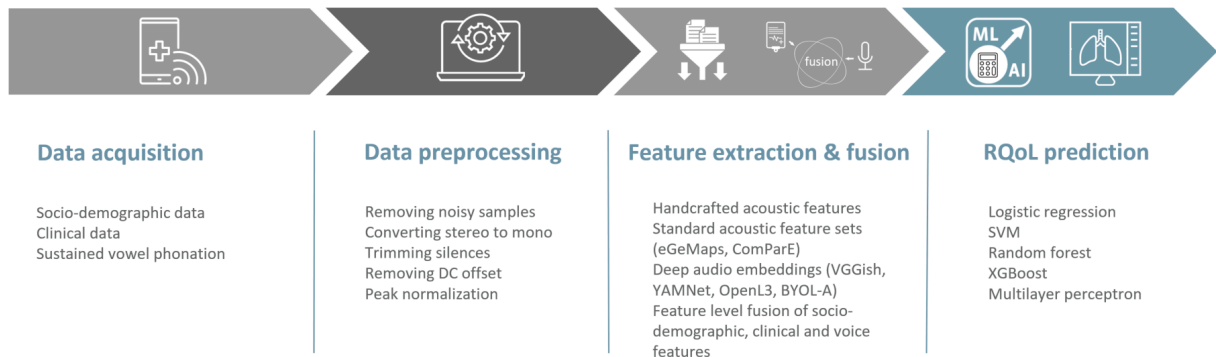


Figure 5 Workflow of RQoL monitoring

Supplementary information

Supplementary Table 1 Handcrafted audio features

ID	Feature	Domain	Computation parameters
1-26	MFCC	Cepstral	Mean and standard deviation of 13 MFCCs
27	RMS power	Time	None
28	Zero crossing rate	Time	None
29	Crest factor	Time	None
30	Dominant frequency	Spectral	None
31	Spectral centroid	Spectral	None
32	Spectral rolloff	Spectral	None
33	Spectral spread	Spectral	None
34	Spectral skewness	Spectral	None
35	Spectral kurtosis	Spectral	None
36	Spectral bandwidth	Spectral	None
37	Spectral flatness	Spectral	None
38	Spectral standard deviation	Spectral	None
39	Spectral slope	Spectral	None
40	Spectral decrease	Spectral	None
41	Maximum phonation time	Time	None
42-44	Aperiodicity features	Time	Fraction of locally unvoiced frames, Number of voice breaks, Degree of voice breaks
45-54	Tremor	Time	Frequency contour magnitude, amplitude contour magnitude, frequency tremor cyclicity, amplitude tremor cyclicity, frequency tremor frequency, amplitude tremor frequency, frequency tremor intensity index, amplitude tremor intensity index, frequency tremor power index, amplitude tremor power index
55-59	Jitter	Time	Local, local absolute, RAP, ppq5, ddp
60-65	Schimmer	Time	Local, local [dB], apq3, apq5, apq11, dda
66	Detrended fluctuation analysis	Nonlinear dynamics	None
67	Shannon entropy	Nonlinear dynamics	None
68	Harmonic to noise ratio	Time	None
69-70	Fundamental frequency (F0)	Prosodic	Mean, standard deviation
71-72	F0 contour	Prosodic	Mean, standard deviation

Supplementary Table 2 Characteristics of the deep audio embeddings

Audio embedding	Learning method	Dataset	Input	Embedding size
VGGish	Supervised	Audio Set	64 band log-mel spectrograms	128
YAMNet	Supervised	Audio Set	64 band log-mel spectrograms	1024
OpenL3	Self-supervised	Music/environmental subset of the Audio Set	128/256 band mel-spectrograms	512/6144
BYOL-A	Self-supervised	Audio Set	64 band log-mel spectrograms	512/1024/2048