UNIVERSITÀ
di VERONA
Scuola di dottorato
in SCIENZE UMANISTICHE

UNIVERSITÀ DEGLI STUDI DI VERONA
Dept. Of Human Sciences
PhD in Human Sciences XXXIIIV Cycle

PHD THESIS

Exploring the Role of Emotion Recognition Algorithms in Relation to Burnout and in Supporting Digital Care and Patient Empowerment

Course coordinator: Prof. Lorenzo Bernini
Tutor: Prof. Riccardo Sartori
PhD student: Andrea Buccoliero

La borsa di dottorato è stata cofinanziata con le risorse del PNRR:

- per il DM 351 nell'ambito della Missione 4 ("Istruzione e ricerca") – Componente 1 ("Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università"), Investimento 3.4. ("Didattica e competenze universitarie avanzate") e Investimento 4.1 ("Estensione del numero di dottorati di ricerca e dottorati innovativi per la pubblica amministrazione e il patrimonio culturale") - progetto M4C1 –Inv. 3.4 e progetto M4C1 – Inv. 4.1
- per il DM 352, nell'ambito della Missione 4 ("Istruzione e Ricerca") – Componente 2 ("Dalla Ricerca all'Impresa"), Investimento 3.3 ("Introduzione di dottorati innovativi che rispondono ai fabbisogni di innovazione delle imprese e promuovono l'assunzione dei ricercatori da parte delle imprese") – progetto M4C2 Investimento 3.3

Abstract

This doctoral thesis examines the role of emotion recognition algorithms, with a specific focus on Speech Emotion Recognition (SER), in relation to burnout syndrome, treated here as a focal but bounded construct, and to the broader objectives of digital care and patient empowerment. Burnout, officially recognised by the World Health Organization (WHO) as an occupational phenomenon, is increasingly prevalent across diverse professional sectors and is associated with severe psychological, organisational, and societal costs. Despite its impact, diagnostic practices remain largely dependent on self-report measures, which, while valuable, are limited by subjectivity, cultural biases, and their inability to capture dynamic changes in real-world contexts.

The research adopts a multidisciplinary perspective that integrates psychology, occupational health, and artificial intelligence. It is guided by the premise that chronic stress and burnout manifest in subtle but measurable changes in vocal production, reflected in acoustic, prosodic, and temporal features. By developing a digital application for collecting structured voice samples and psychometric data, this work investigates the potential of voice biomarkers to complement established diagnostic instruments, such as the Maslach Burnout Inventory (MBI) and the Oldenburg Burnout Inventory (OLBI).

At the core of the thesis is a conceptual model that situates SER within a layered framework, encompassing data collection, feature extraction, classification through machine learning, validation against psychometric benchmarks, and application within digital health platforms. This structure ensures that computational methods are not only technically accurate but also psychologically interpretable and clinically relevant. Empirical analyses demonstrate that AI-driven voice analysis can enhance the validity and sensitivity of burnout assessment, providing non-invasive, scalable, and continuous monitoring capabilities that traditional tools alone cannot achieve.

The study also addresses the ethical and practical dimensions of implementing SER in healthcare and occupational settings. Considerations of privacy, transparency, algorithmic bias, and user trust are positioned as fundamental prerequisites for adoption, highlighting that technological progress must be accompanied by ethical safeguards and regulatory compliance. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

By situating emotion recognition within the broader discourse of digital health, this research suggests that SER may enrich psychological assessment, inform occupational well-being strategies, and contribute to patient empowerment in carefully bounded ways. Ultimately, the thesis suggests that emotion recognition algorithms may serve not only as scientific tools for advancing affective computing but also as potentially useful instruments for informing healthcare delivery, supporting earlier and more personalised forms of care in appropriate contexts where appropriately validated and implemented.

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Prof. Riccardo Sartori, for his invaluable guidance, critical insights, and unwavering encouragement throughout my doctoral journey. His expertise has been a decisive factor in shaping the quality and direction of this work. My sincere thanks also go to Prof. Andrea Ceschi, whose thoughtful advice and continuous support provided both inspiration and motivation during key phases of my research.

I am grateful to my colleagues and collaborators at the University of Verona and GPI S.p.A., whose interdisciplinary perspectives and shared commitment to innovation enriched this project in many ways. I also thank my fellow PhD colleagues, who created a stimulating and supportive environment that made these years more rewarding both personally and professionally.

This work has been carried out within the framework of the National Recovery and Resilience Plan (PNRR) and was supported by specific funding allocated to doctoral research programs.

I would also like to acknowledge the indirect contribution of the TALIA[1] project (inTelligent vocAL biomarkers for pathology early detectIon and follow-up Assessment), which provided essential context for the development of the research.

Finally, I would like to thank all the participants and professionals who generously contributed their time and expertise to the empirical studies. Without their collaboration, the translation of theoretical constructs into practical applications would not have been possible.

[1] https://www.gpigroup.com/news/ai-e-voce-2-milioni-per-il-progetto-coordinato-da-gpi/

I also thank my wife, Francesca, whose unwavering support, patience, and understanding have been an invaluable source of strength throughout this demanding doctoral journey. Her belief in me and her continuous encouragement made every challenge surmountable.

Table of Contents

List of Figures

List of Tables

Abbreviations and Glossary

ACM	Association for Computing Machinery
AI	Artificial Intelligence
ANS	Autonomic Nervous System
ASR	Automatic Speech Recognition
CAM / Grad-CAM	Class Activation Map / Gradient-weighted Class Activation Mapping
CNN	Convolutional Neural Network
CPM	Component Process Model (Scherer's model of emotion)
CREMA-D	Crowd-sourced Emotional Multimodal Actors Dataset
DL	Deep Learning
EMOVO	Italian Emotional Speech Database
GDPR	General Data Protection Regulation
GPI	GPI S.p.A. (azienda)

HNR	Harmonic-to-Noise Ratio
HPA	Hypothalamic–Pituitary–Adrenal axis
IEEE	Institute of Electrical and Electronics Engineers
KEDS	Karolinska Exhaustion Disorder Scale
LSTM	Long Short-Term Memory (neural network architecture)
MBI	Maslach Burnout Inventory
MCC	Matthews Correlation Coefficient
MFCC	Mel-Frequency Cepstral Coefficients
ML	Machine Learning
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
RIR	Room Impulse Response
SER	Speech Emotion Recognition
SHAP	SHapley Additive exPlanations
SMBQ	Shirom–Melamed Burnout Questionnaire
TAFFC	IEEE Transactions on Affective Computing
VTLP	Vocal Tract Length Perturbation
WHO	World Health Organization
XAI	Explainable Artificial Intelligence
XLSR	Cross-Lingual Speech Representations

1. Introduction

Before delving into the substantive sections of this *introduction*, it is essential to provide a comprehensive overview of its structure and how each part incrementally builds the research argument, establishing a logical flow from broad contextualization to specific methodological objectives.

The foundational section is titled "*Context and Relevance of the Problem*". It initiates the discussion by firmly situating the study within the broader landscape of public health and occupational well-being. It underscores the pervasive and escalating challenge of burnout, a syndrome recognised globally for its profound impact on individuals, organisations, and healthcare systems alike. The section emphasises the global scale, complexity, and urgency of addressing burnout, highlighting its multifaceted origins that stem from demanding work environments, increased administrative burdens, and evolving societal expectations. By establishing the critical need for effective interventions, this section sets the stage for the subsequent exploration of the phenomenon.

Section 2, "Social, Technological, and Organisational Changes", delves into the transformative socio-technological shifts currently reshaping the healthcare landscape. It illustrates how rapid advancements in digital health technologies, including telemedicine, wearable sensors, and artificial intelligence, coupled with dynamic organisational restructuring and evolving societal demands for personalised care, have created fertile ground for the strategic application of Speech Emotion Recognition (SER). This section elucidates how SER, with its potential to assess emotional states from vocal cues objectively, presents a promising avenue for early detection and intervention in conditions such as burnout. Crucially, it also supports the necessity of anchoring SER development in robust psychological theories, ensuring that technological innovation is guided by a deep understanding of human emotion and cognition, thereby creating a conceptual bridge to the subsequent theoretical exploration.

The subsequent Section 3, titled "*Psychological Theories of Emotion*", serves as a theoretical cornerstone and provides a comprehensive overview of the various theoretical frameworks that have been developed to explain the nature, origins, and functions of emotions. This section will explore the historical progression of these theories, from early philosophical perspectives to more contemporary cognitive and neuroscience-based models. It will examine the core tenets of each theory, highlighting their strengths, limitations, and the empirical evidence that supports or refutes them. Key theories to be discussed

include, but are not limited to, the James-Lange, Cannon-Bard, Schachter-Singer, and appraisal theories (e.g., Lazarus, Ortony, Clore, & Collins), as well as discrete emotion theories and dimensional theories. Furthermore, the section will explore how these diverse perspectives contribute to our understanding of the complex interplay between physiological arousal, cognitive interpretation, behavioural responses, and subjective experience in the generation and regulation of emotions.

Moving from theoretical foundations to practical application, Section 4 focuses on "*Vocal Biomarkers and the 'Tone of Voice'*" as measurable and clinically relevant indicators of emotional states. It defines what constitutes a vocal biomarker in the context of psychological well-being, discussing specific acoustic features (e.g., pitch, loudness, speaking rate, voice quality) that have been empirically linked to various emotional states and psychological conditions, including stress, anxiety, and depression. This section elaborates on the physiological mechanisms that lead to these vocal changes, thereby demonstrating the scientific basis for using voice analysis as a non-invasive tool for emotional assessment and the early detection of burnout-related vocal signatures.

Section 5, "*Speech Emotion Recognition: State of the Art and Challenges*", provides a critical and comprehensive review of the state-of-the-art in SER technology. It highlights current achievements in the field, showcasing advancements in machine learning algorithms, deep learning architectures, and the development of sophisticated acoustic models that have significantly improved the accuracy and robustness of SER systems. Concurrently, this section candidly addresses persistent methodological challenges, including issues related to data scarcity, annotation inconsistencies, cross-cultural variability in emotional expression, privacy concerns, and the need for greater ecological validity in real-world applications. This balanced perspective underscores both the promise and the ongoing hurdles in advancing SER technology for clinical and occupational health contexts.

Following the theoretical and technological discussions, Section 6 provides a comprehensive understanding of burnout syndrome in depth, addressing this critical occupational health issue in detail. Titled "Burnout Syndrome: Clinical Definition and Measurement", it begins by describing its clinical definition, drawing from supported diagnostic criteria and theoretical models that conceptualise burnout as a multidimensional construct characterised by emotional exhaustion, depersonalization (or cynicism), and a reduced sense of personal accomplishment. This section examines prominent theoretical models, such as Maslach's Burnout Inventory, and discusses various measurement tools used in research and

clinical practice to assess the severity and prevalence of burnout. This detailed examination provides the necessary clinical context for understanding how SER technology can contribute to the identification and management of this condition.

Finally, Section 7 presents the specific "*Research hypotheses and objectives*" that guide the empirical investigation. This section clearly articulates the central questions the study aims to answer, delineating the expected relationships between variables and the specific outcomes anticipated from the research. The hypotheses are grounded in the theoretical frameworks and empirical evidence discussed in the preceding sections, ensuring a clear and direct link between the problem identified, the proposed solution, and the research questions. This section leads seamlessly into the methodology chapter, outlining the practical steps and approaches to be employed in testing the stated hypotheses and achieving the research objectives.

1.1 Context and Relevance of the Problem

As stated by the WHO in the International Classification of Diseases, over the past decade, the escalating prevalence and profound societal impact of emotional and behavioural disorders have become a critical concern, particularly for individuals operating within high-stress occupational environments, healthcare systems, and academic institutions (WHO, 2019). Global epidemiological data paint a concerning picture: longitudinal observations reveal a consistent and alarming upward trajectory in stress-related disorders, including burnout, across multiple professional sectors.

Burnout, officially recognised by the World Health Organisation (WHO, 2019) as an occupational phenomenon resulting from poorly managed chronic workplace stress, extends far beyond simple fatigue. It is a complex and debilitating syndrome characterised by three core, interrelated dimensions: emotional exhaustion, depersonalization (or cynicism), and a reduced sense of personal accomplishment (Grossi et al., 2015).

Emotional exhaustion refers to feelings of being emotionally overextended and depleted of one's emotional resources, often manifesting as a profound sense of tiredness that sleep does not alleviate. Depersonalization involves a cynical or detached response to one's job, colleagues, and clients, sometimes accompanied by irritability or a lack of empathy. Finally, a reduced sense of personal

accomplishment is characterised by feelings of incompetence and a lack of achievement and productivity at work.

As Grossi et al. (2015) emphasise, these three dimensions are cumulatively debilitating, often culminating in a "stress-related exhaustion disorder." This disorder not only impacts psychological well-being but also has significant physiological consequences, including increased susceptibility to illness, chronic pain, and sleep disturbances. While its onset is often insidious, gradually developing over time, burnout progressively erodes an individual's overall well-being. The consequences are far-reaching, leading to marked declines in professional performance and productivity, strained interpersonal relationships both within and outside the workplace, and a general deterioration in overall mental and physical health (Khammissa, Fourie, & Lemmer, 2022). Unaddressed, burnout can contribute to higher rates of absenteeism, presenteeism (being physically present but unproductive), and ultimately, turnover within organisations.

Empirical evidence, including insights from the "Preliminary Qualitative Study on AI and Burnout: Diagnostic Potential of Vocal Biomarkers", consistently highlights alarmingly high prevalence rates among high-demand professional groups. Healthcare providers, educators, and emergency personnel, who are frequently exposed to emotionally taxing situations and heavy workloads, often report burnout rates ranging from 20% to 50% (Dyrbye et al., 2017). Such statistics, consistently observed across diverse socio-cultural contexts and geographical regions, unequivocally underscore not only the immense scale of the problem but also the urgent and critical necessity for a multi-pronged approach encompassing early detection mechanisms, targeted and evidence-based interventions, and robust, sustained preventive measures to mitigate its detrimental effects on individuals and society at large. The economic burden associated with lost productivity, increased healthcare utilisation, and disability claims further amplifies the imperative for effective strategies to combat this pervasive occupational hazard.

Within this pressing context, voice-based emotional assessment emerges as a potential compelling avenue for identifying early, often imperceptible markers of psychological strain. A growing body of evidence supports the capacity of speech analysis to detect subtle alterations in affect and cognition associated with stress and mood disorders (Faurholt-Jepsen et al., 2021). This approach offers a non-invasive and continuous method for monitoring mental well-being, potentially enabling interventions before conditions escalate.

Advances in automated feature extraction tools have significantly advanced this field. Tools such as openSMILE, for instance, have enabled the precise capture of a wide array of acoustic parameters relevant to these early indicators (Eyben et al., 2010). These parameters can include pitch variability, speaking rate, voice quality, and energy fluctuations, all of which have been shown to reflect different aspects of emotional and cognitive states. The ability to automatically extract these features with high fidelity is crucial for the scalability and objectivity of voice-based assessment.

More recently, the integration of deep learning–based Speech Emotion Recognition (SER) frameworks have expanded the potential for scalable, real-time detection of such markers in naturalistic settings (Costantini et al., 2022). Deep learning models, with their ability to learn complex patterns from large datasets, can identify nuanced emotional cues that might be missed by traditional rule-based or machine learning approaches. This enables the development of robust and adaptable systems capable of analysing speech in diverse environments, ranging from telehealth consultations to ambient monitoring in daily life. The real-time capability of these frameworks further enhances their utility, offering immediate insights into emotional states and facilitating timely responses to emerging psychological distress (Lentini et al, 2025).

Human speech operates as a complex, multimodal signal in which communicative meaning emerges from the interaction of both lexical-semantic and non-lexical elements. As observed by Madanian et al. (2023), speech simultaneously conveys the propositional content of language and an affective layer, with the latter expressed through prosodic and paralinguistic cues that can serve as sensitive indicators of psychological states. Objective acoustic parameters—such as fundamental frequency (pitch), intensity, speech rate, and spectral characteristics—have been shown to vary systematically with cognitive load and affective condition (Costantini et al., 2022). More advanced descriptors, including Mel-Frequency Cepstral Coefficients (MFCCs), jitter, and shimmer, further capture fine-grained modulations in voice production, reflecting underlying neuromotor and emotional processes (Taguchi et al., 2018). Empirical evidence suggests that these vocal correlates are not merely incidental by-products of speech but are tightly coupled with mental health status. For instance, Low et al. (2020) documented that reduced pitch variability, monotonous delivery, lower intensity, and slower speech patterns are frequently associated with depressive symptomatology. Collectively, these findings underscore the potential of paralinguistic and prosodic markers—spanning from macro-level features like pitch and intensity to micro-level

perturbations such as jitter and shimmer—as robust, non-invasive indices for detecting subtle variations in cognitive and affective states, often before they manifest in overt behavioural or self-reported symptoms.

These acoustic markers, as synthesised in *A Scoping Review on the Use of Voice Biomarkers for Emotional Assessment* (Buccoliero et al., 2024), provide an invaluable, non-invasive diagnostic channel capable of detecting subclinical changes in mental state well before overt behavioural manifestations emerge. This capability is of particular importance in occupational settings where stigma, lack of resources, or organisational culture can hinder timely help-seeking.

The feasibility of such assessments has been dramatically enhanced by recent developments in Artificial Intelligence (AI) and Machine Learning (ML), which have advanced the ability to model, analyse, and interpret complex vocal signals. Studies, including those by Bakhshi et al. (2020), demonstrate how sophisticated algorithms can capture subtle acoustic patterns associated with stress and burnout, even in noisy, ecologically valid environments. By integrating explainability features, these models not only improve classification accuracy but also enhance interpretability, a critical factor for clinical and organisational adoption.

Concurrent socio-economic, demographic, and organisational transformations further increase the relevance of these technological capabilities. The accelerated digitalisation of work processes, the ageing of the global population, and the widespread adoption of remote and hybrid work arrangements have fundamentally reshaped psychosocial risk profiles. As argued in "What Makes Patients Engaged: An Umbrella Review and Multilevel Perspective on Patient Engagement" (Bassi et al., 2025), these structural and cultural shifts necessitate scalable, evidence-based, and theory-informed tools—such as Speech Emotion Recognition (SER)—that can operate across diverse contexts. Such systems have the potential to bolster individual resilience, inform proactive occupational health policies, facilitate continuous monitoring, and deliver timely, personalised interventions that can mitigate the onset or progression of burnout.

The intersection of rising burnout prevalence across different populations and professional groups, the advancement of voice-based biomarker research for stress and mental health assessment (Danhof-Pont, van Veen, & Zitman, 2011; Sara et al., 2023), and the rapid evolution of AI-driven analytical methods

(Grządzielewska, 2021; Wang, Boumadane, & Heba, 2021) creates a timely and necessary opportunity to develop integrated digital solutions. Such systems have the potential to improve early detection, strengthen patient and worker empowerment, and ultimately contribute to more resilient healthcare and organisational infrastructures.

While the transformative potential of digital technologies in healthcare and organisational contexts is widely acknowledged, a more critical perspective is necessary to fully understand their implications. In particular, the narrative of digital empowerment—often associated with increased accessibility, personalization, and patient engagement—risks overlooking structural inequalities and unintended consequences that may, paradoxically, exacerbate existing vulnerabilities. One of the most pressing concerns is that of digital exclusion, which refers not only to the lack of access to digital devices or internet connectivity, but also to broader socio-economic and demographic disparities that influence individuals' ability to benefit from digital innovations. As highlighted by Wilson-Menzfeld et al. (2025), digital exclusion is a multifaceted phenomenon shaped by age, income, education, and health status, and may lead to the systematic marginalisation of already disadvantaged populations. In this sense, digital health solutions—if not carefully designed and implemented—risk reinforcing, rather than reducing, inequalities in access to care.

1.2 Social, Technological, and Organisational Changes

Contemporary societies are currently undergoing profound and multifaceted transformations, as underscored by MacLachlan et al. (2019), who identify "changes in social structures and systems" that necessitate a critical re-evaluation and active reshaping to foster a more inclusive and sustainable future. These shifts are not isolated but somewhat interlinked and multidimensional, impacting various aspects of human endeavour and societal organisation.

A significant area of transformation is observed in the nature and structure of work itself, where traditional paradigms are evolving under the influence of technological change, globalisation, and shifting workforce demographics. Greenhalgh et al. (2005) highlight how the diffusion of innovation within organisations is reshaping supported practices, while Graffigna and Barello (2018) emphasise that these systemic shifts necessitate new approaches to skill development and active participation to ensure equitable opportunities and sustainable prosperity. This aligns with broader perspectives on

organisational adaptation and engagement, which stress the importance of data-driven and preventive strategies in navigating socio-economic transformations (Greene & Hibbard, 2012).

Simultaneously, the **philosophy and delivery of healthcare** are undergoing a substantial paradigm shift. There is an increasing emphasis on patient-centred and integrative frameworks, as highlighted by Gartner et al. (2022). This move signifies a departure from purely disease-focused models towards a more holistic understanding of health that considers the individual's unique needs, preferences, and social determinants. This integrative approach seeks to blend various medical disciplines and therapeutic modalities to offer comprehensive and personalised care, ultimately aiming for improved health outcomes and patient satisfaction.

Beyond the realms of work and healthcare, broader **societal determinants of health and well-being** are gaining increasing recognition. As MacLachlan and McVeigh (2021) emphasise, these determinants are inherently linked to achieving the Sustainable Development Goals. This perspective acknowledges that factors such as education, income, housing, environmental quality, and social support systems have a profound influence on an individual's health status and overall quality of life. Addressing these systemic factors is crucial for building resilient and healthy communities.

Demographic change, particularly the "ageing of the population," stands out as a pivotal trend shaping societal priorities and the allocation of resources, as underscored by Barnes (2005). This progressive demographic shift is intricately linked to a global surge in chronic diseases and neurodegenerative conditions (Wallace et al., 2022; Watson & Leverenz, 2010), alongside a notable increase in stress-related disorders. Concurrently, labour markets are undergoing profound structural changes. Transformations in work organisation increasingly necessitate cognitively demanding, knowledge-intensive, and frequently emotionally taxing activities (Golubinski, Oppel, & Schreyögg, 2020). This occurs within an accelerating pace of work, which critically challenges both individual resilience and the robustness of organisational systems. Taken together, these interconnected dynamics not only intensify existing psychosocial risk factors but also significantly broaden the scope and complexity of vulnerabilities to emotional strain, professional burnout, and various other mental health challenges. The convergence of an ageing population, rising health burdens, and evolving work demands presents a multifaceted challenge for public health, economic productivity, and social welfare systems, necessitating comprehensive strategies that address both individual well-being and systemic resilience.

Furthermore, these intricate transformations are significantly influenced by complex processes of innovation diffusion within health systems, as extensively detailed by Greenhalgh et al. (2005). The adoption and spread of new technologies, practices, and organisational models within healthcare are not linear but rather involve a dynamic interplay of social, economic, and political factors. This underscores the profound interrelationship between socio-economic change, ongoing healthcare reform efforts, and the overarching pursuit of collective well-being. Understanding and managing these diffusion processes are vital for effectively translating scientific advancements into tangible improvements in public health and societal flourishing.

Technological innovation, particularly within the burgeoning domain of **digital health**, has emerged as a dual force: both a primary driver of systemic transformation in healthcare and a powerful counterbalance to its inherent risks. Digital health interventions are remarkably diverse, encompassing a broad spectrum of technologies designed to enhance health outcomes and streamline the delivery of care.

At one end of this spectrum are sophisticated "wearable sensing devices for detecting and processing acoustic signals in healthcare" (Mallegni et al., 2022). These devices, often seamlessly integrated into daily life, can continuously monitor vital acoustic information, offering early detection capabilities for various conditions and providing valuable data for diagnostic purposes. Beyond acoustic analysis, wearable technology extends to monitoring other physiological parameters, enabling proactive health management and personalised feedback for users.

Complementing these sensing devices are mobile health (mHealth) applications, which have rapidly become ubiquitous tools for supporting patient engagement and self-monitoring (Cozad et al., 2022). These applications empower individuals to take a more active role in their health journey by providing interfaces for tracking symptoms, medication adherence, diet, and exercise. They can also deliver personalised health information, educational content, and reminders, fostering greater adherence to treatment plans and promoting healthier lifestyles.

Another pivotal component of digital health is telemedicine platforms. As meticulously described by Chiang, Starren, and Demiris (2021), these platforms leverage communication technologies to enable remote consultations, diagnoses, and monitoring. Telemedicine has proven instrumental in extending

specialist care beyond traditional geographical and logistical barriers, making healthcare more accessible to underserved populations and improving convenience for all patients. It facilitates virtual visits with physicians, specialists, and therapists, reducing the need for in-person appointments and mitigating associated travel and wait times.

Furthermore, significant **advances in AI-assisted biosensing** are driving "the convergence of traditional and digital biomarkers" (Arya et al., 2023). This convergence represents a paradigm shift in medical diagnostics and personalised care. By integrating multimodal physiological, cognitive, and affective data—collected from diverse sources including traditional lab tests, imaging, and advanced digital sensors—AI algorithms can identify subtle patterns and correlations. This comprehensive data integration facilitates earlier and more accurate disease detection, predicts disease progression, and enables the development of highly personalised treatment plans tailored to the unique biological and behavioural profiles of individual patients. This synergistic approach promises to revolutionise preventive medicine, chronic disease management, and the delivery of precision healthcare.

"The Effectiveness of a Home-Based Computerised Cognitive Training in Parkinson's Disease: A Pilot Randomised Cross-Over Study" (Tagliente et al., 2024), to which I directly contributed, offers compelling empirical evidence of these benefits, showing that structured, home-based cognitive training protocols can yield measurable improvements in attention, executive function, and mood among individuals living with neurodegenerative disorders. These outcomes underscore the capacity of technology to extend high-quality, personalised interventions beyond traditional clinical environments, thereby enhancing accessibility, adherence, and continuity of care.

From an organisational perspective, these technological advancements are time by time integrated into **comprehensive workplace health strategies**. These strategies leverage digital monitoring tools within robust occupational health frameworks, aligning with the core tenets of Population Health Management. This approach is characterised by its "proactive, preventive, and data-driven approaches to health and well-being" (MacLachlan et al., 2019). The integration of such systems facilitates continuous, non-invasive assessment of critical mental and emotional parameters. This capability is crucial for detecting early signs of stress accumulation and accurately predicting potential burnout trajectories. Consequently, organisations can implement timely and targeted interventions, effectively mitigating long-term adverse effects on employee well-being and productivity (Graffigna & Barello, 2018; Greene & Hibbard, 2012).

This proactive stance not only supports individual employee health but also contributes to a more resilient and sustainable organisational environment.

Thus, the contemporary landscape is undergoing a profound transformation, driven by the convergence of societal shifts, technological advancements, and organisational paradigm changes. This dynamic interplay is not merely incremental but rather a fundamental reshaping of "social structures and systems" (MacLachlan et al., 2019), aimed at fostering greater participation and inclusion across various domains. Within the healthcare sector, this evolution is particularly salient, manifesting in the widespread adoption of participatory models of care.

A cornerstone of these emerging models is the enhanced capacity for self-monitoring and the provision of immediate, highly personalised feedback. These capabilities are primarily powered by an array of sophisticated digital tools, including eHealth platforms (Barello et al., 2016) and patient portals (Irizarry, De Vito Dabbs, & Curran, 2015). These technologies empower individuals to become more proactive and engaged participants in their own health journeys. By providing accessible data and tailored insights, they enable individuals to contribute to the co-design and implementation of healthcare interventions actively.

The integration of self-monitoring capabilities represents a paradigm shift in healthcare delivery, moving away from a solely reactive approach to a more proactive and preventative one. Individuals are encouraged to be no longer passive recipients of care but active agents in managing their well-being. This aims at facilitating real-time data collection through wearable devices, mobile applications, and connected health sensors, which provide a continuous stream of information about vital signs, activity levels, sleep patterns, and other health indicators.

Coupled with self-monitoring is the crucial element of personalised feedback. This feedback transcends generic advice, leveraging algorithms and artificial intelligence to analyse individual data and deliver insights that are specific to the user's health profile, goals, and behaviours. For example, a patient with diabetes might receive personalised recommendations for diet and exercise based on their glucose readings, or an individual striving for weight loss might get tailored suggestions for calorie intake and activity targets. This immediate and relevant feedback serves as a motivator, reinforcing positive behaviours and guiding individuals towards healthier choices.

eHealth platforms, as highlighted by Barello et al. (2016), serve as comprehensive ecosystems for managing health information and facilitating communication between patients and healthcare providers. These platforms often incorporate features such as secure messaging, online appointment scheduling, prescription refills, and access to educational resources. They can also integrate with various self-monitoring devices, aggregating data into a centralised, user-friendly interface. This holistic approach streamlines administrative tasks and empowers patients with readily available information and tools to manage their health proactively.

Similarly, patient portals, as described by Irizarry, De Vito Dabbs, & Curran (2015), are secure online platforms that grant patients direct access to their electronic health records (EHRs). This access includes lab results, medication lists, immunisation records, and clinical notes. Beyond simply viewing information, many portals offer functionalities that allow patients to communicate with their care teams, request referrals, and even contribute to their health information. By demystifying medical jargon and providing transparency, patient portals foster a sense of ownership and accountability in individuals regarding their health data.

Ultimately, these digital tools are instrumental in fostering a collaborative approach to healthcare. By providing accessible data and tailored insights, eHealth platforms and patient portals empower individuals to contribute to the co-design and implementation of healthcare interventions actively. This means individuals can provide valuable input on treatment plans, express their preferences, and work in partnership with their healthcare providers to develop personalised strategies that align with their lifestyle and goals. This collaborative model not only enhances patient satisfaction but also yields more effective and sustainable health outcomes.

As articulated by Jørgensen et al. (2018), such participatory approaches are instrumental in cultivating **patient empowerment**. This signifies a pivotal shift away from the traditional model where individuals are passive recipients of medical care. Instead, patients are increasingly recognised and supported as active partners in managing and enhancing their own well-being. This paradigm shift not only promotes a more person-centred approach to healthcare but also holds the potential to improve health outcomes, foster greater patient satisfaction, and optimise resource utilisation within healthcare systems. The emphasis on individual agency and collaborative care is redefining the relationship between patients and

providers, paving the way for a more integrated, responsive, and effective healthcare ecosystem.

Digital ecosystems can foster sustained engagement by providing "persuasive features for patient engagement through mHealth applications in managing chronic conditions" (Almutairi, Vlahu-Gjorgievska, & Win, 2023), equipping individuals with the knowledge, skills, and motivational resources necessary to make informed health decisions. This participatory approach — supported by systematic frameworks for eHealth-driven patient engagement (Barello et al., 2016) and grounded in the principles of patient empowerment (Acuña Mora et al., 2022) — offers the dual advantage of improving clinical outcomes and reducing pressures on healthcare systems, as engaged individuals are more likely to adhere to preventive measures and self-care routines.

These transformations are not isolated events but rather emerge from a complex interplay of social structures, technological systems, and individual agency (MacLachlan et al., 2019). A continuous feedback loop characterises this dynamic relationship: evolving social needs, such as the demand for more accessible or personalised healthcare, act as a primary driver for technological innovation. In response, new technological capacities, such as artificial intelligence and advanced data analytics, subsequently reshape organisational strategies across various sectors (Greenhalgh et al., 2005).

These organisational shifts, in turn, have profound implications for the lived experiences of individuals. In both work and healthcare contexts, these changes reinforce the increasing importance of concepts such as "patient empowerment" and active participation in care processes (Graffigna & Barello, 2018; Groene et al., 2010). Individuals are no longer passive recipients but are encouraged to be active stakeholders in managing their health and professional development.

Effectively addressing this multifaceted and multidirectional feedback loop necessitates highly interdisciplinary approaches. Bridging traditional silos between social science, clinical research, and data science is crucial to developing interventions that are not only technologically sophisticated and efficient but also deeply socially and culturally responsive (Gartner et al., 2022; Wind et al., 2022). This integrated perspective ensures that innovation serves human needs and societal well-being, rather than merely advancing technology for its own sake.

The interlocking forces of demographic evolution, digital innovation, and adaptive organisational

change create a dual imperative. On one hand, these changes elevate the risk environment for burnout and other mental health issues, demanding vigilant monitoring and responsive care systems. On the other hand, they present an unprecedented opportunity to deploy technology-driven, evidence-based, and contextually nuanced solutions capable of fostering resilience, optimising the allocation of resources, and strengthening population health outcomes at scale. Recognising and strategically leveraging this balance will be central to the next generation of health policy, workplace well-being programs, and integrated care models. In the context of this thesis, these macro-level transformations also underscore the necessity of grounding subsequent discussions—particularly on psychological theories of emotion—in an understanding of how socio-technological change shapes both the opportunities and design imperatives for Speech Emotion Recognition systems.

Closely related to this issue is the challenge of digital health literacy, which concerns individuals' capacity to understand, interpret, and effectively use digital health information and tools. The increasing complexity of digital platforms, combined with the proliferation of data-driven interfaces, may create barriers for users who lack the necessary cognitive, technical, or educational resources. This is particularly relevant in the context of emotionally and cognitively demanding conditions such as burnout, where individuals may already experience reduced attentional capacity and decision-making abilities. Furthermore, as Ziebland, Hyde, and Powell (2021) argue, digital health technologies often generate unintended consequences that complicate their intended benefits. Among these, information overload and alert fatigue represent critical risks: systems designed to monitor, notify, and support users can inadvertently produce continuous interruptions, excessive feedback, and heightened cognitive burden. Rather than alleviating stress, such mechanisms may contribute to emotional strain, reduced adherence, and even disengagement from care pathways.

Taken together, these considerations suggest that the integration of digital solutions into healthcare and organisational systems must be approached with a nuanced and reflexive perspective. Technological innovation should not be interpreted as inherently beneficial, but rather as contingent upon its alignment with users' capabilities, contexts, and needs. This implies the necessity of adopting inclusive design principles, promoting digital literacy, and carefully evaluating the cognitive and emotional load imposed by digital interventions. In doing so, it becomes possible to move beyond a purely optimistic view of digital transformation and towards a more balanced framework that acknowledges both its opportunities and its limitations.

1.3 Psychological Theories of Emotion

It is essential to establish the conceptual foundation that connects supported emotion theories to the practical objectives of this thesis, explicitly clarifying how each framework has been operationalised or empirically tested in SER and voice biomarker studies relevant to burnout detection. These models not only explain the mechanisms of emotional experience but also inform concrete design decisions in SER, such as which acoustic markers to prioritise and how to interpret them in burnout-related contexts. A rigorous exploration of emotion—particularly in the intertwined contexts of SER and voice biomarker research—demands engagement with supported psychological theories capable of explaining the physiological, cognitive, and behavioural substrates of affective phenomena.

Emotions are complex, multidimensional, and dynamic phenomena that emerge from a continuous interplay between various internal and external factors. This interplay encompasses neurophysiological activation, which involves specific brain regions and neurotransmitter systems; intricate cognitive appraisal processes, through which individuals evaluate the significance of events; and a range of expressive modalities, including readily observable cues like facial expressions, nuanced body language, and the intricate patterns of speech (Ekman, 1992; Pfister & Robinson, 2010).

Among these expressive modalities, speech holds particular significance as it conveys not only explicit lexical content but also a rich emotional component that can subtly yet powerfully reveal the speaker's internal mental and affective states (Madanian et al., 2023). The ability of speech to encode and transmit emotional information forms the foundational theoretical basis for the development and application of Speech Emotion Recognition (SER) systems. These sophisticated computational systems are designed to analyse a multitude of acoustic, prosodic, and temporal features inherent in the human voice. Key features commonly analysed include pitch (the perceived highness or lowness of a sound), intensity (the loudness or energy of a sound), and Mel-Frequency Cepstral Coefficients (MFCCs), which represent the short-term power spectrum of a sound, particularly relevant for speech processing. By systematically analysing these features, SER systems aim to derive scientifically grounded and clinically meaningful interpretations of emotional states, offering objective insights into an individual's emotional landscape (Costantini et al., 2022; Taguchi et al., 2018). The utility of SER extends across various fields, including mental health assessment, human-computer interaction, customer service analytics, and security

applications.

The **James–Lange theory,** a foundational concept in the study of emotion, proposes that our subjective emotional experiences are a direct consequence of our perception of physiological changes within our bodies (James, 1890, as cited in Pfister & Robinson, 2010). This implies a bottom-up process where bodily sensations precede and give rise to emotional feelings. For instance, according to this theory, we don't cry because we are sad; instead, we feel sorry because we cry and perceive the associated physiological changes, such as increased heart rate, tear production, and muscle tension.

In the specialised field of Speech Emotion Recognition (SER), the James–Lange theory offers a compelling framework. It suggests that quantifiable vocal modulations—such as alterations in pitch, amplitude, speech rate, spectral energy distribution, and timbre—can serve as discernible indicators of underlying autonomic nervous system activation. This is because the autonomic nervous system directly influences the muscles and organs involved in speech production. When an individual experiences an emotional state, the autonomic nervous system responds with specific physiological changes, many of which can manifest in their voice.

Empirical research consistently supports this premise. For example, Costantini et al. (2022) conducted a study within occupational burnout settings and successfully identified stress-related shifts in both pitch and amplitude variations in speech. Furthermore, their study utilised sophisticated algorithmic analyses, explicitly focusing on acoustic features like Mel-Frequency Cepstral Coefficients (MFCCs). MFCCs are widely used in SER because they effectively represent the spectral envelope of a sound, which is crucial for distinguishing different vocal qualities. The consistent and reproducible patterns of autonomic arousal revealed through these analyses—such as elevated pitch variability or changes in speech rate—directly illustrate the physiological foundation proposed by the James–Lange theory. These findings suggest that the objective, measurable characteristics of a person's voice can provide tangible evidence of their internal emotional state, aligning with the theory's core assertion that physiological responses are primary to subjective emotional experience. The ability to detect and interpret these vocal markers holds significant implications for developing more advanced and accurate SER systems, which can have applications in areas like mental health monitoring, human-computer interaction, and even lie detection.

In contrast, the **Cannon–Bard theory** posits that “emotional experience and physiological arousal occur

simultaneously and independently" through the mediation of subcortical structures (Pfister & Robinson, 2010). Applied to Speech Emotion Recognition (SER), this framework supports the concurrent analysis of expressive vocal changes and autonomic markers without assuming a unidirectional causal relationship. Taguchi et al. (2018) provide an illustrative case: in the context of neonatal cry analysis, concurrent yet independent changes in vocal tone and physiological distress indicators were observed, reflecting the Cannon–Bard principle that affective and physiological components are co-activated but separable in origin. In the paper "*Analysing neonatal vocal expression: methodological approaches to identifying neurological and psychiatric signatures*" (Saha et al., 2025), authors (including myself) support the relationship. This study is reported in Chapter 7 of this thesis.

Appraisal theories, particularly those championed by Lazarus, posit cognitive evaluation as the quintessential antecedent to emotional experience (Pfister & Robinson, 2010). This theoretical framework suggests that emotions are not merely direct responses to stimuli but rather emerge from an individual's subjective interpretation and assessment of a situation. Within the realm of Speech Emotion Recognition (SER), this principle holds significant implications. Prosodic features (e.g., pitch, intensity, duration, speech rate) and spectral variations (e.g., formants, spectral centroid, bandwidth) in a speaker's voice can serve as powerful indicators of their appraisal of a situation. These acoustic cues, often operating beneath the level of conscious awareness, encode the speaker's subjective interpretations, providing insights that extend far beyond the literal, lexical content of their utterances. Research by Low et al. (2020) provides compelling empirical evidence in support of integrating appraisal theory into SER methodologies. Their work supported that combining situational metadata—information about the context, circumstances, and personal relevance of an event—with traditional acoustic features, including both prosodic and spectral parameters, yielded a substantial improvement in the algorithmic accuracy of emotion classification. This finding directly aligns with the core tenet of appraisal theory, which emphasises the primacy of subjective evaluation processes in shaping emotional responses. By understanding how individuals appraise their surroundings and how these appraisals are manifested acoustically, SER systems can achieve a more nuanced and accurate understanding of human emotion, moving beyond superficial classifications to capture the intricate interplay between cognition, context, and vocal expression. This approach underscores the need for SER models to consider not just what is said, but how it is said, and the underlying cognitive processes that drive those vocalisations.

From a dimensional perspective, emotion is often conceptualised as existing along continuous axes

rather than as discrete categories. One of the most influential frameworks in this regard is **Russell's Circumplex Model of Affect** (Russell, 1980). This model posits that emotions can be understood as points within a two-dimensional space defined by two orthogonal axes: "valence" and "arousal." Valence refers to the pleasantness or unpleasantness of an emotion, ranging from highly negative (e.g., sadness, anger) to highly positive (e.g., joy, excitement). Arousal, on the other hand, describes the intensity or activation level of an emotion, spanning from low arousal (e.g., calmness, boredom) to high arousal (e.g., excitement, fear). This circumplex structure suggests that emotions are not isolated experiences but somewhat interconnected and can be ordered around a circle. Building upon this foundational framework, recent research has sought to operationalise these abstract emotional dimensions using measurable indicators. Pérez-Toro et al. (2023) exemplify this approach by demonstrating how acoustic indicators can be mapped to affective coordinates within Russell's model. Specifically, their work explored the relationship between features of speech, such as spectral centroid (a measure related to the brightness or sharpness of a sound) and speech rate (the speed at which a person speaks), and the corresponding levels of valence and arousal. This innovative methodological approach allowed them to achieve cross-linguistic emotion recognition, meaning their system could identify emotions regardless of the language being spoken. By analysing universal acoustic patterns rather than relying on language-specific lexical or semantic cues, they moved towards a more generalizable understanding of emotional expression.

This focus on developing language-independent assessment tools has significant implications, particularly within fields such as burnout detection. The ability to accurately and consistently identify emotional states across diverse linguistic and cultural contexts is crucial for creating robust and widely applicable diagnostic and monitoring systems. Traditional self-report measures of burnout can be subjective and influenced by cultural norms or reporting biases. By leveraging objective acoustic indicators mapped onto a well-supported dimensional model of emotion, researchers can develop more reliable and generalizable methods for assessing emotional states linked to burnout. This alignment with the broader objective of developing universal assessment tools underscores the potential of such methodologies to advance our understanding and detection of complex psychological phenomena.

Scherer's Component Process Model (CPM) offers a comprehensive framework for understanding emotions as dynamic episodes arising from continuous appraisal processes. According to Scherer (2009), these emotional episodes are not static states but rather involve a coordinated interplay of various

components: physiological changes (e.g., heart rate, skin conductance), expressive behaviours (e.g., facial expressions, vocalisations), and subjective feelings (e.g., conscious experience of joy, sadness). The CPM posits that these components are intricately linked and unfold over time, influenced by an individual's ongoing evaluation of their environment.

This framework is particularly relevant when considering the impact of prolonged stress on emotional expression. For instance, Higuchi et al. (2022) conducted research that aligns with the CPM's predictions by observing systematic alterations in vocal parameters in individuals experiencing chronic stress, including burnout. Their findings supported changes in fundamental frequency (F0), which relates to pitch, as well as harmonic-to-noise ratio, indicating voice quality, and temporal speech patterns, such as speech rate and pauses. These vocal modifications serve as tangible evidence that emotional states, particularly those associated with chronic stress, manifest across multiple, interdependent channels, precisely as predicted by the CPM. The consistency between Higuchi et al.'s observations and Scherer's model highlights the utility of the CPM in providing a comprehensive understanding of how emotional experiences are embodied and expressed, particularly in challenging psychological states such as burnout.

Despite the rapid advancements and the proliferation of theoretical models in the field of Sentiment and Emotion Recognition (SER), a critical lacuna persists: "a significant gap in psychological understanding regarding how SER algorithms function and their effectiveness" (Costantini, Cesarini, & Casali, 2022). This fundamental disconnect highlights a pervasive challenge in bridging the seemingly disparate domains of computer science and psychology. To effectively bridge this divide, it is imperative to integrate psychological constructs directly into every stage of algorithm development. This includes, but is not limited to, informed feature selection that reflects psychological theories of emotion and sentiment, the construction of model architectures that are psychologically plausible, and the development of interpretability frameworks that resonate with human cognitive processes. This urgent need for a "multidisciplinary approach that can combine psychological and computer science knowledge" in algorithm design has been emphatically underscored.

Such a deep and systematic integration is not merely an academic exercise; it is anticipated to profoundly enhance diagnostic accuracy in real-world applications, particularly within mental health assessment and intervention. By embedding psychological principles, the outputs generated by SER algorithms are

expected to become more meaningful, offering insights that are clinically relevant and understandable to practitioners. Furthermore, this integration is crucial for ensuring the explainability of these complex models, a vital component for fostering trust and facilitating responsible use in sensitive areas, such as mental health. Ultimately, a psychologically informed approach is crucial for maintaining ethical soundness, mitigating biases, and ensuring that technological advancements promote human well-being.

Within the scope of this thesis, these operationalised principles serve as guiding tenets, shaping every decision made. This will dictate which acoustic features are prioritised for analysis, moving beyond mere technical availability to select features with demonstrable psychological relevance. It will also influence how classification models are structured, potentially leading to novel architectures that better capture the nuances of human emotion. Crucially, the application of interpretability methods, such as SHAP value analysis (Lundberg & Lee, 2017), will be informed by these principles, ensuring that the insights gleaned from these analyses are not only statistically robust but also clinically meaningful and actionable. This comprehensive approach aims to maintain the highest standards of scientific rigour while maximising the clinical relevance and ethical integrity of the research.

1.4 Vocal Biomarkers and the "Tone of Voice"

Vocal biomarkers are defined as acoustic characteristics of the voice which undergo modifications when the subject is under the influence of a workload (Ruiz, Legros, & Guell, 1990). These subtle yet significant alterations in voice patterns offer an objective and unobtrusive window into an individual's psychological and physiological condition. They represent a groundbreaking frontier in health monitoring, providing insights that traditional assessment methods might miss.

In contemporary healthcare, the domains of well-being and Human-Machine Interface, particularly in SER and burnout detection, utilise vocal biomarkers as a non-invasive and highly scalable tool for monitoring emotional well-being. Their utility is especially pronounced in emerging healthcare models such as telemedicine, where remote patient monitoring is crucial, as well as in occupational health settings for proactive employee well-being management (Shah et al., 2025; Danhof-Pont, van Veen, & Zitman, 2011; Langer et al., 2022). The ability to assess emotional and physiological states from voice recordings offers significant advantages in terms of accessibility, cost-effectiveness, and data collection efficiency.

The core acoustic-prosodic indicators that are typically analysed include:

- **Fundamental frequency (F0)**: Also known as pitch, F0 reflects the rate of vibration of the vocal folds and can vary significantly with emotional state.
- **Jitter**: This refers to the cycle-to-cycle variation in fundamental frequency, often indicating vocal instability.
- **Shimmer**: Similar to jitter, shimmer measures the cycle-to-cycle variation in amplitude of the vocal signal.
- **Harmonic-to-noise ratio (HNR)**: The HNR quantifies the ratio of periodic (harmonic) to aperiodic (noisy) components in the voice, providing insights into the regularity of vocal fold vibration.
- **Spectral centroid:** This parameter describes the "centre of mass" of the voice's spectrum, indicating the overall brightness or darkness of the sound.
- **Formant frequencies**: These are the resonant frequencies of the vocal tract, which are crucial for speech perception and can be affected by articulatory precision.
- **Speech rate:** The speed at which words are spoken can be a strong indicator of arousal or cognitive load.
- **Articulation precision** refers to the clarity and distinctness of speech sounds, which can be compromised under stress.
- **Pause distribution**: The frequency, duration, and placement of pauses in speech can reveal cognitive processing and emotional states.

These parameters collectively reflect the intricate interplay between affective states, respiratory control, laryngeal biomechanics, and articulatory processes (Costantini, Cesarini, & Casali, 2022). For instance, increased emotional arousal may lead to changes in respiratory patterns, which in turn affect vocal fold tension and alter the fundamental frequency (F0). Similarly, muscle tension induced by stress can impact articulatory precision and laryngeal control, manifesting as changes in jitter, shimmer, or HNR.

The underlying mechanisms mediating alterations in these vocal parameters are deeply rooted in neurophysiological responses to stress. Notably, dysregulation of the hypothalamic-pituitary-adrenal (HPA) axis, the body's central stress response system, plays a significant role. Chronic stress can lead to

sustained activation of the HPA axis, influencing various physiological systems that affect voice production. Furthermore, shifts in the autonomic nervous system (ANS) balance, particularly an increase in sympathetic nervous system activity and a decrease in parasympathetic activity, can directly impact vocal fold tension, respiratory control, and vocal tract muscle activity (Grossi et al., 2015; Khammissa, Fourie, & Lemmer, 2022). These neurophysiological changes ultimately translate into detectable modifications in the acoustic properties of the voice, making vocal biomarkers a powerful tool for monitoring and understanding human psychological and physiological states.

The concept of "tone of voice" extends far beyond mere variations in prosody; it constitutes a holistic and integrated profile encompassing a complex interplay of acoustic parameters. These include, but are not limited to, pitch (fundamental frequency and its variability), intensity (vocal energy and amplitude), timbre (vocal quality and resonance characteristics), harmonic structure (the relationship between fundamental frequency and overtones), spectral texture (the distribution of energy across the frequency spectrum), rhythmic variation (speaking rate, pause duration, and their fluctuations), and dynamic range (the overall variability in vocal expression) (Giddens et al., 2013). This multifaceted understanding of tone of voice is crucial for a comprehensive assessment of vocal expression.

In contexts of chronic stress and burnout, empirical research has consistently reported systematic and measurable alterations in these acoustic parameters. Specifically, studies indicate a reduced variability in fundamental frequency (F0), suggesting a more monotonic vocal delivery, alongside a lowered mean pitch. Furthermore, everyday observations include the elongation of pauses, which may indicate potential respiratory inefficiencies or cognitive processing delays, as well as a measurable decrease in overall vocal energy. These changes are often accompanied by increased spectral roughness, which can be indicative of vocal fold irregularities, and altered harmonic-to-noise ratios, indicating a less clear and more breathy vocal quality (Kouba et al., 2023).

These observed acoustic alterations are not isolated phenomena but are highly consistent with the known pathophysiological profile associated with prolonged activation of the hypothalamic–pituitary–adrenal (HPA) axis, a key component of the body's stress response system. Chronic HPA axis activation can lead to a cascade of physiological changes, including fatigue-induced respiratory inefficiency, which directly impacts vocal production and control. Moreover, the sustained physiological strain can result in diminished neuromotor control of phonation, impairing the precise coordination required for clear and

dynamic vocalisation (Grossi et al., 2015). This link between physiological stress responses and specific vocal changes underscores the potential of vocal analysis as a diagnostic tool.

Consequently, such multidimensional acoustic markers are increasingly recognised as early-stage indicators of stress risk. Their integration into assessment protocols aligns seamlessly with contemporary integrative stress models that emphasise the profound interconnections between physiological, cognitive, and behavioural domains. By providing objective and quantifiable data on an individual's stress state, these vocal markers offer a non-invasive and potentially continuous method for monitoring well-being, paving the way for earlier intervention strategies and personalised stress management approaches. The emerging field of vocal biomarker research thus holds significant promise for both clinical and occupational health settings.

Cutting-edge Speech Emotion Recognition (SER) research, exemplified by studies leveraging benchmark datasets such as the RAVDESS corpus (Luna-Jiménez et al., 2021), is making significant strides in discerning subtle affective states. This research contributes to that advanced deep learning architectures, most notably the Xception model and its numerous variants (Chollet, 2017; Liao et al., 2024), possess a remarkable capacity to capture nuanced spectral–temporal cues. These cues are critical for identifying and understanding complex emotional disturbances like burnout, anxiety, and depression. The ability of these models to process and interpret intricate patterns in speech opens new avenues for early detection and intervention in mental health.

To enhance the practical utility and trustworthiness of these sophisticated SER models, the integration of explainable AI (XAI) methods has become paramount. Techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM; Selvaraju et al., 2017), Local Interpretable Model-agnostic Explanations (LIME; Ribeiro et al., 2016), SHapley Additive exPlanations (SHAP; Lundberg & Lee, 2017), and occlusion sensitivity analysis (Uchiyama et al., 2023) are crucial for demystifying the 'black box' nature of deep learning. These XAI methods serve to reveal the specific spectral bands, temporal frames, and prosodic contours that are most influential in the model's classification decisions. For instance, Grad-CAM can visually highlight particular frequency ranges in a spectrogram that strongly contribute to a burnout classification. At the same time, SHAP values can quantify the individual impact of various acoustic features on the model's output.

Such transparency is not merely an academic exercise; it is essential for the successful integration of SER technologies into clinical settings. By understanding why a model makes a particular prediction, practitioners can gain a deeper insight into a patient's emotional state, complementing traditional diagnostic methods. This fosters practitioner trust, which is a critical barrier to adoption for any new technology in healthcare. Furthermore, meeting stringent regulatory compliance requirements, as emphasised by studies like Shah et al. (2025), necessitates a clear understanding of model behaviour. Regulatory bodies demand accountability and transparency, particularly when AI systems are employed in sensitive domains, such as healthcare. The explainability provided by these XAI methods ensures that SER models are not only accurate but also auditable and ethically sound, paving the way for their widespread and responsible application in mental health diagnostics and monitoring.

The integration of multimodal Artificial Intelligence (AI) techniques is gaining significant traction due to their potential to enhance robustness in emotion recognition, particularly in applications such as burnout detection. A burgeoning consensus advocates for systems that fuse "facial gesture and paralanguage analysis" with complementary data streams. These additional data sources can include physiological measures, such as heart rate variability, which offers insights into autonomic nervous system activity, and contextual behavioural metadata, encompassing details about the user's environment or recent interactions. The rationale behind this multimodal approach is to create a more comprehensive and resilient understanding of an individual's emotional state, moving beyond the limitations of single-modality analysis.

In the specific context of burnout detection, the benefits of such multimodal fusion are particularly pronounced. State-of-the-art frameworks exemplify this by combining audio, visual, and textual modalities. For instance, audio analysis can capture vocal cues indicative of stress or fatigue, visual analysis can track subtle facial expressions and changes in body language, and textual analysis can process written or spoken communication to identify emotional content and sentiment. This synergistic combination significantly enhances system resilience under ecologically valid conditions. These real-world scenarios are often characterised by confounding factors, such as acoustic interference (e.g., background noise), variable recording setups (e.g., different microphone or camera qualities), and heterogeneous speaker populations (e.g., diverse accents, speaking styles, or cultural expressions).

The capacity for these multimodal AI systems to generalise across diverse occupational and cultural

contexts is profoundly valuable, especially for large-scale mental health surveillance and preventive interventions. In occupational settings, employees from various roles and backgrounds may express burnout differently, and a multimodal system can better account for these nuances. Similarly, cultural variations in emotional expression necessitate a flexible and adaptable approach to communication. By integrating multiple streams of data and learning from diverse examples, these systems can develop a more universal understanding of burnout indicators, facilitating early identification and intervention across broader populations. This adaptability is crucial for implementing effective and equitable mental health support programs on a global scale.

From a clinical perspective, a robustly defined and validated "tone of voice" profile holds immense potential for advancing mental health monitoring and intervention. Such a profile could underpin continuous, unobtrusive monitoring systems specifically designed for stress and burnout, enabling early detection through objective vocal biomarkers (Danhof-Pont, van Veen, & Zitman, 2011; Langer et al., 2022). These sophisticated systems would move beyond subjective self-reports, offering a more consistent and real-time assessment of an individual's psychological state.

The key advantage of these monitoring systems lies in their ability to enable real-time risk stratification. By continuously analysing vocal cues, clinicians and caregivers could identify individuals at high risk of escalating stress or burnout before symptoms become debilitating. This proactive approach would facilitate timely intervention, preventing symptom escalation that could otherwise severely compromise an individual's daily functioning, professional performance, or overall health outcomes (Khammissa, Fourie, & Lemmer, 2022). Early intervention could range from targeted psychological support to adjustments in the work environment or lifestyle, ultimately mitigating the long-term impact of chronic stress.

Realising this transformative potential, however, necessitates addressing several critical challenges that span technological, methodological, and ethical domains.

Firstly, the development of standardised, cross-linguistic, and noise-tolerant feature extraction protocols is paramount. This objective, as highlighted by Costantini, Parada-Cabaleiro, Casali, & Cesarini (2022), is crucial for ensuring the widespread applicability and reliability of vocal biomarkers. Such protocols must be meticulously designed to identify and measure vocal features consistently across diverse

populations, accounting for variations in age, gender, dialect, and cultural background. Furthermore, they must exhibit robust performance in various real-world environments, effectively mitigating the confounding effects of background noise, reverberation, and individual vocal variations (e.g., changes due to common colds or temporary voice fatigue). Achieving this standardisation involves rigorous testing and validation across large, heterogeneous datasets to establish universally accepted benchmarks for vocal biomarker extraction.

Secondly, alongside technical advancements in feature extraction, establishing rigorous interpretability frameworks is crucial. As articulated by Low, Bentley, & Ghosh (2020), these frameworks must extend beyond merely delivering high predictive accuracy for conditions like stress and burnout. They must, critically, provide clinically actionable insights. This necessitates a shift from purely statistical correlations to a deeper understanding of the underlying psychological and physiological mechanisms reflected in vocal changes. For instance, an increase in vocal tremor might correlate with heightened anxiety. Still, an interpretability framework would aim to explain why that tremor occurs (e.g., due to increased sympathetic nervous system activation affecting laryngeal muscles) and what specific clinical interventions it might suggest. This comprehensive understanding allows clinicians to move beyond simple diagnostic labels and develop precise, personalised, and effective therapeutic strategies. The ultimate goal is to bridge the gap between raw acoustic data and meaningful clinical recommendations, thereby transforming vocal information from a raw signal into an influential diagnostic, prognostic, and even therapeutic tool. This includes not only identifying individuals at risk but also monitoring treatment efficacy and providing objective feedback on a patient's progress.

1.5 Speech Emotion Recognition: State of the Art and Challenges

Speech Emotion Recognition (SER) stands as a pivotal sub-discipline within the broader, interdisciplinary field of affective computing, drawing foundational knowledge from computer science, psychology, and cognitive science (Tao & Tan, 2005). At its core, SER involves the sophisticated application of Machine Learning and Deep Learning algorithms to extract and interpret emotional states embedded within human speech (Costantini, Cesarini, & Casali, 2022). This burgeoning field holds immense promise across a diverse range of applications. In mental healthcare, SER offers potential for the early detection of various mental health disorders and can aid in clinical diagnostics by providing objective insights into a patient's emotional well-being. Furthermore, its real-time emotional monitoring

capabilities are revolutionising telehealth services, enabling more personalised and responsive care. Beyond clinical settings, SER is increasingly being integrated into organisational well-being strategies, fostering healthier and more productive work environments.

A confluence of significant technological advancements primarily fuels the remarkable acceleration of SER research and development. Foremost among these are the continual improvements in computational modelling, which provide the underlying infrastructure for increasingly complex analytical tasks. Simultaneously, the maturation of deep learning architectures, exemplified by breakthroughs from researchers such as Dal Rí, Ciardi, & Conci (2023) and Atila & Şengür (2021), has dramatically enhanced the accuracy and robustness of emotional state detection. Complementing these algorithmic strides is the escalating availability of high-fidelity emotional speech corpora. These comprehensive datasets, such as EMOVO (Costantini et al., 2014), CREMA-D (Cao et al., 2014), and RAVDESS (Livingstone & Russo, 2018), provide the essential training material for machine learning models, enabling them to learn and generalise across a broad spectrum of emotional expressions and vocal nuances. The convergence of these factors positions SER as a critical area of ongoing research with profound implications for human-computer interaction and emotional intelligence in technology.

A cornerstone in the standardisation of SER evaluation is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). This dataset contains precisely curated recordings encompassing eight core emotional states—calm, happy, sad, angry, fearful, surprised, disgusted, and neutral—each delivered at two intensity levels (Shah et al., 2025). The methodological rigour in capturing balanced, multimodal emotional expressions has supported RAVDESS as a benchmark for algorithm training, validation, and comparative analysis across architectures. The controlled design of this dataset allows researchers to systematically assess feature extraction strategies, generalisation capacity, and the sensitivity of models to prosodic and spectral variations.

Methodologically, contemporary Speech Emotion Recognition (SER) systems are increasingly underpinned by sophisticated deep learning paradigms. Among these, Convolutional Neural Networks (CNNs) have emerged as a foundational architecture due to their efficacy in processing raw audio signals and extracting salient features. More advanced variants, such as the Xception architecture, further enhance this capability. Xception, as detailed by Chollet (2017) and further elaborated in recent work by Liao et al. (2024), is particularly noteworthy for its innovative use of "depth-wise separable

convolutions." This architectural design allows Xception to efficiently capture hierarchical spectral-temporal representations, which are crucial for discerning subtle emotional nuances within speech.

The practical application and performance of these deep learning models have been rigorously tested. For instance, Shah et al. (2025) presented compelling evidence that Xception-based feature extraction, when judiciously paired with supervised classifiers, can yield remarkable accuracy. Their research specifically highlighted the effectiveness of Fine Gaussian Support Vector Machines (FGSVM) in conjunction with Xception, achieving recognition accuracies of over 93%. This high level of accuracy underscores the potential of such integrated systems in real-world SER applications.

Furthermore, the robustness and generalizability of these models can be significantly bolstered through strategic data augmentation techniques. Atmaja and Sasou (2022) supported the substantial benefits of incorporating various augmentation strategies to improve model resilience. These strategies include pitch shifting, which alters the fundamental frequency of the speech without changing its duration; time-stretching, which adjusts the playback speed of the audio; and noise injection, which introduces controlled amounts of background noise into the training data. Such augmentations are vital for training models that can perform reliably across diverse acoustic environments and speaker variations, thereby mitigating issues related to variability and domain shifts commonly encountered in complex SER tasks.

Translating high accuracy from controlled laboratory conditions to field applications in Speech Emotion Recognition (SER) remains a significant challenge due to several inherent complexities of "real-world environments." These environments introduce a multitude of confounding factors, including pervasive acoustic noise, the heterogeneity of recording devices, significant cross-linguistic variability, and pronounced class imbalance. All these elements collectively degrade the stability and reliability of SER models, as extensively documented by research (Atila & Şengür, 2021; Bilotti et al., 2024; Costantini, Cesarini, & Casali, 2022).

Beyond these external environmental factors, the very nature of spontaneous speech in natural contexts presents additional hurdles. Such speech often contains overlapping or rapidly transitioning affective states, which are notoriously tricky for conventional SER systems to accdelineate and classify accurately. To overcome these challenges and preserve performance in ecologically valid settings, advanced domain adaptation strategies are crucial. These strategies aim to bridge the gap between the controlled training

data and the unpredictable characteristics of real-world speech. Furthermore, sophisticated noise-robust feature engineering techniques are crucial for extracting salient emotional cues from noisy or corrupted audio signals, ensuring that the model can still discern emotional states despite imperfections in the input.

Ultimately, addressing these complexities necessitates the design and implementation of highly adaptable and resilient SER systems. These systems must possess the capability to flexibly adapt to variable prosodic contours (such as intonation and rhythm), fluctuating speech rates, and subtle contextual cues that profoundly influence emotional expression. Ensuring such robustness across diverse operational settings—ranging from call centres to clinical diagnostics and human-computer interaction—is paramount for the successful deployment and practical utility of SER technology. This continuous effort in research and development is vital for moving SER from a promising laboratory concept to a dependable and widespread application in various fields.

Despite considerable progress and significant advancements in methodologies and computational power, the field of Speech Emotion Recognition (SER) continues to grapple with fundamental limitations. A primary challenge lies in the absence of standardised, ecologically valid, and truly end-to-end SER pipelines. Such pipelines are crucial for achieving reliable and accurate emotion detection across the vast spectrum of human interaction, which is inherently influenced by diverse sociolinguistic nuances, varying cultural expressions of emotion, and highly dynamic situational contexts. The current lack of such robust systems significantly impedes the transition of SER from laboratory settings to real-world applications.

Beyond technical hurdles, the widespread adoption and large-scale deployment of Speech Emotion Recognition (SER) technologies are complicated by a complex set of unresolved ethical considerations. Paramount among these is the protection of voice data privacy, given that such data “can reveal a person’s emotional and psychological state” and must therefore be anonymised, securely stored, and handled in compliance with regulations such as GDPR. Informed consent remains a cornerstone of ethical research, requiring that participants “are fully informed about the nature of the research, the type of data being collected, how it will be used, and their rights to withdraw” (Buccoliero, 2025). The risk of algorithmic bias, particularly from non-representative training datasets, poses a substantial threat to fairness and accuracy, with potential discriminatory effects if left unaddressed (Meissen et al., 2024).

SER research is navigating a complex landscape defined by the need to harmonise algorithmic sophistication, contextual adaptability, and ethical accountability. This thesis aims to contribute to this trajectory by developing transparent, high-accuracy, context-aware SER systems explicitly tailored to detect early markers of burnout, thus enabling proactive interventions in both clinical and occupational health domains.

1.6 Burnout Syndrome: Clinical Definition and Measurement

Burnout Syndrome, as a focal but bounded construct, was initially considered a vague occupational complaint; it has since become a critical concern with profound implications for individual mental health, organisational productivity, and overall societal well-being (Heinemann & Heinemann, 2017). Its growing recognition stems from an increasing understanding of its far-reaching consequences, impacting not only the afflicted individuals but also the broader economic and social fabric.

Despite its pervasive presence in clinical discussions, academic research, and managerial strategies, burnout continues to grapple with a fundamental challenge: the absence of universally accepted diagnostic criteria and a standardised classification system (Nadon, De Beer, & Morin, 2022). This lack of consensus creates a significant obstacle to advancing the field. Without clear and consistent diagnostic markers, the comparability of empirical studies is severely hampered. Researchers employ diverse methodologies and criteria, leading to fragmented findings that are difficult to synthesise and apply across different contexts.

Furthermore, this diagnostic ambiguity directly impedes the development of coherent, evidence-based prevention and intervention strategies (Heinemann & Heinemann, 2017; Nadon et al., 2022). Health professionals and organisations struggle to implement effective measures when the very definition and presentation of burnout remain open to interpretation. This variability makes it challenging to design targeted programs, assess their efficacy, and ultimately, provide consistent and reliable support to those at risk or experiencing burnout.

Consequently, prevalence estimates for burnout exhibit considerable variation across different countries, economic sectors, and the methodological approaches used in studies. This wide range of reported

figures poses substantial challenges for policymakers who aim to allocate resources effectively and for health professionals who strive to understand the true scope of the problem within their respective populations. The inconsistencies underscore the urgent need for a unified understanding of burnout to facilitate more accurate assessment, effective intervention, and robust policy development.

Clinically, the assessment of burnout predominantly relies on validated self-report psychometric instruments. Among these, the Maslach Burnout Inventory (MBI) is widely considered the "gold standard" for its comprehensive measurement of three core dimensions: emotional exhaustion, depersonalization (often referred to as cynicism), and a reduced sense of personal accomplishment (Parker & Tavella, 2022). These dimensions capture the multifaceted nature of burnout, encompassing emotional depletion, interpersonal detachment, and a diminished sense of efficacy.

While the Maslach Burnout Inventory (MBI) remains the most widely recognised and utilised instrument for assessing burnout, a range of alternative assessment tools exists, each contributing unique perspectives and nuances to the understanding of this complex syndrome. These diverse instruments underscore the multifaceted nature of burnout and the varying theoretical frameworks employed in its study.

One such alternative is the Karolinska Exhaustion Disorder Scale (KEDS). The KEDS distinguishes itself by focusing specifically on "exhaustion disorder," a concept rooted in prolonged exposure to stress and subsequent profound fatigue (Grossi, Perski, Osika, & Savic, 2015). Unlike the MBI's tripartite model of emotional exhaustion, depersonalization, and reduced personal accomplishment, the KEDS zeroes in on the core physiological and psychological depletion that characterises this specific state of prolonged stress. Its emphasis is often on the physical and cognitive manifestations of exhaustion, providing a different lens through which to understand the debilitating effects of chronic stress.

Similarly, the Shirom–Melamed Burnout Questionnaire (SMBQ) provides an alternative theoretical approach to assessing burnout. The SMBQ measures burnout across three distinct dimensions: physical fatigue, emotional exhaustion, and cognitive weariness (Danhof-Pont, van Veen, & Zitman, 2011). While "emotional exhaustion" overlaps with a key component of the MBI, the SMBQ's inclusion of "physical fatigue" and "cognitive weariness" highlights the tangible and debilitating impact of burnout on an individual's physical and mental capacity. This broader scope acknowledges that burnout extends

beyond emotional depletion, encompassing a profound sense of physical tiredness and a noticeable decline in cognitive function, such as difficulty concentrating or making decisions.

These alternative tools, despite often capturing constructs that partially overlap with those assessed by the MBI, are crucial because they frequently emphasise distinct aspects of burnout. This diversity in assessment highlights the inherent complexity of the phenomenon and the variety of theoretical conceptualisations that inform its study and measurement. The existence of these different instruments enriches the research landscape by allowing for a more granular and comprehensive understanding of burnout, acknowledging its various forms, manifestations, and underlying mechanisms. Researchers and practitioners can select the most appropriate tool based on their specific research questions, clinical focus, or the particular theoretical model of burnout they wish to explore.

However, a significant challenge in the field is the absence of universally applied diagnostic thresholds for these measures. This lack of standardisation contributes substantially to marked heterogeneity in prevalence estimates across studies and significantly complicates meta-analytic synthesis of research findings (Parker & Tavella, 2022). The varying cut-off scores used in different studies make it difficult to compare results and draw firm conclusions about the true prevalence and impact of burnout.

Furthermore, the inherent reliance on subjective self-report in these instruments introduces several methodological concerns. Issues such as recall bias, where individuals may inaccurately remember or report their symptoms, can skew results. Cultural differences in symptom expression also pose a challenge, as the way individuals perceive and articulate their experiences of burnout can vary significantly across cultural contexts. Moreover, self-report measures may fail to adequately capture subclinical presentations of burnout, where individuals experience symptoms that are impactful but do not meet formal diagnostic criteria, potentially leading to an under-recognition of the true scope of the problem. These limitations underscore the need for continued research into objective measures and standardised diagnostic criteria to enhance the accuracy and comparability of burnout assessments.

Recent research increasingly highlights the biopsychosocial nature of burnout, implicating dysregulation across multiple physiological systems. Perturbations in the hypothalamic-pituitary-adrenal (HPA) axis, altered autonomic nervous system (ANS) balance, cardiovascular irregularities, immune system dysfunction, and metabolic and endocrine disruptions, including those beyond cortisol secretion, have

all been observed (Danhof-Pont et al., 2011; Khamissa et al., 2022). These biological correlates reinforce the need for multimodal assessment strategies. However, causal mechanisms remain under debate due to the predominance of cross-sectional study designs and insufficient longitudinal monitoring.

The search for objective and non-invasive biomarkers has intensified, with vocal analysis emerging as a particularly promising avenue. Although no not yet clinically validated vocal biomarker for burnout currently exists, advances in AI-driven speech analytics have supported the capacity to detect subtle prosodic, spectral, and temporal shifts associated with chronic stress and emotional dysregulation (Grządzielewska, 2021; Langer et al., 2022). Vocal markers offer unique advantages: they can be captured passively and repeatedly in naturalistic contexts, enabling ecologically valid, continuous monitoring that could complement traditional questionnaires and physiological testing (Low et al., 2020).

To overcome the inherent limitations of current self-report measures in assessing burnout, qualitative studies and expert panels strongly advocate for the adoption of integrated diagnostic frameworks. These advanced frameworks propose to merge validated self-report instruments with objective physiological and behavioural indicators. Notably, the inclusion of voice-derived features is highlighted as a promising avenue to enhance the accuracy and comprehensiveness of burnout assessment (Guarrasi et al., 2025; Low, Bentley, & Ghosh, 2020).

Further augmenting this approach, multimodal designs are increasingly recognised for their potential to revolutionise burnout tracking. By combining vocal, visual, and movement features (Galatzer-Levy et al., 2021; Muzammel et al., 2021), these designs directly address the shortcomings of approaches that rely solely on self-report. A key advantage of multimodal designs is their ability to enable continuous, ecologically valid tracking of burnout trajectories across diverse contexts. This constant monitoring capability offers a dynamic and nuanced understanding of an individual's burnout state, surpassing snapshot assessments. This paradigm shift aligns seamlessly with broader precision health initiatives, where the ultimate goal is to generate personalised, data-rich profiles. These comprehensive profiles are then utilised to guide proactive and highly targeted interventions, ultimately fostering a more personalised and practical approach to mental well-being (Gómez-Vilda et al., 2022). The integration of such diverse data streams holds promise for developing more robust and predictive models to identify, monitor, and mitigate burnout.

In conclusion, the complexities of defining, measuring, and monitoring burnout necessitate an interdisciplinary, data-driven approach to understanding and addressing this phenomenon. By uniting psychometric assessments, physiological indicators, and innovative vocal analytics into a cohesive framework, researchers and practitioners can enhance diagnostic specificity, detect early warning signs, and design tailored intervention strategies for both clinical populations and high-risk occupational groups.

1.7 Research Hypotheses and Objectives

The overarching objective of this doctoral research is to advance the understanding and application of Speech Emotion Recognition (SER) as a tool for assessing emotions and psychological states in various contexts, including applied, clinical, and technological settings. The project examines how vocal biomarkers can serve as reliable, ethical, and interpretable indicators of affective states, with a specific focus on detecting burnout, addressing emotional dysregulation, and monitoring mental health. The research is developed through four complementary studies, each contributing to a refined theoretical and empirical understanding of emotion recognition from voice.

Primary Hypothesis (H1) – Emotional Signatures in Voice: Specific acoustic and prosodic features (e.g., pitch variability, MFCCs, speech rate, jitter, shimmer) systematically correlate with discrete and dimensional emotional states (valence and arousal), supporting their role as quantifiable biomarkers of affective expression.

Secondary Hypothesis (H2) – burnout, as a focal but bounded construct, and Emotional Dysregulation: Voice patterns reflecting reduced pitch range, slower tempo, and lower spectral energy are predictive of burnout-related emotional exhaustion and chronic stress, distinguishing affected individuals from controls with statistically significant accuracy.

Exploratory Hypothesis (H3) – Human-AI Interpretability Alignment: The inclusion of psychologically meaningful acoustic features enhances the interpretability and user trust of SER outputs, aligning computational inferences with human emotion perception and clinician judgments. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Exploratory Hypothesis (H4) – Predictive Validity of Vocal Biomarkers: Cross-validation results demonstrate that SER-based emotional profiles can predict changes in self-reported affect, stress indices, and psychometric scores, suggesting potential for longitudinal monitoring and early detection of emotional imbalance. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Exploratory Hypothesis (H5) – Integration into Digital Health Ecosystems: SER tools can be integrated into digital health platforms to support early screening and personalized intervention for emotional and psychological well-being, confirming the translational potential of the research. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

The expected contributions include:

- Scientific Contribution: Integration of psychological and computational paradigms for emotion recognition, offering new empirical evidence on voice as a diagnostic medium.
- Methodological Contribution: Development of a validated, interpretable, and bias-aware SER framework. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.
- Applied Contribution: Creation of a prototype for emotional assessment integrated within healthcare digital platforms.
- Ethical Contribution: Formulation of best-practice guidelines for responsible AI use in psychological and clinical settings.

These hypotheses and objectives guided the main research papers, aligning the doctoral project with its ultimate purpose: establishing SER as a valid, ethical, and interdisciplinary bridge between human psychology and artificial intelligence. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

1.8 Thesis Outline

The preceding sections have articulated an intricate, multi-layered framework integrating theoretical,

empirical, and methodological perspectives essential to this doctoral investigation. The discourse commenced with an examination of the escalating societal, clinical, and occupational imperatives to address burnout, synthesising insights from public health, occupational psychology, and biomedical sciences to underscore its multifaceted biopsychosocial character. The analysis of socio-technological and organisational shifts illuminated the profound influence of digital transformation on emotional well-being—both as a potential source of strain and as an avenue for innovative, preventive health strategies—thereby situating digital tools at the nexus of monitoring, intervention, and patient empowerment.

By interweaving foundational psychological theories of emotion with emerging evidence on vocal biomarkers, the work has supported a robust conceptual framework for interpreting how speech-derived indicators can serve as non-invasive, scalable proxies for affective and cognitive states associated with burnout. The review of state-of-the-art Speech Emotion Recognition (SER) methodologies, along with their attendant challenges of accuracy, domain generalisation, interpretability, and ethical stewardship, positions this research within the vanguard of applying artificial intelligence in precision health contexts.

The critical appraisal of burnout's clinical definitions and measurement paradigms revealed persistent conceptual ambiguities and methodological fragmentation, underscoring the need for integrative, multimodal diagnostic frameworks. Such frameworks should conjoin psychometric evaluation with physiological and behavioural metrics—particularly those obtained through advanced speech analytics—to enhance diagnostic specificity, temporal sensitivity, and ecological validity.

From this interdisciplinary synthesis, the research hypotheses and objectives have emerged with clarity, aiming to advance AI-driven voice analysis as both a diagnostic adjunct and a catalyst for patient engagement, while contributing to ongoing debates on explainable and ethically aligned AI in healthcare.

The dedicated period within GPI's R&D laboratories was meticulously focused on the comprehensive investigation, innovative design, and robust development of a specialised application named "Talking About". The core purpose of this application is the systematic acquisition of diverse datasets, specifically curated for the rigorous training and subsequent preliminary support for an Artificial Intelligence agent. This sophisticated tool facilitates the collection of a wide array of heterogeneous data, encompassing crucial vocal samples that capture nuanced prosodic and acoustic features. These detailed psychometric

assessments provide insights into cognitive and emotional states, as well as essential demographic information to contextualise the collected data.

Figure 1 - GPI's Talking About Mock-up

This pivotal phase of the research encompassed several critical stages. Initially, a meticulous identification of critical acoustic features was made, which were hypothesised to be indicative of emotional states or other psychological attributes. This was followed by the careful design and iterative training of various machine learning models, leveraging advanced algorithms to discern intricate patterns within the acquired data. To empirically validate the algorithm's efficacy and robustness in real-world scenarios, a comprehensive pilot study was conducted with meticulous implementation. This study was specifically designed to test the algorithm's capacity to accurately detect and classify emotions within ecologically valid speech contexts, ensuring that the findings were generalizable and applicable to natural human communication.

Throughout the entirety of these research activities, there was a steadfast and close collaboration with GPI's R&D department. This synergistic partnership was instrumental in guaranteeing not only methodological rigour in every step of the research process but also a strict adherence to the highest technological standards. This alignment with GPI's internal technological frameworks was paramount, as it ensured seamless compatibility and facilitated the subsequent large-scale implementation of the developed AI agent, paving the way for its integration into broader applications and systems.

1.9 Overview of publications

The thesis encompasses the outcomes of 4 of my studies and my contribution to the related publications:

Study_1. Riccardo Sartori, Francesca Marinaro, Francesco Tommasi, Andrea Buccoliero, Mattia Zene, Syed Adil Hussain Shah e Andrea Ceschi. *A scoping review on the use of voice biomarkers for emotional assessment*; 2025. EJPS 2025. https://doi.org/10.1027/1015-

5759/a000915

Study_2. Giuseppe Lentini, Paolo Ranzi, Isabella Della Torre, Ismaela Avellino, Andrea Buccoliero, Antonio Colangelo; *A Focused Approach for Speech Emotion Recognition in Real-World Environments*. IADIS Conference. Big Data Analytics, Data Mining and Computational Intelligence 2025. ISBN (Book): 978-989-8704-70-2

Study_3. Tagliente S, Minafra B, Aresta S, Santacesaria P, Buccoliero A, Palmirotta C, Lagravinese G, Mongelli D, Gelao C, Macchitella L, Pazzi S, Scrutinio D, Baiardi P e Battista P (2025). *Effectiveness of a home-based computerised cognitive training in Parkinson's disease: a pilot randomised crossover study*. Front. Psychol. 15:1531688. doi: 10.3389/fpsyg.2024.1531688

Study_4. Syed Taimoor Hussain Shah, Syed Adil Hussain Shah, Andrea Buccoliero, Iqra Iqbal Khan, Syed Baqir Hussain Shah, Angelo Di Terlizzi, & Giacomo Di Benedetto (2025). *Analysing Neonatal Vocal Expression: Methodological Approaches to Identifying Neurological and Psychiatric Signatures*. Journal of Multiscale Neuroscience, 4(2): 158-176. DOI: https://doi.org/10.56280/1703023560

In addition to these studies, the period spent at GPI allowed me to work and contribute to the following publications, too:

Study_5. Bassi, C., Marinaro, F., Anarbaeva, A., Buccoliero, A., Scandola, M., Sartori, R., Ceschi, A. (2025, September 3). Developing and Validating an AI Model to Detect Burnout Risk from Vocal Biomarkers. https://doi.org/10.17605/OSF.IO/HMK4F

Study_6. Chiara Bassi; Francesca Marinaro; Akylai Anarbaeva; Andrea Ceschi; Andrea Buccoliero; Ismaela Avellino; Isabella Della Torre; Syed Adil Hussain Shah; Giuseppe Lentini; Michele Scandola; Riccardo Sartori. TALIA (Intelligent Vocal Biomarkers for Pathology Assessment): A Prospective Study to Validate an AI Model for burnout, as a focal but bounded construct, Risk Classification. EJPA - 2025

Study_7. Shah, S.A.H.; Shah, S.T.H.; Khaled, R.; Buccoliero, A.; Shah, S.B.H.; Di Terlizzi, A.; Di Benedetto, G.; Deriu, M.A. *Explainable AI-Based Skin Cancer Detection Using CNN, Particle Swarm Optimisation and Machine Learning*. J. Imaging 2024, 10, 332. https://doi.org/10.3390/jimaging10120332

Study_8. S. T. Hussain Shah; Syed Adil Hussain Shah, K. Panagiotopoulos, J. Pigueiras-del-Real,

K. Qayyum, S. B. Hussain Shah, A. Buccoliero, A. Di Terlizzi, M. A. Deriu, "Explainable Emotion Recognition Using Xception-Based Feature Extraction and Supervised Machine Learning on the RAVDESS Dataset" 2025 IEEE Medical Measurements & Applications (MeMeA), Chania, Greece, 2025, pp. 1-6, doi: 10.1109/MeMeA65319.2025.11068008.

Study_9. Rashidi, M.; Arima, S.; Stetco, A.C.; Coppola, C.; Musarò, D.; Greco, M.; Damato, M.; My, F.; Lupo, A.; Lorenzo, M.; et al. *Prediction of Parkinson's Disease Using Long-Term, Short-Term Acoustic Features Based on Machine Learning.* Brain Sci. 2025, 15, 739. https://doi.org/10.3390/brainsci15070739

Study_10. Chiara Bassi, Francesca Marinaro, Andrea Buccoliero, Anna Maria Meneghini, Riccardo Sartori, Andrea Ceschi. Preliminary Qualitative Study on AI and burnout, as a focal but bounded construct: Diagnostic Potential of Vocal Biomarkers. https://hdl.handle.net/11562/1163887

Study_11. Chiara Bassi, Andrea Buccoliero, Anna Maria Meneghini, Riccardo Sartori, Francesco Tommasi, Andrea Ceschi. *What Matters to You? Exploring Determinants of Patient Engagement Through a Multilevel Review*. 2025. https://hdl.handle.net/11562/1158651

Although not included among the main studies of this thesis due to timing constraints, it is important to acknowledge the relevance of the Study_5 by Bassi et al. (2025), which was only recently finalized, submitted, and formally registered as a protocol on Open Science Framework. This work represents a significant advancement in the research trajectory developed throughout the thesis, as it moves more directly toward the empirical validation of AI-based models for detecting burnout risk from vocal biomarkers. While the present thesis primarily establishes the conceptual, methodological, and translational foundations of voice-based assessment, this study extends that line of inquiry by operationalizing and testing a concrete predictive framework. Its exclusion from the main body of the dissertation is therefore not indicative of a lack of relevance but rather reflects the temporal boundaries of the doctoral process. In this sense, the study can be interpreted as a continuation and consolidation of the thesis contributions, providing a more targeted and prospective step toward the type of validation that the thesis itself frames as necessary for future research.

In synthesis, the variety of studies I contributed to describes how this research incorporates a multidisciplinary approach, bridging psychology, artificial intelligence, and digital health to explore digital care frameworks, advancing patient empowerment while addressing complex emotional and

behavioural health challenges.

2. Literature Review

Speech Emotion Recognition (SER) broadly refers to the Machine Learning and Deep Learning (ML/DL) algorithms employed to identify emotional states from human speech. The literature reveals a variety of supported speech analysis and classification techniques applied to extract emotions from signals. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

The literature review, thoroughly detailed in this chapter, adopts a scoping review approach to map these techniques, outlining key elements that define the algorithmic assessment of emotions. These elements include the databases used for emotion recognition, the concepts and types of emotions considered, empirical studies on SER, and their associated limitations. The study emphasises current perspectives and practices to provide recommendations for future advancements.

The review culminated in the comprehensive paper A scoping review on the use of voice biomarkers for emotional assessment, which has been accepted for publication in the European Journal of Psychological Assessment. The entire study is reported within Chapter 4 of this thesis, providing an in-depth analysis and synthesis of existing research in the field. The review meticulously examined various methodologies, findings, and applications of voice biomarkers, highlighting their potential and limitations in the precise and objective evaluation of human emotions. The insights garnered from this rigorous review form a foundational component of the broader research presented in this thesis, informing subsequent chapters and contributing to a deeper understanding of the subject matter.

2.1 Theoretical Framework

The intersection of affective computing, psychology, and digital health provides fertile ground for investigating voice as a biomarker for emotional assessment (Tao & Tan, 2005; Eyben et al., 2015). This interdisciplinary approach not only holds significant theoretical importance for advancing emotion science (Russell, 1980; Schirmer, 2018) but is also increasingly recognised as crucial for applications in occupational health. Specifically, it offers promising avenues for preventing work-related syndromes, such as burnout —a pervasive issue in modern workplaces (Taguchi et al., 2018; Low et al., 2020).

Within this evolving paradigm, voice biomarkers are conceptualised far beyond mere indicators of momentary affective states. Instead, they are viewed as integral components of sophisticated digital infrastructures designed for early detection, continuous monitoring, and proactive preventive intervention (Shinohara et al., 2021; Faurholt-Jepsen et al., 2021). This strategic positioning of voice-based technologies aligns seamlessly with a broader, global trend towards developing scalable, technology-enabled solutions. Such solutions are critical for alleviating the growing burden on traditional healthcare systems, which often struggle to provide sufficient resources for mental and emotional well-being. Concurrently, these digital tools empower individuals to take a more active role in managing their own emotional health, fostering self-awareness and providing accessible support (Muzammel, Salam, & Othmani, 2021; Pérez-Toro et al., 2023). The ability of voice analysis to provide objective, non-invasive, and continuous data makes it a particularly valuable asset in this push towards more accessible and personalised health management.

The theoretical foundation for this research posits that emotions are not merely internal, subjective experiences, but rather complex phenomena that manifest physically and are communicated through various external channels, notably speech (Ekman, 1992; Russell, 1980). This perspective moves beyond a purely cognitive or introspective view of emotion, emphasising its embodied nature and its role in social interaction.

The human voice, in particular, serves as a rich repository of paralinguistic features that offer profound insights into an individual's affective states. These features include, but are not limited to, pitch (the fundamental frequency of the voice), intonation (the rise and fall of pitch), prosodic rhythm (the pattern of stressed and unstressed syllables), temporal dynamics (such as speech rate, pauses, and speech onset time), and spectral patterns (the distribution of energy across different frequencies in the voice) (Costantini, Cesarini, & Casali, 2022; Pfister & Robinson, 2010). Unlike self-report instruments, which can be susceptible to biases, conscious manipulation, or limitations in introspection, these paralinguistic markers often provide access to emotional nuances that are either unconsciously expressed or difficult for individuals to articulate verbally.

The significance of these vocal cues lies in their direct reflection of the intricate interplay between autonomic nervous system activity, cognitive appraisal processes, and social expressive behaviours. For instance, changes in heart rate, breathing patterns, and muscle tension, which are regulated by the

autonomic nervous system, can directly influence vocal parameters like pitch and loudness. Simultaneously, cognitive appraisals of a situation—how an individual interprets and evaluates an event—shape the specific emotional response, which in turn influences vocal expression. Furthermore, social display rules and the communicative intent of the speaker also modulate how emotions are vocalised. This multifaceted integration makes speech a vibrant, dynamic, and ecologically valid medium for detecting and analysing human emotion in naturalistic settings (Schirmer, 2018; Giddens et al., 2013). Consequently, analysing these vocal features allows researchers to gain a more comprehensive understanding of emotional experience that transcends what can be gleaned from overt verbal content alone.

The theoretical framework for this study is built upon a comprehensive understanding of emotion, drawing from several foundational psychological perspectives. The James–Lange theory serves as a cornerstone, positing that physiological responses precede and contribute to the subjective experience of emotion. This view emphasises the crucial role of bodily changes—such as heart rate acceleration, muscle tension, or skin conductance—as direct antecedents to the feeling of an emotion (Giddens et al., 2013). Thus, an individual might first experience an increased heart rate and then interpret that physiological change as fear.

In contrast, appraisal theories highlight the cognitive dimension of emotional experience. These theories argue that emotions are not merely automatic physiological reactions but are instead shaped by an individual's evaluation and interpretation of environmental events (Pfister & Robinson, 2010). For instance, encountering a growling dog might elicit fear if it is perceived as a threat, but amusement if it is perceived as playful. This cognitive appraisal process involves assessing the significance of an event in relation to one's goals, well-being, and coping potential.

To provide a more nuanced understanding of emotional states, Russell’s circumplex model of affect offers a dimensional approach to understanding emotions. This model organises emotions along two primary orthogonal axes: arousal (ranging from low to high activation) and valence (ranging from unpleasant to pleasant) (Russell, 1980). This allows for the mapping of a wide range of emotions within a two-dimensional space, where, for example, "excitement" is characterised by high arousal and high valence. In contrast, "sadness" is characterised by low arousal and low valence. This dimensional view provides a valuable framework for understanding the underlying structure of emotional experience.

More contemporary integrative models further synthesise these perspectives, recognising that emotion is a complex interplay of physiological, cognitive, and social dimensions (Schirmer, 2018). These models acknowledge that while bodily changes and cognitive appraisals are vital, social context, cultural norms, and individual learning also significantly shape how emotions are experienced, expressed, and regulated. This holistic view underscores the dynamic and multifaceted nature of emotional processes.

Within this rich theoretical landscape, Speech Emotion Recognition (SER) emerges as a critical methodological and conceptual bridge. SER explicitly synthesises these diverse psychological insights with advanced computational techniques (Schuller, Steidl, & Batliner, 2009). By analysing acoustic markers—such as pitch, intensity, speaking rate, and vocal quality—SER offers a non-invasive and objective means to infer emotional and cognitive states from speech (Rejaibi et al., 2022). This approach is grounded in the understanding that emotional states manifest in measurable physiological and behavioural changes, which are acoustically encoded in the voice.

Therefore, the theoretical framework underpinning this research posits a bidirectional integration between psychological constructs and computational methodologies. On one hand, supported psychological theories of emotion—including the James–Lange perspective, appraisal theories, the circumplex model, and integrative frameworks—provide the foundational understanding and conceptual guidance for the development and refinement of computational models used in SER. These theories inform the selection of relevant acoustic features and the interpretation of algorithmic outputs. On the other hand, the algorithmic outputs generated through SER feed back into psychological theory and applied practice. The objective, data-driven insights derived from acoustic analysis can enrich our understanding of emotional processes, validate existing theories, and potentially uncover new nuances in emotional expression and cognition. Furthermore, these computational insights have direct implications for various applied domains, such as human-computer interaction, mental health diagnostics, and behavioural analysis, thereby strengthening the practical utility of psychological research. This synergistic relationship ensures that both theoretical understanding and practical application are continuously refined and advanced.

2.2 Review of Existing Research

The increasing interest in voice biomarkers for emotional assessment is a testament to the advancements in digital signal processing and artificial intelligence. As highlighted in the scoping review by Buccoliero et al. (2025), a substantial body of research is dedicated to this emerging field. The core principle revolves around the premise that human emotional states manifest in subtle, yet detectable, alterations in vocal characteristics.

Empirical investigations consistently demonstrate that specific acoustic features serve as reliable indicators of emotional expression. These features include, but are not limited to, Mel-Frequency Cepstral Coefficients (MFCCs), which represent the short-term power spectrum of a sound, providing a robust representation of timbre. Prosodic variation, encompassing elements such as pitch, rhythm, and intonation, offers crucial insights into the emotional nuances conveyed through speech. Additionally, micro-perturbations in vocal fold vibration, quantified by measures like jitter (variability in pitch period) and shimmer (variability in amplitude), have been shown to correlate with various emotional states. Spectral descriptors, which analyse the distribution of energy across different frequencies, further enrich the understanding of vocalic emotional markers.

The convergence of these acoustic features allows for effective differentiation between various emotional states, ranging from joy and excitement to sadness and anger. Beyond mere emotional classification, the utility of these biomarkers extends to predicting clinical conditions. For instance, research by Taguchi et al. (2018), Low et al. (2020), and Pérez-Toro et al. (2023) has independently supported the potential of voice biomarkers in identifying indicators of depression, stress, and even post-traumatic stress disorder (PTSD). This diagnostic potential underscores the transformative impact these technologies could have on mental health screening and monitoring.

A key advantage driving the widespread adoption of voice biomarkers is their inherent non-invasiveness. Unlike traditional diagnostic methods that may require physical contact or specialised equipment, voice data can be collected passively and remotely, minimising patient discomfort and logistical hurdles. The ease of collection further contributes to their appeal, as simple recording devices, such as smartphones or personal computers, can capture the necessary vocal samples. This accessibility naturally paves the way for seamless integration with modern healthcare technologies. The work by Di Cesare et al. (2024) and Tracey et al. (2023) exemplifies how these markers can be effectively integrated into mobile health (mHealth) applications and telehealth platforms, enabling continuous monitoring and early intervention

in various clinical and non-clinical settings. This integration holds significant promise for democratizing access to mental health support and enhancing the proactive management of emotional well-being.

The landscape of emotion recognition through vocal analysis has seen significant advancements, with recent research highlighting both the robustness of acoustic markers and the sophistication of contemporary machine learning and deep learning architectures (Dal Rí, Ciardi, & Conci, 2023). Higuchi et al. (2022) supported systematic changes in vocal dynamics within populations experiencing chronic stress, underscoring the physiological impact of stress on speech patterns. This finding contributes to the growing body of evidence linking psychological states to observable vocal characteristics.

Further expanding the applicability of these models, Hansen et al. (2022) provided compelling evidence for the portability of Speech Emotion Recognition (SER) models across diverse linguistic boundaries. Their research utilised transfer learning techniques, effectively showcasing the cross-cultural potential of these models, which is crucial for developing universally applicable diagnostic and analytical tools. This portability suggests that the fundamental acoustic cues for emotion may transcend language-specific phonetic variations.

The development and widespread adoption of benchmark datasets have been pivotal in propelling research in this field. Datasets such as RAVDESS (Livingstone & Russo, 2018), EMOVO (Costantini et al., 2014), CREMA-D (Cao et al., 2014), and DAIC-WOZ (Gratch et al., 2014) have played a crucial role by providing standardised resources for model training and evaluation. These shared resources enable consistent replication of studies and facilitate robust cross-study comparisons, which are essential for scientific progress. The consistent reporting of high classification accuracies, frequently surpassing 80–90%, within studies utilising these datasets further provides preliminary support for the effectiveness of current SER methodologies.

Moreover, the integration of semantic content with paralinguistic features has been a significant leap forward in enhancing predictive accuracy (Ma et al., 2024). Advances in natural language processing (NLP) have allowed researchers to move beyond purely acoustic analysis, incorporating the meaning and context of spoken words. This holistic approach, which combines what is said with how it is said, provides a more comprehensive understanding of emotional states and significantly enhances the accuracy of emotion recognition systems. This synergistic approach, combining acoustic analysis and

semantic understanding, represents a promising direction for future research and applications in areas such as mental health monitoring, human-computer interaction, and customer service analytics.

Nevertheless, the existing literature on Speech Emotion Recognition (SER) and vocal biomarkers continues to struggle with substantial methodological heterogeneity, resulting in persistent fragmentation. This fragmentation manifests across various crucial aspects of research design and implementation. For instance, emotion elicitation protocols vary widely, with some studies relying on acted emotions for controlled environments, while others utilise spontaneous speech gathered from clinical or experimental contexts. This divergence in elicitation methods often contributes to inconsistent and incomparable outcomes across investigations (Costantini, Cesarini, & Casali, 2022; Taguchi et al., 2018).

Furthermore, the affective taxonomies employed to categorise and describe emotions exhibit considerable variability. Researchers adopt different models of emotion, leading to a lack of a standardised framework for classifying emotional states from vocal cues. This absence of a unified taxonomy complicates meta-analyses and the generalisation of findings. Similarly, feature selection strategies differ significantly, with researchers exploring a vast array of acoustic features, including prosodic, spectral, and voice quality parameters. The optimal set of features for robust SER remains a subject of ongoing debate, contributing to the heterogeneity of results (Pfister & Robinson, 2010; Schuller, Steidl, & Batliner, 2009).

Validation procedures also present a significant source of methodological divergence. Studies employ a variety of cross-validation techniques, dataset sizes, and evaluation metrics, making direct comparisons of model performance challenging. The predominance of laboratory-based paradigms, while offering controlled experimental conditions, often constrains ecological validity. This limitation hinders the direct translation of findings into real-world clinical or occupational environments, where speech patterns and emotional expressions are far more complex and nuanced (Low et al., 2020; Min et al., 2023).

A critical gap also exists in the collaboration between computational scientists and psychologists. This limited interdisciplinary exchange often results in a disconnect between advancements in algorithmic optimisation and the practical clinical interpretability of SER models (Schirmer, 2018). While

computational scientists focus on improving model accuracy and efficiency, psychologists are more concerned with the clinical relevance, reliability, and validity of these technologies in diagnostic or therapeutic contexts.

Finally, many existing studies in SER remain exploratory in scope, lacking the rigorous longitudinal validation and integration into supported real-world diagnostic workflows necessary for widespread clinical adoption (Rejaibi et al., 2022; Shinohara et al., 2021). The absence of long-term studies tracking the consistency and predictive power of vocal biomarkers over time limits their utility in monitoring disease progression or treatment efficacy. Moving forward, greater emphasis on standardised methodologies, interdisciplinary collaboration, and robust real-world validation will be crucial for advancing the field of SER and unlocking the full potential of vocal biomarkers in clinical and occupational settings.

2.3 Identification of Research Gaps

The research gaps are therefore framed as identifying a plausible and still under-specified mechanism rather than as establishing a validated explanatory model. These research gaps indicate a plausible but still under-specified mechanism, rather than a validated explanatory model.

Several critical challenges currently impede the widespread adoption and clinical utility of Speech Emotion Recognition (SER) systems within psychological and healthcare contexts. Addressing these limitations is paramount for advancing the field and realising the full potential of SER as a diagnostic and therapeutic tool. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Insufficient Integration with Psychology: A significant shortcoming lies in the predominant focus on technical model accuracy at the expense of robust theoretical grounding in psychology. Many existing studies prioritise algorithmic sophistication over deep psychological understanding of human emotion. This often leads to outputs with limited clinical interpretability, failing to account for the nuanced, subjective, and context-dependent nature of emotional experience. The neglect of the subjective complexity of human emotion, as highlighted by Costantini et al. (2022) and Lin et al. (2022), means that current SER models may misrepresent or oversimplify intricate emotional states, rendering them

less useful for clinical decision-making. Future research must bridge this gap by incorporating supported psychological theories of emotion, cognitive processes, and psychopathology into the development of the SER model.

Lack of Standardisation: The current landscape of SER research is characterised by a notable lack of standardisation, which severely impedes comparability and progress. Divergent datasets, varied feature extraction methods, and a multitude of evaluation metrics make it exceptionally difficult to synthesise findings across studies. The absence of universally accepted benchmarks prevents meaningful aggregation of results, making it challenging to identify the most effective approaches or to track improvements consistently (Low et al., 2020; Min et al., 2023). Establishing standard protocols for data collection, annotation, feature representation, and evaluation is crucial for fostering cumulative science and accelerating the development of robust and reliable SER systems.

Scarce Real-World Validation: Despite promising laboratory results, a significant limitation of SER research is the scarcity of real-world validation. Only a minority of studies have rigorously tested SER applications in ecologically valid settings that mirror actual clinical practice, workplaces, or telehealth platforms. The vast majority of datasets are collected under controlled laboratory conditions, which, while applicable for initial model development, often fail to capture the variability, noise, and complexity inherent in naturalistic speech (Rejaibi et al., 2022; George & Ilyas, 2024). This reliance on laboratory data severely limits the generalizability and practical utility of current SER systems in diverse real-world environments where emotional expressions are often spontaneous, subtle, and intertwined with contextual factors.

Ethical and Privacy Concerns: The integration of SER into clinical and personal applications raises substantial ethical and privacy concerns that remain inadequately addressed. Issues of data security, ensuring informed consent from individuals whose vocal data is being analysed, and mitigating algorithmic bias are recurrent but often receive insufficient attention, despite their centrality for clinical adoption (Buccoliero, 2023). The potential for misidentification of emotional states, misuse of sensitive emotional data, and the exacerbation of existing biases (e.g., gender, accent-based) underscore the urgent need for robust ethical guidelines, transparent data governance frameworks, and rigorous bias detection and mitigation strategies before widespread deployment.

Cultural and Linguistic Variability: Emotional expression through speech is highly context-dependent and varies significantly across cultures and languages. Yet, systematic cross-cultural preliminary support for SER systems remains underdeveloped. Current SER systems are disproportionately trained on English-language or Western datasets, severely constraining their global applicability (Hansen et al., 2022; Pérez-Toro et al., 2023). This linguistic and cultural bias can lead to inaccuracies when applied to non-Western populations or other languages, potentially misinterpreting emotional cues or failing to recognise culturally specific expressions. Future efforts must prioritise the development of diverse, multilingual, and multicultural datasets, along with culturally informed models, to ensure equitable and globally applicable SER technologies.

In addition to these five core issues, the review also identifies a lack of consensus regarding which psychological constructs should be prioritised, particularly the long-standing tension between categorical and dimensional models of emotion (Russell, 1980). Furthermore, there is limited exploration of longitudinal trajectories of emotional change, with most empirical studies focusing on cross-sectional snapshots rather than temporal dynamics, thus constraining insights into how vocal and affective markers evolve (Danhof-Pont et al., 2011; Faurholt-Jepsen et al., 2021).

2.4 Conceptual Model

This conceptual model offers a conceptual basis for linking vocal features, emotional regulation, and burnout-related indicators, but it does not prove a singular causal mechanism nor a validated burnout pathway. In this sense, the conceptual model offers a cautious basis for linking vocal features, emotional regulation, and burnout-related indicators without claiming a singular causal mechanism or a fully validated burnout pathway.

To address these limitations, this thesis advances a conceptual model that systematically integrates psychological and computational perspectives. The model is articulated across five interdependent layers:

- Input Layer: Acquisition of voice data in both structured and naturalistic environments, capturing variability across contexts and devices. Attention is given to balancing experimental control with ecological validity.

- Feature Extraction Layer: Derivation of acoustic and prosodic markers (e.g., MFCCs, jitter, shimmer, spectral centroid, harmonic-to-noise ratio) informed by psychological theories of emotion. This layer operationalises theoretical constructs into measurable parameters.
- Classification Layer: Deployment of ML/DL architectures to map features onto categorical (discrete emotions such as anger, sadness, joy) and dimensional (valence, arousal, dominance) models of affect. Hybrid architectures integrating convolutional, recurrent, and transformer-based networks are considered.
- Validation Layer: Triangulation of algorithmic outputs with validated psychometric tools, including the Maslach burnout, as a focal but bounded construct, Inventory (MBI), the Oldenburg Burnout Inventory (OLBI), and other context-specific instruments. Emphasis is placed on cross-validation, generalisation across populations, and integration with gold-standard psychological measures.
- Application Layer: Translation into digital health platforms that support early detection, personalised intervention, and longitudinal monitoring of emotional health. This includes integration into mobile applications, telehealth systems, and workplace wellness platforms.

This architecture is designed to safeguard both scientific rigour and clinical applicability, thereby mediating between computational precision and psychological interpretability. It emphasises iterative refinement, in which empirical findings inform theoretical frameworks and vice versa, ensuring cumulative progress.

2.5 Research Positioning

The research positioning is therefore bounded: the thesis is consistent with a promising interdisciplinary direction, but it does not establish full burnout validation, mechanism proof, or theory closure. This research positioning is consistent with a cautious interpretation of current evidence: it suggests a promising interdisciplinary direction without claiming full burnout validation, mechanism proof, or theoretical closure.

This doctoral research is strategically positioned at the dynamic intersection of psychology, occupational health, and artificial intelligence. Its overarching aim is to significantly advance Speech Emotion Recognition (SER), ensuring its development as a tool that is not only scientifically rigorous but also

profoundly clinically relevant. While a substantial portion of existing literature within this domain tends to prioritise computational novelty and technical advancements, the distinctive and crucial contribution of this research lies in its deliberate re-contextualization of SER. This re-contextualization is firmly rooted within a robust, psychologically informed framework, meticulously ensuring deep alignment with supported theories of emotion and best practices in clinical application (Costantini et al., 2022; Min et al., 2023).

The project is poised to make significant advances in the broader field of SER through several key initiatives. Firstly, it proposes and advocates for rigorous methodological standardisation and enhanced reproducibility in both data collection and feature extraction processes. This commitment to consistency is vital for building a reliable and trustworthy body of research. Secondly, it actively seeks to enhance interdisciplinary integration, fostering a more seamless and productive collaboration between the traditionally distinct fields of psychology and computer science. This bridging of disciplines is crucial for developing holistic and practical solutions to SER. Finally, and of paramount importance, the research prioritises the ethically responsible implementation of SER technologies. This critical focus includes meticulous attention to safeguarding individual privacy, ensuring truly informed consent from participants, and actively working towards comprehensive bias mitigation (Low et al., 2020). By addressing these ethical considerations proactively, the research aims to ensure that SER development is not only innovative but also equitable and protective of human well-being.

This thesis makes a substantial contribution to the rapidly evolving field of digital health by addressing fundamental gaps in current approaches and prioritising clinical validity. Its core achievement lies in the development of scalable, non-invasive, and contextually sensitive tools designed explicitly for emotional assessment. This innovation significantly advances the broader discourse surrounding digital health solutions for mental well-being.

A key focus of this research is to highlight the immense potential of voice biomarkers. These objective indicators, derived from vocal patterns, are presented as powerful aids in supporting early detection and prevention strategies for conditions like burnout and other related psychological states (Danhof-Pont, van Veen, & Zitman, 2011; Shinohara et al., 2021). This emphasis on early intervention aligns seamlessly with the overarching goals of patient empowerment and the digital transformation currently sweeping across healthcare systems worldwide. By ensuring the ethical integration of speech emotion

recognition (SER) technologies into routine clinical practice, this work aims to enhance individual engagement in their own health journeys (Graffigna & Barello, 2018; Greene & Hibbard, 2012).

Furthermore, this thesis strategically positions SER as a vital bridge, connecting theoretical psychological science with practical application. It enriches our understanding of human emotions and mental health while simultaneously addressing urgent societal needs, particularly in the domains of mental health support and workplace well-being (MacLachlan et al., 2019; Min et al., 2023). This dual orientation, encompassing both rigorous academic advancement and tangible practical applicability, is what truly defines the unique and significant contribution of the present research.

3. Methodology

The research architecture is structured but heterogeneous rather than fully unified in a single design logic.

3.1 Research philosophy

The research architecture is structured but heterogeneous rather than fully unified in a single design logic.

This doctoral research is situated within an interdisciplinary paradigm that actively integrates psychology, occupational health, and artificial intelligence. The philosophical orientation underpinning the study is pragmatic, reflecting the dual necessity of ensuring methodological rigour and maximising practical applicability in real-world settings. By adopting a sensible approach, the research acknowledges that no single paradigm or method can capture the full complexity of human emotional expression and burnout phenomena. Instead, it combines the strengths of different approaches to yield insights that are both theoretically robust and practically meaningful.

The project is simultaneously **theory-driven** and **technology-oriented**. On the theoretical side, it builds on psychological models of emotion, occupational health frameworks, and validated psychometric tools that have long been used to assess stress, burnout, and emotional regulation. On the technological side, it leverages advanced computational methods. ranging from classical signal processing techniques to

contemporary deep learning algorithms, to detect and classify emotional states from voice data. The dual orientation legitimises **methodological pluralism**, integrating quantitative analyses, computational modelling, and clinical validation to construct a comprehensive picture of the phenomenon under study.

This philosophical stance also acknowledges the limitations of traditional self-report measures, which are prone to bias and underreporting. For conditions such as burnout, which may develop insidiously and manifest subtly in daily communication, self-report alone may be insufficient. Therefore, computational tools are proposed as complementary, non-invasive methods that augment traditional diagnostic practices while maintaining psychological validity and clinical relevance.

3.2 Overall research design

The overall design should be understood as a thesis-by-publication architecture: structured at the programme level, but heterogeneous across studies. The included studies play different roles, including foundational, methodological, adjacent, and cross-domain functions. Such a design unfolds across four interconnected stages, each corresponding to a peer-reviewed publication. Taken together, they form a coherent trajectory in which each stage explicitly builds upon the preceding one, progressively extending both the methodological scope and its application.

The first study (Sartori et al., 2025) of the doctoral research adopted a **scoping review methodology** to synthesise the state of the art in Speech Emotion Recognition (SER). Following the methodological framework of Arksey and O'Malley (2005), the review mapped the existing literature on voice biomarkers and algorithms for emotional assessment, and collected and analysed 23 studies. This review revealed a lack of standardisation in approaches, with heterogeneous datasets, algorithms, and feature sets, as well as persistent conceptual ambiguities and challenges of interpretability regarding the relation between acoustic features and emotions. Moreover, it underscored the absence of a unified, psychology-based foundation in existing SER developments (Min et al., 2023; Costantini et al., 2022, as cited in Sartori, 2025). By highlighting these methodological inconsistencies and theoretical gaps, the scoping review provided the theoretical and critical foundation for empirical experimentation, ensuring that subsequent phases of the research would address concrete knowledge gaps rather than replicate existing limitations.

The second study of the doctoral research translated the insights of the scoping review into a **computational investigation** explicitly aimed at addressing ecological validity (Lentini et al., 2025). During the research period at GPI S.p.A., an application named *Talking About* was developed to collect heterogeneous datasets (voice samples, psychometric assessments, and demographic data) for training and testing SER models under real-world conditions. A key innovation was the integration of Voice Activity Detection (VAD) modules to manage background noise and environmental variability, thereby ensuring robust speech detection (Braun & Tashev, 2021). This stage directly confronted the limitations of laboratory-based paradigms by testing models in noisy, variable, and multilingual contexts, thereby enhancing generalizability. Furthermore, the design and training of machine learning models incorporated advanced deep learning architectures capable of handling spectral–temporal features, paving the way for scalable and clinically relevant applications.

The third study of the research (Tagliente e al., 2024) expanded the technical **development into a clinical setting**, with digital health tools tested as feasible adjuncts to standard care. For example, studies in the medical domain have adopted pilot and quasi-experimental designs to validate voice and multimodal biomarkers in patient populations, including Parkinson's and mood disorders, showing how AI-supported interventions can be integrated into treatment pathways (Shinohara et al., 2021; Higuchi et al., 2022; Tonn et al., 2022). Although not directly focused on burnout, such trials demonstrate that digital interventions can be ethically and effectively integrated into healthcare delivery, providing empirical evidence for their feasibility, safety, and potential to enhance patient care.

The fourth study (Shah et al., 2025) further extended the methodological framework into **translational applications** by exploring the use of SER in sensitive clinical populations. Recent studies have supported that computational voice analysis can be successfully applied beyond occupational and adult mental health contexts to paediatric and neurological settings, including work on neonatal and developmental vocal expression (Taguchi et al., 2018; Higuchi et al., 2022; Tonn et al., 2022). This contributes to the generalizability of SER methodologies, indicating their potential for early detection of neurological and psychiatric conditions, while also reinforcing the broader clinical utility of voice biomarkers as non-invasive diagnostic tools (Shinohara et al., 2021; Takano et al., 2023).

Taken collectively, these four stages establish a heterogeneous, yet unified research design oriented towards the **development, validation, and ethical integration** of voice-based emotion recognition as a

tool for burnout detection, digital care, and patient empowerment. Each stage builds upon the previous one, reinforcing a progressive logic that moves from conceptual synthesis to computational robustness, to clinical feasibility, and finally to translational expansion.

The methodological approach developed in this thesis is comprehensive and rigorous, synthesising theoretical foundations, computational innovation, empirical validation, and ethical safeguards into a coherent framework. By combining literature mapping, algorithmic development, clinical experimentation, and translational applications, the project achieves both scientific depth and practical relevance. The progression across the four publications illustrates a cumulative research trajectory that advances from conceptual synthesis to technical enhancement, to clinical feasibility, and finally to translational generalisation.

The coherence of this methodological design ensures that the contributions are not fragmented but interconnected, demonstrating originality, robustness, and relevance. Beyond advancing academic knowledge, the approach also addresses pressing societal needs in mental health and occupational well-being by proposing scalable, ethical, and contextually sensitive tools. This dual contribution—to both science and practice—underscores the value of the research and its potential to shape the future of digital healthcare and patient empowerment.

3.3 Methods overview

The methodological framework of this thesis is deliberately multifaceted, combining literature synthesis, computational modelling, empirical trials, and translational applications. Each method was carefully selected to respond to specific research questions while ensuring coherence across the overall project design.

Data Collection. Data were gathered through multiple complementary pathways. 1) First, systematic literature searches were conducted across databases such as Scopus, Web of Science, and PsycINFO to support the scoping review. 2) Second, open-source speech emotion corpora and purpose-built datasets, such as the Emozionalmente corpus, provided training and testing material for SER models. 3) Third, clinical data were collected in the Parkinson's trial, where participants underwent neuropsychological assessments alongside a digital cognitive training intervention. 4) Fourth, neonatal cry datasets were

compiled and analysed to assess methodological extensions into developmental neuroscience. Additional data were obtained through a custom application, developed in collaboration with GPI S.p.A., which enabled the systematic collection of structured voice samples alongside psychometric measures (Maslach Burnout Inventory, Oldenburg Burnout Inventory) and socio-demographic variables. This multi-modal data strategy ensured both breadth and depth.

Integration with Psychometric Measures. A defining feature of the project was the alignment of computational outputs with validated psychometric instruments. In the context of burnout, assessment relied on supported inventories such as the Maslach Burnout Inventory (MBI) and the Oldenburg Burnout Inventory (OLBI), widely recognised for their psychometric robustness in occupational health research (Danhof-Pont, van Veen, & Zitman, 2011). In parallel, studies in neurological populations employed standardised neuropsychological batteries and disease-specific scales, such as the MDS-UPDRS for patients with Parkinson's disease, to provide robust clinical benchmarks (Goetz et al., 2008). In neonatal and developmental contexts, developmental scales and expert clinical annotations served as gold standards against which computational predictions were validated (Shinohara et al., 2021; Higuchi et al., 2022). This triangulation of self-report, clinical scales, and computational measures guaranteed that findings retained both psychological validity and clinical interpretability.

Analytical and Computational Techniques. The project employed a wide range of signal processing techniques, including Mel-Frequency Cepstral Coefficients (MFCCs), spectral descriptors, formant and prosodic features, and temporal dynamics (Rejaibi et al., 2022). These features were then used as input for machine learning models, ranging from classical classifiers such as Support Vector Machines to advanced deep learning architectures, including Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and hybrid models (Costantini et al., 2022; Min et al., 2023; Muzammel et al., 2021). Model performance was validated through cross-validation protocols, cross-corpus evaluations, and transfer learning experiments using datasets like DAIC-WOZ, RAVDESS, and AVi-D, ensuring robustness against variability in speakers, environments, and linguistic contexts (Hansen et al., 2022; Lin et al., 2022). Such methodological diversity allowed for a comprehensive evaluation of model robustness, accuracy, and generalisation across heterogeneous conditions.

Statistical Analysis in Clinical Contexts. In Parkinson's clinical studies, rigorous statistical techniques were employed to evaluate the effects of interventions. For instance, when comparing demographic and

clinical subgroups, Mann–Whitney U tests and Chi-Square tests were used to account for non-normal distributions and categorical variables (Ozkanca et al., 2019; Hansen et al., 2022). In other analyses, ANOVAs were employed to examine differences in depression and vocal biomarker scores across age groups and conditions, confirming systematic variations with strong statistical significance. To ensure robustness against false positives, post-hoc corrections, such as the Benjamini–Hochberg procedure, are recommended within validation frameworks for voice-based biomarkers. These methodological safeguards guaranteed that the findings were not only statistically valid but also retained clinical interpretability, reinforcing their translational value in patient care.

Computational Validity and Explainability. Interpretability was prioritised through the adoption of explainable AI methods, most notably SHAP (Shapley Additive exPlanations), which enables the attribution of model predictions to specific acoustic features (Lundberg & Lee, 2017; Lin et al., 2022; Schultebraucks et al., 2022). By linking algorithmic outputs to psychologically interpretable constructs, transparency and trust in the models were strengthened. This methodological choice is particularly significant in sensitive domains such as healthcare, where clinical adoption depends not only on performance metrics but also on the interpretability of computational outputs (Low et al., 2020).

3.4 Ethical considerations

Ethical considerations permeated every stage of the project, reflecting a commitment to safeguarding participants and ensuring socially responsible research outcomes. All studies involving human subjects adhered to the principles of the Declaration of Helsinki, with protocols reviewed and approved by institutional ethics committees (e.g., IRB clearance for the Parkinson's trial). Participants were fully informed about the scope of the research and signed detailed consent forms clarifying data use, storage procedures, and their right to withdraw at any time without consequences.

Data management practices were designed to **ensure strict compliance with the General Data Protection Regulation (GDPR)**. Anonymisation procedures were applied systematically, encryption safeguarded both storage and transmission, and access was restricted to authorised personnel only. Regular audits and data security checks were conducted to minimise the risk of breaches. Special emphasis was placed on ensuring that sensitive voice data—capable of revealing not only content but also emotional state—was treated with the highest level of confidentiality (Arksey & O'Malley, 2005;

Braun & Tashev, 2021; Hansen et al., 2022; Lin et al., 2022). Data management practices were meticulously designed and implemented to ensure strict compliance with the General Data Protection Regulation (GDPR), a cornerstone of modern data privacy legislation. A multi-faceted approach was employed, integrating several key protective measures. Anonymisation procedures were applied systematically at the earliest possible stage of data processing, rendering individual subjects unidentifiable while preserving the analytical utility of the data. Furthermore, robust encryption protocols safeguarded all data, both at rest in storage and during transmission across networks, preventing unauthorised interception or access.

Access to this sensitive data was rigorously restricted to only authorised personnel, a small, carefully vetted group with a legitimate need to access the information for research purposes. This principle of least privilege was strictly enforced. To maintain this high level of security, regular audits and comprehensive data security checks were conducted regularly. These proactive measures aimed to identify and rectify any potential vulnerabilities, thereby minimising the risk of data breaches and ensuring ongoing compliance with evolving security standards.

A particular emphasis was placed on the handling of sensitive voice data. This type of data presents unique challenges, as it has the capacity to reveal not only the explicit content of communication but also subtle cues about emotional states, health conditions, and even identity through vocal characteristics. Consequently, this sensitive voice data was treated with the highest possible level of confidentiality and protection, employing specialised safeguards beyond general data management protocols. This approach aligns with best practices in research ethics and data privacy, drawing upon supported methodologies (Arksey & O'Malley, 2005; Braun & Tashev, 2021; Hansen et al., 2022; Lin et al., 2022) to ensure both research integrity and the utmost respect for participant privacy.

Another central ethical concern was mitigating **algorithmic bias**. Datasets were curated to ensure diversity in terms of language, demographic characteristics, and clinical profiles, thereby reducing the risk of models reproducing discriminatory outcomes. Bias detection and correction techniques were integrated into the modelling process, and cross-linguistic evaluations were conducted to verify fairness across populations. By explicitly addressing these concerns, the research aimed to promote equitable applications of SER technologies (Costantini et al., 2022; Hansen et al., 2022; Min et al., 2023).

Notably, the project emphasises that SER should be deployed only as a complementary and **non-invasive tool to support professional clinical judgment**, never as a substitute. Its role is to enhance early detection, facilitate monitoring, and provide additional layers of evidence to clinicians, rather than automating decision-making in isolation. Transparency in dissemination was maintained throughout the research process, with explicit acknowledgement of limitations, risks, and future challenges.

4. Publication 1

The initial research effort culminates in a comprehensive publication titled "A scoping review on the use of voice biomarkers for emotional assessment." This publication serves as a foundational exploration into the burgeoning field of utilising vocal cues as objective indicators of emotional states. It systematically maps the existing literature, identifying key methodologies, prominent findings, and emerging trends in the application of voice-based technologies for discerning human emotions. The review highlights the diverse range of voice features, including pitch, prosody, amplitude, and speaking rate, that researchers are investigating as potential biomarkers for various emotions, such as joy, sadness, anger, and fear. Furthermore, it delves into the different analytical techniques employed, from traditional signal processing to advanced machine learning algorithms, in extracting and interpreting these vocal markers. The scope of the review extends to examining the potential applications of such advancements across various domains, including mental health monitoring, human-computer interaction, and personalised emotional feedback systems. By synthesising the current knowledge base, this publication aims to provide a critical overview of the opportunities and challenges associated with leveraging voice for emotional assessment, thereby setting the stage for future research and development in this promising area.

Paper Title: A scoping review on the use of voice biomarkers for emotional assessment
Status: Published https://doi.org/10.1027/1015-5759/a000915
Journal: European Journal of Psychological Assessment
My Contribution: 50% - Conceptualisation, Methodology, Validation, Formal Analysis, Investigation, Writing (Original Draft, Review & Editing).

A scoping review on the use of voice biomarkers for emotional assessment

Abstract

Speech Emotion Recognition (SER) is an umbrella term that encompasses all Machine Learning & Deep Learning (ML/DL) algorithms used for the specific task of extracting emotional states from human speech. In the literature, various techniques have been utilised to extract emotions from signals, including well-supported speech analysis and classification techniques. Using the scoping review method, the

paper maps techniques for speech-based emotion recognition. In doing so, it presents elements defining the use of algorithms to assess emotions, e.g., databases used for emotion recognition, notions and types of emotions considered, and empirical investigations made toward SER and related limitations. The contribution places particular emphasis on existing perspectives and practices to offer a series of recommendations for future developments.

Keywords: emotion, machine learning, deep learning, speech-emotion recognition, scoping review.

Recently, advances in digital technology and Artificial Intelligence (AI) have enabled the acquisition of a large amount of voice recordings, coupled with a more detailed analysis of speech (Pfister & Robinson, 2010). The term Artificial Intelligence, coined by John McCarthy in 1955, denotes the science of building *intelligent machines*, that is, machines capable of performing actions that fall within the scope of human intelligence. Machine Learning (ML) is a subset of AI that applies statistical methods to iteratively learn patterns and relationships from data without explicitly being programmed. In other words, ML algorithms learn to make decisions or predictions based on the training data set provided as input. Deep Learning (DL) is a branch of Machine Learning that aims to mimic the functioning of the human brain through the development of artificial neural networks to solve complex problems. Artificial neural networks consist of a set of interconnected neurons, distinguished into several layers (i.e., an input layer, an output layer and one or more hidden layers) (Aggarwal et al., 2022). ML and DL methods have been used to assess a person's emotional state automatically.

Emotions are expressed through speech, facial expressions, and other non-verbal clues (Pfister & Robinson, 2010). Speech contains two main insights into what has been spoken (i.e., lexical content) and how it is spoken (i.e., emotional component), which can reveal the mental state of an individual (Madanian et al., 2023). Accordingly, advances in digital technology have realised the so-called Speech Emotion Recognition (SER), a process of determining speakers' emotional state from acoustic speech features, irrespective of the semantic content. It refers to a sub-discipline of the interdisciplinary field of affective computing, which combines evidence-based knowledge from computer science, psychology, and cognitive science (Tao & Tan, 2005). SER algorithms encompass speech signal extraction, acoustic feature extraction, feature selection, and classification.

SER algorithms are *capable* of learning the relationship between input (i.e., acoustic features) and output

(i.e., emotion or mood disorder). This relationship is understood in a supervised manner when the training data used to feed the algorithm has been previously labelled; this is called Supervised Learning. Alternatively, it is Unsupervised Learning, in which case the learning system must detect existing patterns without pre-existing labels.

On the surface, the central hypothesis is that a set of objectively measurable parameters in voice reflects the person's emotional state (Costantini et al., 2022). SER algorithms are then developed and trained in various ways to assess such parameters via quantitative features, which can include emotional aspects such as pitch, intensity, or represent specific coefficients of emotion (e.g., the Mel-Frequency Cepstral Coefficients, MFCCs) (Taguchi et al., 2018). For example, speech patterns analysed via SER are used as indicators of mental state (e.g., focusing on lower pitch, monotonous speech, lower sound intensity, and lower speech rate can indicate the presence of depressive symptoms, Low et al., 2020). On a global level, we are witnessing an increase in attention on SER as a pioneering artificial intelligence agent for detecting mood disorders through voice analysis.

For instance, Madanian and colleagues (2022) reported a preliminary analysis demonstrating the feasibility of automatic vocal emotion recognition for mental health purposes. Vocal signals were labelled with emotions that are relevant to mental health. In addition, the acoustic features extracted from vocal signals and used to feed ML algorithms are related to the symptoms of mental disorders (e.g., MFCCs) as reported in the literature, specifically in the field of computer science and mental health. Other studies have shown that automatic emotion recognition, specifically that which relies on speech, can support the assessment of mood disorders. Harati and colleagues (2018) utilised the relationship between short-term emotions and long-term depressive states to develop predictive models incorporating emotion-based features.

Despite the ever-expanding literature on the topic, there is a dearth of psychological knowledge on how SER works and whether they are effective in providing indications of an individual's emotions. While we witness an increase in interest, the literature lacks a common ground for SER developments, with scholars and digital engineers developing different and separate metrics that encompass distinct patterns for emotional recognition (Costantini et al., 2022). Moreover, SERs are evaluated and implemented in experimental settings that differ from one study to another, with various methodological approaches, which leave the literature without a unique framework for interpreting the existing perspectives (Min et

al., 2023). On the surface, the increasing development of SER algorithms requires psychology to evaluate such developments and provide substantial indications on how scholars and practitioners can develop and use these algorithms in pragmatic, ethical, and legal terms, while also ensuring the validity and reliability of these algorithms.

In this article, we aim to address the current state of the literature by presenting the findings of a scoping review on voice biomarkers and SER, which assesses emotion state. Mapping the existing perspectives on the use of voice as a proxy for accessing emotions via biomarkers and artificial intelligence agents can represent an initial step for psychology to approach the exponential technological evolution. We do so by addressing a general question on what the evidence-based knowledge on the use of voice biomarkers is to detect emotion. Following the above-described issues in SER, this question can be broken down into four sub-questions:

(i) What are the current perspectives on the use of voice biomarkers for emotion recognition?
(ii) What is the empirical evidence-based knowledge on the assessment?
(iii) Which are the existing algorithms for speech emotion recognition?
(iv) Which ethical concerns appear to be relevant in the use of voice biomarkers?

In the following sections, we begin by introducing introduce the methods used to search and extract data related to our research questions. We continue by presenting our results following each sub-question separately. Finally, we discuss our synthesis of the literature and provide indications for future research and practical perspectives on the use and development of SER. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Materials and methods

Based on the methodological indications by Arksey and O'Malley (2005), our study used the scoping review approach to synthesise the existing literature in the field of SER, in order to understand the use of the human voice as a proxy to assess emotional state, and correspondingly to identify voice biomarkers. The purpose of a scoping review is to provide an overview of the available research evidence by mapping the known in a particular area of interest.

Pragmatically, scoping reviews are based on a systematic approach for data collection, but inclusion and exclusion criteria are more flexibility, and the synthesis is based on the research questions rather than methodological rigor. As such, a scoping review follows five main steps: namely a) identify the research question, b) identify relevant studies and select them, c) chart the data (data extraction process), d) collate, summarize, and report the results, e) consult stakeholders, to validate findings from the scoping review. Considering that the topic of this review (i.e., vocal biomarkers and algorithms to assess emotional state) is relatively new and under-investigated, the scoping review is an appropriate methodology for addressing our broad research questions.

Based on the research questions, the keywords used in the search string were defined considering terms such as “Speech”, “Speech Emotion Recognition”, “Emotion”, “Voice”, “Biomarkers” “Artificial Intelligence Agents” “SER” using Boolean operators. The search string has been used in specific fields of psychology, neuroscience, cognitive sciences, digital engineering for papers in English without any limitations for the type of manuscript (e.g., research article, review). The search phase has been conducted using Scopus, Web of Science (WoS) and PsychInfo. No restrictions were set for search-time range and study type. After removing duplicates, we continued with title and abstract screening using the following inclusion criteria: a) studies that used voice to assess emotion; b) studies that deployed SER algorithms to recognize emotion from voice; c) studies that defined voice biomarkers to identify emotions. Items collected were read and analyzed for full-text evaluation prior synthesis. In doing this, we extracted evidence from the selected studies, through a structured table, built ad hoc for this study. It reports, for each study, the authors and year, the country where the study was conducted, the number of participants, the type of study, the main objectives, the main results, the acoustic features (voice biomarkers), the speech-eliciting task, the language, the scales, the pre-existing datasets, the software, the study design, and the AI algorithms. Although the meaning of the terms ‘voice’ and ‘speech’ is different, in the reviewed articles the authors present these algorithms often using the two terms interchangeably and focusing on the features of voice.

Figure 2 shows the process and the results of our search analysis. The search yielded an initial $N = 322$. Of these, $n = 242$ were excluded after title screening. Then, the abstracts of the remaining $n = 80$ items were analyzed, and $n = 53$ were excluded as they did not meet the inclusion criteria. Finally, the full texts of the remaining $n = 27$ items were studied, and four were excluded. The final sample consisted in $n = 23$ studies, included in our review for further analysis.

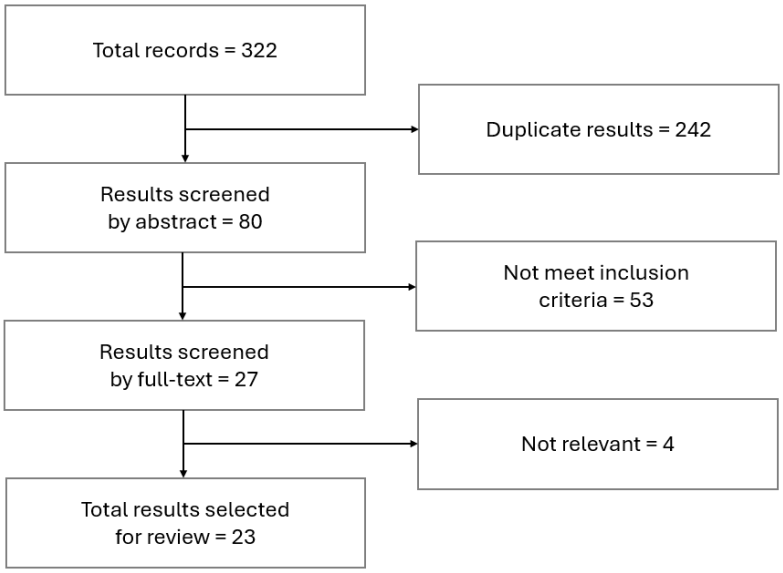

Figure 2 - Process and results of the data collection

Results

Overview of the collected data

Two studies (Giddens et al., 2013; Low et al., 2020) were literature reviews and one of them (Low et al., 2020) used a systematic review approach, while the rest were empirical studies. Among them $n = 11$ were cross-sectional studies, $n = 7$ were quasi-experiment studies. The remaining three adopted more complex approaches with only $n = 1$ study (Min et al., 2023) using a mixed method approach combining cross-sectional and a longitudinal approach. $N = 1$ paper (Faurholt-Jepsen et al., 2021) reporting a quasi-experimental combined with a longitudinal approach and $n = 1$ study (Pérez-Toro et al., 2023) comprising two cross-sectional studies and one quasi-experiment study. The procedures of the empirical studies complied with the Declaration of Helsinki or were approved by the ethics committee. In addition, the informed consent was signed by participants in 16 of the collected studies.

Five studies took place in the United States (Galatzer-Levy et al., 2021; Giddens et al., 2013; Low et al., 2020; Schultebraucks et al., 2022; Zhang et al., 2020), four in Japan (Higuchi et al., 2022; Taguchi et al., 2018; Takano et al., 2023; Shinohara et al., 2021), three in China (Di et al., 2021; Lin et al., 2022;

Miao et al., 2022), three in Germany (Pérez-Toro et al., 2022; Perez-Toro et al., 2023; Tonn et al., 2022), two in France (Muzammel et al., 2021; Rejaibi et al., 2022), two in Denmark (Faurholt-Jepsen et al., 2021; Hansen et al., 2022), one in Czech Republic (Kouba et al., 2023), one in Israel (Wasserzug et al., 2023), one in South Korea (Min et al., 2023), one in Turkey (Ozkanca et al., 2019). Table 1 summarizes the main features of the studies analyzed.

Table 1 - overview of the collected items.

Author(s), year	Country	Participants	Type of study	Main Goals	Main Results
Faurholt-Jepsen et al., 2021	Denmark	180 participants: 121 patients with diagnosis of Bipolar Disorder, 21 unaffected first-degree relatives, and 38 healthy controls.	Empirical Study	Investigate whether voice features could discriminate between: 1. BD, unaffected first-degree relatives (UR) and healthy control individuals (HC); 2. affective states within BD.	It was shown that voice features can discriminate BD from HC with high sensitivity (0.79), but with low specificity (0.54), and that voice features significantly can differentiate between UR and HC. Within patients with BD, mania was rather specifically discriminated from euthymia.
Galatzer-Levy et al., 2021	United States	20 patients (suicide attempt).	Empirical Study	Examine measurements extracted from video interviews to quantify facial, vocal, and movement behaviors in relation to suicide risk severity.	Suicide severity was associated with multiple visual and auditory markers including speech prevalence, overall expressivity, and head movement measured as head pitch variability and head yaw variability.
Giddens et al., 2013	United States		Review	Study the effects of various forms of stress upon the healthy voice.	The results indicate that stress can have a significant impact on voice characteristics and this impact can vary based on individual factors such as personality type and gender. The nature of the stressor as well as the context in which the stress occurs must be taken under consideration.
Hansen et al., 2022	Denmark	40 patients with major depressive disorder, 42 healthy controls, 25 patients in remission (to validate and test the model).	Empirical Study	Evaluate depression and remission from voice using transfer learning.	Speech Emotion Recognition model was able to accurately discriminate between healthy controls and depressed patients (obtaining an Area Under the Curve - AUC of 0.71).
Higuchi et al., 2022	Japan	102 patients with depression and 129 healthy adults.	Empirical Study	Detect depression using a composite index (MDDI) of vocal acoustic properties.	The proposed index (MDDI) distinguished between depressed and normal subjects with ~90% accuracy in the training set, and ~80% accuracy in the test set.
Kouba et al., 2023	Czech Republic	10 licensed air traffic controllers.	Empirical Study	Identify and monitor air traffic controllers' fatigue levels based on voice analysis.	Voice analysis is able to identify differences in wakefulness and fatigue, as its results correlate with changes in brain activity.
Lin et al., 2022	China	57 patients with major depressive disorder and 47 healthy individuals (to evaluate the model performance).	Empirical Study	Use the biological information of speech, combined with deep learning to build a binary-classification model of depression in the elderly.	Vocal biomarkers extracted from raw speech signals have high potential for the early diagnosis of depression in older adults. The initial sensitivity and specificity of the DL model were respectively 82.14% and 80.85%.

Low et al., 2020	United States		Systematic Review	Study the use of speech for automated assessments across a broader range of psychiatric disorders.	The majority of studies focus on MDD, PTSD, and bipolar disorder; certain disorders have been less studied such as eating and anxiety disorders. 63% of studies-built machine learning predictive models, and the remaining 37% performed null-hypothesis testing only. 32% of studies used AVEC data sets. Models' performance is a function of sample size, dataset's preprocessing, feature selection, and model used.
Miao et al., 2022	China	DAIC-WOZ dataset: 189 participants.	Empirical Study	Identify depression using machine learning and deep learning models.	The study found that the CNN yielded the best classification by using fused features in the DAIC-WOZ dataset.
Min et al., 2023	South Korea	104 patients with bipolar disorder or major depressive disorder.	Empirical Study	Assess suicidality based on acoustic voice features of psychiatric patients using artificial intelligence.	The within-person classification model outperformed the between-person classification model. The between-person classifier was able to detect high suicidality with 69% accuracy, whereas the within-person model was able to predict worsening suicidality over 2 months with 79% accuracy.
Muzammel et al., 2021	France	DAIC-WOZ dataset: 189 participants, but only 182 participants have been used: 30% depressed (major depressive disorder) - 70% non-depressed.	Empirical Study	Analysis of different deep learning architectures for depression recognition.	LSTM based models performed better as compared to CNN based models. The fusion of deep audio features with visual features leads to better performance comparing to their fusion with word embedding textual features. The best outperforming model is based on the model-level fusion using an LSTM network of the deep audio and the visual features (accuracy of 77.16%).
Ozkanca et al., 2019	Turkey - United States	921 Parkinson's patients: 603 nondepressed, 318 depressed.	Empirical Study	Classify depressed and nondepressed subjects using their voice features and PD severity.	The models achieved accuracies as high as 0.77 in classifying depressed and nondepressed subjects accurately using their voice features and PD severity. Significant correlation between PD severity and depression presence was found.
Pérez-Toro et al., 2022	Germany	50 healthy control subjects, 25 Depressive-Parkinson's Disease patients and 35 Non Depressive-Parkinson's Disease patients.	Empirical Study	Classify depressed and non-depressed Parkinson's patients using a combination of speech analysis and natural language processing methods.	The automatic classification of depressed and non-depressed Parkinson's patients showed F-scores of up to 0.77, this suggests that speech and language information are directly associated to the depression state of PD patients.
Pérez-Toro et al., 2023	Germany	IEMOCAP: ten speakers; Customer service in banking call-centers: 1285 male and 1078 female; Depression in PD: 25 depressive PD patients; 35 Non-depressive PD patients; ADReSS Challenge Dataset (AD): AD patients and healthy controls.	Empirical Study	Evaluate scenarios such as customer satisfaction and assessment of patients with neurodegenerative diseases, using deep learning techniques and the Arousal-Valence plane.	The classification of depression in PD considering each model separately obtained the highest results using the arousal for acoustics and the valence for linguistics. The early fusion of the arousal and valence models improved the results. The classification of AD patients is not directly linked to emotional, mood, or affective labels. The classification of AD produces higher results using the arousal information for acoustics and the valence information for linguistics.

Rejaibi et al., 2022	France	3 dataset: DAIC-WOZ: 189 participants, but only 182 participants have been used; RAVDESS: 24 actors; AVi-D: 292 participants.	Empirical Study	Detect depression and predict its severity level from speech.	The proposed approach (MFCC-based RNN) outperforms the state-of-art approaches on the DAIC-WOZ database with an accuracy of 76.27%. The performances of the proposed approach are evaluated under multi-modal and multi-features experiments. MFCC based high-level features hold relevant information related to depression. Adding visual action units and different other acoustic features further boosts the classification results by 20% and 10% to reach an accuracy of 95.6% and 86%, respectively.
Schultebraucks et al., 2022	United States	81 trauma survivors.	Empirical Study	Classify major depressive disorder (MDD) and posttraumatic stress disorder (PTSD) using machine learning-based computer vision, semantic, and acoustic analysis.	It's possible to identify digital markers that can be utilized with DL models to classify MDD and PTSD status. The algorithm discriminates PTSD status with an AUC of 0.90 as well as depression status with an AUC of 0.86.
Shinohara et al., 2021	Japan	14 healthy individuals and 30 patients with major depression.	Empirical Study	Assess the psychological issues of individuals with major depressive disorders using emotional components contained in their voices. Define and evaluate two indices: vitality and mental activity (derived from emotions).	A significant negative correlation existed between the vitality extracted from the voices and HAM-D scores ($r=-0.33$, $p<0.05$). Mental activity was not validated because continuous data could not be collected sufficiently for the same participants in both the healthy and the patient group.
Taguchi et al., 2018	Japan	36 patients with major depressive disorder and 36 healthy controls.	Empirical Study	Discriminate between depressive patients and healthy controls using vocal acoustic features.	The second dimension of MFCC was significantly different between groups and allowed the discrimination between patients and controls with a sensitivity of 77.8% and a specificity of 86.1%.
Takano et al., 2023	Japan	110 subjects with depression and bipolar disorder.	Empirical Study	Classify patients in different symptom groups based on acoustic features of their speech.	The study shows that it is possible to separate different symptom groups with an accuracy of 79%. The results suggest that voice from speech can estimate the symptoms associated with depression.
Tonn et al., 2022	Germany	163 participants affected by different degrees of depression.	Empirical Study	Evaluate the measurements of the presence or absence of depressive mood in participants by comparing the analysis of speech parameters with the results of the PHQ-9.	The study shows that there is a high correlation between the depression severity classification as measured by the PHQ-9 and the depressive vocal scores as measured by the VoiceSense analysis system ($r=0.41$; $p<.001$).
Wasserzug et al., 2023	Israel	40 patients with major depressive disorder and 104 healthy controls.	Empirical Study	Design and validate a prototype of automatic speech analysis based on algorithms for classifying the speech features related to MDD.	The findings revealed robust significant differences between the vocal depression scores of MDD patients and the equivalent vocal depression scores of non-clinical participants as well as significant differences between patients in the acute phase and in remission.

Zhang et al., 2020	United States	222 participants.	Empirical Study	Explore and validate features extracted from recorded voice samples of depressed subjects as digital biomarkers for suicidality, psychomotor disturbance and depression severity.	The findings show that: 1. prosody and acoustic features were not strongly predictive, with respect to the presence of psychomotor retardation; 2. all voice features had predictive power, with respect to the presence of suicidal ideation; 3. all voice features had relatively similar predictive power, with respect to predicting depression severity. Voice features extracted from audio of depressed subjects were able to predict PHQ9 question 9 and total scores with an area under the curve of 0.85.

In most of the studies, the same language was used to train and test the AI algorithms deployed. In detail, the most widely used languages were Japanese, English, and German. In four studies, different languages were used in the two phases. Hansen and colleagues (2022) performed a cross-lingual study, they used different languages to train and validate the AI model, i.e., English and German were used to train the model to predict emotion, and Danish was used to validate the model in the detection of depression. This allows us to generalise the model to new participants and languages. This is important for clinical implementations, where models need to handle a variety of languages, dialects, and accents. Schultebraucks and colleagues (2022) collected vocal data from patients who were fluent in English, Spanish or Mandarin; and they used this dataset to train and test the model. Rejaibi and colleagues (2022) used an English dataset to train and test the model, and a German dataset to generalise the model (i.e., to test the performance of the model on other datasets). Lastly, Pérez-Toro and colleagues (2023) used different languages; English dataset to train the model and Spanish and English datasets to test the model. The proposed approach improved the accuracy of the models, suggesting that acoustic information can be utilised across different languages.

Pre-existing datasets

The majority of the empirical studies analyzed performed data collection during the studies, defining the inclusion and exclusion criteria and selecting participants. Seven studies used pre-existing datasets to train the AI models and evaluate their performance. Three of them (Miao et al., 2022; Muzammel et al., 2021; Rejaibi et al., 2022) used the Distress Analysis Interview Corpus Wizard-of-Oz dataset (DAIC-WOZ, Gratch et al., 2014). It was introduced in the Audio/Visual Emotion Challenge and Workshop in 2017 (AVEC 2017). It is composed of clinical interviews aiming to investigate different psychological distress conditions such as depression, anxiety, and post-traumatic stress disorder. The dataset is available upon request. Rejaibi and colleagues (2022), in addition to the DAIC-WOZ dataset, also used

AVi-D corpus (Valstar et al., 2013; Valstar et al., 2014), and RAVDESS dataset (Livingstone & Russo, 2018). AVi-D corpus was introduced during the AVEC 2014 challenge, and it is a depression database. The RAVDESS dataset contains actors reciting phrases with different emotions. The RAVDESS dataset and the AVi-D corpus are public datasets. The RAVDESS dataset was also used by Hansen and colleagues (2022), along with two other similar datasets: CREMA-D (Cao et al., 2014), and EMO-DB (Burkhardt et al., 2005) datasets. The datasets are publicly available. Lin and colleagues (2022) used the Oizys dataset to train the model; it contains 1589 unique speakers, who are asked to answer questions. The dataset is proprietary due to privacy issues associated with health data. Ozkanca and colleagues (2019) used the mPower Voice Dataset (Bot et al., 2016) in the phases of training and testing of the model; it contains 921 unique subjects who provided an answer to the third question of the Parkinson's Disease Questionnaire (PDQ-8). The dataset is available upon request. Pérez-Toro and colleagues (2023) used four public datasets: 1. The Interactive Emotional Dyadic Motion Capture dataset (IEMOCAP, Busso et al., 2008), that consists of dyadic sessions where actors perform improvisations or scripted scenarios, specifically selected to elicit emotional expressions, is used to discriminate different quadrants in the arousal-valence plane. 2. Customer service in banking call-centers: it consists of recordings from banking call-centers, that were annotated according to two different emotion labels: positive (satisfied) and negative (dissatisfied). 3. Depression in PD: it contains spontaneous speech from 60 speakers, 25 Depressive PD patients and 35 Non-Depressive PD patients. 4. ADReSS Challenge Dataset (Luz et al., 2020) was created for the Alzheimer's speech classification task in the Interspeech ADReSS challenge 2020, the recordings consisted of spontaneous speech from participants with and without a diagnosis of Alzheimer's Dementia. The IEMOCAP dataset and the ADReSS Challenge Dataset are available upon request.

Scales

Assessment scales were used in most of the studies analyzed, in order to label the voice data or to collect more information about the study participants. Depending on the objectives of the study, different assessment scales were implemented. Notably, most studies that have looked at depression (i.e., presence of depression or severity of depression) have used the Hamilton Rating Scale for Depression (HAMD, Hamilton, 1960) and the Patient Health Questionnaire-9 (PHQ-9, Spitzer et al., 1999). The studies that analyzed suicide risk and suicidality used the Beck Scale for Suicide Ideation questionnaire (Beck et al., 1979; Kliem et al., 2017). The MDS Unified Parkinson's Disease Rating Scale (MDS-UPDRS, Goetz et

al., 2008) was used in the studies that focused on depression in Parkinson's patients. Table 2 shows in detail the scales used and the reason why they were used.

Table 2 - overview of the features, tasks and characteristics of the studies using SER.2

Reference	Features	Task	Language	Scales	Pre-existing datasets	Software	Study design	AI algorithms
Hansen et al., 2022	13 MFCCs.	Sentences spoken by actors to train the model. Interviews to test the model.	English and German for training; Danish for validation and test.	17-item Hamilton Rating Scale for Depression, used to define the severity of depression.	CREMA-D, RAVDESS and EMO-DB datasets to train the model.	N.A.	Quasi-experiment study.	Speech Emotion Recognition (SER) model was used to predict depression (Transfer learning).
Lin et al., 2022	Log-Mel spectrogram and latent features.	Reading a fixed text.	Mandarin.	Self-Rating Depression Scale and Hamilton Rating Scale for Depression (completed before the acoustic data collection).	Oizys: to train the model. It contains 1589 unique speakers. Answer questions.	N.A.	Quasi-experiment study.	Two different models were used: 1. DepAudioNet: audio-based method on depression classification (baseline model); 2. Depression Classifier Model.
Min et al., 2023	Features for the between-person classification model: sex, age, past suicide attempts; root-mean-square energy, F1, F0 bandwidth, F2 bandwidth, mean pitch, magnitude error, MFCC1, MFCC5, MFCC15, MFCC22, MFCC24, and MFCC27. Features for the within-person classification model: sex, age, past suicide attempts; root-mean-square energy, MFCC1, MFCC8, MFCC9, MFCC19, MFCC20, MFCC28, MFCC29, MFCC31, MFCC33, MFCC34, MFCC35, and MFCC36.	Clinical interviews.	N.A.	Suicidality was assessed using the Beck Scale for Suicidal Ideation and suicidal behavior using the Columbia Suicide Severity Rating Scale. Other scales used in the study are: Hamilton Depression Rating Scale, Patient Health Questionnaire-9, Beck Anxiety Inventory and Barratt Impulsiveness Scale.	N.A.	N.A.	Cross-sectional study. Longitudinal study.	Two different models were used: 1. Single-layer artificial neural network for the between-person classification model. 2. Extreme gradient boosting machine learning algorithm for the within-person model.
Zhang et al., 2020	Acoustic (derived from eGeMAPS), prosodic and linguistic features.	Repeat one phrase and 30-sec of free speech.	English.	Patient Health Questionnaire-9 (PHQ-9).	N.A.	openSMILE (for acoustic features derived from eGeMAPS).	Cross-sectional study.	Two different models were used: 1. Gradient boosting classifier to predict depression symptomatology; 2. elasticNet linear

								regression model to predict depression severity.
Schulte braucks et al., 2022	Voice prosody, facial, speech content, and movement features (digital biomarkers).	Qualitative interviews.	English, Spanish or Mandarin.	PTSD status was evaluated using the PTSD Checklist for DSM-5 (PCL-5). Depression severity was evaluated using the Center for Epidemiologic Studies of Depression Scale (CES-D).	N.A.	PRAAT Software python library Parselmouth.	Cross-sectional study.	Deep learning networks.
Pérez-Toro et al., 2022	Spectral, prosody, and linguistic features.	Spontaneous speech.	Spanish.	Depression was labeled according to one of the items included in the first part of the MDS-UPDRS scale.	N.A.	N.A.	Cross-sectional study.	Gaussian Mixture Model-Universal Background Model (GMM-UBM) and Support Vector Machine (SVM) classifiers to find differences between depressed and non-depressed Parkinson's patients.
Ozkanca et al., 2019	Two sets of features: the first was derived from AVEC2013; the second was derived from GeMAPS.	Make the single phoneme sound /'a/.		3rd question of the Parkinson's Disease Questionnaire (PDQ-8) to define depression; PD severity was defined using MDS-UPDRS survey.	mPower Voice Dataset.	MATLAB to remove silence; openSMILE to extract features; MRMR algorithm to filter AVEC2013 features.	Cross-sectional study.	Three different types of algorithms were tested to examine depression prediction: Support Vector Machine; Random Forest; Deep Neural Network.
Shinohara et al., 2021	N.A.	Conversation with physicians (patient group); Read 14 type of fixed phrases (healthy people); Read fixed sentences (patient and healthy group)	Japanese.	Patients' depression severity was evaluated using the Hamilton Rating Scale for Depression.	N.A.	Software ST (sensibility technology) used to extract emotions.	Quasi-experiment study.	N.A.
Kouba et al., 2023	Average Voice Energy, Stress %, Upset %, Energy Level, Uneasy Level, Stress Level, Thinking Level, Confidence, Concentration and Anticipation.	Voice was recorded during the simulations.		Karolinska Sleepiness Scale (to define subjective feelings of drowsiness).	N.A.	Layered Voice Analysis (LVA) software.	Cross-sectional study.	N.A.

Higuchi et al., 2022	Acoustic quantities, temporal statistics and statistical functionals (The large openSMILE emotion feature set).	Read 10 set phrases.	Japanese.	The severity of depression was assessed using the Hamilton Depression Rating Scale.	N.A.	openSMILE (to extract features); PCA algorithms (to reduce dimensionality).	Quasi-experiment study.	Logistic regression to classify subjects (i.e., depressed patients and healthy control).
Wasserzug et al., 2023	200 acoustic features: lengths, ranges, slopes, frequencies, values and shapes of pitch, amplitudes and silences.	Participants' mobile phone conversations (45 sec).	Hebrew.	Severity of MDD was assessed on the Hamilton Depression Rating Scale and Clinical Global Impression Scale.	N.A.	VoiceSense.	Quasi-experiment study.	N.A.
Tonn et al., 2022	200 raw voice parameters: lengths, ranges, slopes, frequencies, values, and shapes of pitch-extracted parameters, amplitude-extracted parameters, and silence-extracted parameters.	Answer vocally to a question (e.g. Please say in a few sentences on your social life).	German.	Patient Health Questionnaire-9 (PHQ-9).	N.A.	VoiceSense.	Cross-sectional study.	N.A.
Muzammel et al., 2021	Audio features (MFCCs), visual features (AU) and textual features (Word2Vec).	DAIC-WOZ: clinical interviews.	English.	Participants' data are labeled in terms of depression severity level and a binary depression label (yes or not) using the Patient Health Questionnaire-8 (PHQ-8).	DAIC-WOZ.	N.A.	Cross-sectional study.	Unimodal and multimodal representation for depression recognition: MFCC-CNN (Convolutional Neural Network); MFCC-LSTM (Long-Short Term Memory).
Takano et al., 2023	Glottal flow (NAQ, QOQ), frequency (FFT, formant, MFCC), prosody (jitter, shimmer, fundamental frequency, HNR, energy RMS).	Read fixed phrases.	Japanese.	GRID-HAM-D17 was used to label depressive symptoms.	N.A.	DisVoice (glottal flow), praat (prosody), openSMILE (frequency and prosody).	Cross-sectional study.	K-means method to cluster the subjects into two groups; Decision tree learning to predict the subjects' symptom classes.
Miao et al., 2022	BSF (bispectral features) and BCF (bicoherent features), frequency spectrum, prosodic features, speech quality.	DAIC-WOZ: Clinical interviews.	English.	Patient Health Questionnaire-8 (PHQ-8).	DAIC-WOZ.	COVAREP (to extract frequency spectrum, prosodic features, speech quality). Higher-Order Spectral Analysis (HOSA) (to extract BSF, BCF).	Cross-sectional study.	Three different models were tested: Support Vector Machine; K-Nearest Neighbors; Convolutional Neural Network.

Taguchi et al., 2018	Mean values of the following acoustic features: root mean square of energy, twelve dimensions of MFCC, zero crossing rate, harmonics to noise ratio, fundamental frequency (Feature set of Interspeech 2009 Emotion Challenge).	Read ten digits "012-345-6789" before and after a verbal fluency task.	Japanese.	Quick Inventory of Depressive Symptomatology - Self-Report.	N.A.	openSMILE.	Quasi-experiment study.	N.A.
Rejaibi et al., 2022	MFCCs (visual action units were used in multi-modal experiments).	DAIC-WOZ: clinical interviews; RAVDESS: recite sentences; AVi-D: reading task and answering questions.	English (DAIC-WOZ and RAVDESS); German (AVi-D).	Patient Health Questionnaire-8,PHQ-8 (DAIC-WOZ dataset); Beck Depression Inventory-II, BDI-II (AVi-D dataset)	DAIC-WOZ (to train the model); RAVDESS (transfer learning); AVi-D (to evaluate the model's performance).	N.A.	Cross-sectional study.	MFCC-based RNN (Recurrent Neural Network) to detect depression or to predict the depression severity level.
Pérez-Toro et al., 2023	2D-Mel spectrogram, linguistic features.	IEMOCAP: interactions between speakers; Customer service in banking call-centers: customer opinions; Depression in PD: spontaneous speech; ADReSS Challenge Dataset (AD): spontaneous speech.	English; Spanish.	N.A.	IEMOCAP (to discriminate different quadrants in the arousal-valence plane); Customer service in banking call-centers; Depression in PD; ADReSS Challenge Dataset (AD).	N.A.	Cross-sectional study. Quasi-experiment study.	CNN + Bi-GRU (Convolutional Neural Network + Bidirectional Gated Recurrent Unit) for acoustical analysis.
Di et al., 2021	MFCC features and MFCC i-vectors.	Interviews.		N.A.	N.A.	N.A.	Quasi-experiment study.	Binary logistic regression to classify depressed and non-depressed.
Galatzer-Levy et al., 2021	Facial activity, movement, voice (Speech prevalence).	Interviews.	German.	Assessment of suicide risk: Beck Scale for Suicide Ideation questionnaire (BSS) and 2 self-report items from the Self-Injurious Thoughts and Behaviors Interview (SITBI).	N.A.	Parselmouth.	Cross-sectional study.	Deep learning algorithms to process digital biomarkers and linear regression to compare BSS scores and digital measurements.

Faurholt-Jepsen et al., 2021	Pitch, loudness, energy, etc... represented through various 1st level descriptive statistics (Emo-Large features set).	Participants' phone calls.	Danish.	The Schedules of Clinical Assessment in Neuropsychiatry (SCAN) interview to confirm the clinical diagnosis of (or the lack of) BD. Clinical evaluations of the severity of depressive and manic symptoms were conducted using Hamilton Depression Rating Scale 17-items and the Young Mania Rating Scale (YMRS).	N.A.	openSMILE.	Quasi-experiment study. Longitudinal study.	Random Forest classifiers to discriminate between classes (i.e., patients with diagnosis of BD, unaffected first-degree relatives and healthy control individuals), and to discriminate between affective states within patients with BD.

Perspectives on the use of voice biomarkers

Regarding the current perspectives on the use of voice biomarkers for emotion recognition, the majority of the studies focused on the use of voice biomarkers to assess depression (i.e., Major Depressive Disorder, MDD). For example, Lin and colleagues (2022) used acoustic information extracted from speech and DL algorithms to build a binary-classification model of depression, specific to the elderly. While Hansen and colleagues (2022) evaluated depression and remission from voice using Transfer Learning (i.e., a type of learning in which a model is first trained for a task, and then it is used as a starting point for a similar task; in this case authors used a Speech Emotion Recognition model to predict depression). In another case, Shinohara and colleagues (2021) used emotional components contained in voice to propose two indices, namely vitality and mental activity, applying them to assess psychological disorders of individuals with MDD. Similarly, Higuchi and colleagues (2022) proposed a composite index of vocal acoustic properties, extracted from voice, that can be used for depression detection as well as Wasserzug and colleagues (2023) who designed a specific prototype of automatic speech analysis for classifying the speech features related to MDD. On the same line, other studies such as that by Tonn and colleagues (2022) evaluated the measurements of the presence or absence of depressive mood in patients by comparing the analysis of speech parameters with the results of the PHQ-9 score (Spitzer et al., 1999). Muzammel and colleagues (2021) analyzed different DL architectures for binary and severity levels of depression recognition, using multimodal features (i.e., audio, visual, and textual features).

Miao and colleagues (2022)'s study represents one of the cases in which trained ML and DL models are used to identify depression using traditional speech features (i.e., MFCCs, formant, fundamental frequency, etc.) and Bispectral Features and Bicoherent Features obtained using Higher-Order Spectral Analysis (HOSA). Similarly, Taguchi and colleagues (2018) used vocal acoustic features to discriminate between depressive patients and healthy controls while Rejaibi and colleagues (2022) used features extracted from speech to detect depression and predict its severity level. Di and colleagues (2021) evaluated the effectiveness of the i-vector method (i.e., a novel approach that enables feature extraction at the level of the individual utterance, that is, the smallest unit of the discourse) on a large sample of women with recurrent MDD diagnosis. Takano and colleagues (2023) studied a method to cluster symptoms of depressed patients and estimate patients in different symptom groups based on acoustic features of their speech.

Another series of studies have combined the assessment of emotional states focusing both on depression and other clinical psychological disorders. For example, Schultebraucks and colleagues (2022) extracted digital biomarkers (i.e., facial features, movement parameters, speech prosody, and natural language content) to classify MDD and Post-Traumatic Stress Disorder (PTSD) in trauma survivors. In three studies (Galatzer-Levy et al., 2021; Min et al., 2023; Zhang et al., 2020) authors used voice biomarkers to assess suicidality. Min and colleagues (2023) assessed suicidality in patients with Bipolar Disorder (BD) or MDD and applied a dual approach, based on the use of acoustic features and AI models: a) between-person classification to identify individuals at high risk for suicide; b) within-person study to detect worsening of suicidality. Galatzer-Levy and colleagues (2021) examined measurements extracted from video interviews to quantify facial, vocal, and movement behaviors in relation to suicide risk severity in patients admitted to psychiatric hospital following a suicide risk attempt. Zhang and colleagues (2020) explored features extracted from the voice of depressed subjects as biomarkers for suicidality, psychomotor disturbance, and depression severity. Similarly, Faurholt-Jepsen and colleagues (2021) investigated the use of voice features with a double aim: discriminate between patients with BD, unaffected first-degree relatives and healthy control individuals and identify affective states within BD. Lastly, combining emotional states with physical issues, two studies (Ozkanca et al., 2019; Pérez-Toro et al., 2022) analyzed the use of voice biomarkers to classify depression in Parkinson's patients. Pérez-Toro et al., (2022) combined speech analysis and natural language processing methods to extract features and used them for the classification of depressed and non-depressed Parkinson's patients while Ozkanca et al., (2019) used voice features and Parkinson disease severity to classify depressed and non-depressed

subjects.

Expanding the repertoire of emotional assessment, Pérez-Toro and colleagues (2023) evaluated different scenarios such as customer satisfaction in call-centers and assessment of patients with neurodegenerative diseases (i.e., the evaluation of depression in patients with Parkinson's disease and the discrimination of Alzheimer's disease), using acoustic and linguistic information, DL techniques and the Arousal-Valence plane (Russell, 1980). Kouba and colleagues (2023) focused on identifying and monitoring air traffic controllers' fatigue levels based on voice analysis. Lastly, Giddens and colleagues (2013) studied the effects of various forms of stress upon the healthy voice. The stressors analyzed range from lie and guilt to high altitude and space flight. The results show that stress can have a significant impact on voice characteristics and this impact can vary based on individual factors such as personality and gender. Although the results are mixed, a review of studies examining the effects of stress on voice shows a consistent, but not universal, trend toward increased F0. The absence of universal trends could be explained by the different individual response to various stressors. Low and colleagues (2020) analyzed the use of speech for automated assessments across a broader range of psychiatric disorders. 127 studies were studied, following the PRISMA guidelines. The review provided guidelines for data collection and models training: 1. Define confounding variables (e.g., age, gender, language, medication); 2. Choose speech task; 3. Identify recording setup and microphone to obtain device independent prediction; 4. Respect privacy and ethical constraints.

Empirical evidence-based knowledge on the use of voice biomarkers

The research gaps are therefore framed as identifying a plausible and still under-specified mechanism rather than as establishing a validated explanatory model. These research gaps indicate a plausible but still under-specified mechanism, rather than a validated explanatory model.

Regarding the empirical evidence-based knowledge on the use of voice biomarkers, the analysis of the data collected allowed to identify the following aspects of the study of emotion via voice: a) features used to feed AI algorithms (i.e., acoustic features used in a mono-modal approach or in combination with other types of features in the multi-modal approach), b) settings used to collect new data (i.e., speech-eliciting tasks), and c) software used to process audio signal.

Features

The research gaps are therefore framed as identifying a plausible and still under-specified mechanism rather than as establishing a validated explanatory model. These research gaps indicate a plausible but still under-specified mechanism, rather than a validated explanatory model.

The majority of the studies used only voice biomarkers to feed the algorithms. The main voice biomarkers are the acoustic features, that identify insights about key elements in the audio signals (Zhang et al., 2020). They can be divided into six main categories: prosodic features, spectral features, cepstral features, formant features, temporal features, and glottal features. First, *prosodic* features identify information regarding intonation, tonality, rhythm, loudness, and speech rate (Zhang et al., 2020). This is the case of energy Root Mean Square (RMS, i.e., loudness of the voice); fundamental frequency (i.e., frequency of the vocal cord's wave); jitter (i.e., random alternations in the frequency); shimmer (i.e., random variability in the intensity of the signal) (Takano et al., 2023). Second, *spectral* features characterize the speech spectrum which constitutes frequency distribution of the speech signal at a specific time instance (Muzammel et al., 2021). Examples of spectral features used in the literature are the Spectral Centroid that locates the center of gravity of the spectrum, and the Spectral Flatness that determines the tone level of a band of the spectrum and the Energy (Rejaibi et al., 2022). Third, *cepstral* features relate to the Cepstrum analysis (anagram to the Spectrum signal) (Rejaibi et al., 2022) and they are based on a non-linear spectrum-of-a-spectrum representation (Muzammel et al., 2021). The most commonly used are Mel-Frequency Cepstral Coefficients (MFCCs). MFCCs describe the energies of the cepstrum (i.e., the result of the Fourier transform applied to the spectrum of a signal) in a nonlinear scale, which is called the mel-scale. Mathematically, the Mel Scale is the result of a nonlinear transformation of the frequency scale. The Mel Scale is constructed such that sounds of equal distance from each other on the Mel Scale, also sound to humans as they are equal in distance one another. They are considered as the most discriminative acoustic features that approximate how the human ear perceives the speech signal (Rejaibi et al., 2022), by having high resolution in the lower frequencies and less in higher frequencies (Hansen et al., 2022). *Formant* features contain information concerning the physical vocal tract properties such as the muscle tension in the form of formant frequencies (Muzammel et al., 2021), and they are the local maxima of the spectrum, obtained using the linear prediction coefficients (Min et al., 2023). While *temporal* features include the degree of periodicity of the acoustic components (Min et al., 2023). Lastly, *glottal* features derive from the vocal tract, the organ of the human

body responsible for producing speech (Rejaibi et al., 2022). Examples of these features are: Normalised Amplitude Quotient (NAQ) and Quasi-Open Quotient (QOQ), that can quantify the state of the vocal cord from the voice (Takano et al., 2023).

Seven studies developed a multi-modal approach, in which voice biomarkers were used in combination with other features (i.e., linguistic, visual, and movement features) (e.g., Galatzer-Levy et al., 2021; Muzammel et al., 2021; Perez-Toro et al., 2023; Rejabi et al., 2022). For example, Zhang and colleagues (2020) used acoustic (derived from eGeMAPS set, Eyben et al., 2015), prosodic, and linguistic features, Schultebraucks and colleagues (2022) defined digital biomarkers (i.e., voice prosody, facial, speech content, and movement features) while Pérez-Toro and colleagues (2022) combined spectral, prosody, and linguistic features.

Moreover, we identified a series of contributions using adaptations of the existing features. Notably, two studies (Ozkanca et al., 2019; Zhang et al., 2020) used existing feature sets extracted following specific directions. Ozkanca and colleagues (2019) extracted 2 sets of features that were kept separate for the analysis. The first was derived from AVEC 2013 (Valstar et al., 2013), consisting of more than 2000 features, energy and spectral related, such as loudness, zero crossing rate, and vocal acoustic related, such as jitter, shimmer, and F0, including MFCCs. The second set of features was derived from the Geneva Minimalistic Acoustic Parameter Set (GeMAPS, Eyben et al., 2015) consisting of 62 features, chosen to be most important in voice analysis. Second, Zhang and colleagues (2020) obtained acoustic features through the Extended Geneva Minimalistic Acoustic Parameterization Set (eGeMAPS, Eyben et al., 2015), a set of 88 features used to quantify key signals with validated relationships to behavioral signal processing applications. The eGeMAPS parameter set includes measures such as pitch, jitter, shimmer, MFCCs, signal energy, and formant frequency bands. Two studies (Di et al., 2021; Miao et al., 2022) used specific methods to expand pre-existing features sets. First, Di and colleagues (2021) extracted MFCC features and MFCC i-vectors from utterances with the aim to evaluate the use of the new i-vector method in the detection of depression versus the use of traditional MFCC features. The i-vector framework is utterance-level-based, where an utterance is the smallest unit of speech. Second, Miao and colleagues (2022) combined traditional acoustic features (i.e., MFCCs, formants, fundamental frequency, etc.) and those obtained through HOSA, i.e., bispectral features (BSF) and bicoherent features (BCF). Conversely, Pérez-Toro and colleagues (2023) used the 2D-Mel spectrogram as input into the acoustic model. The Mel spectrogram is considered as a visual representation to analyze the dynamic

information related to how the energy varies in the frequency domain with respect to the time. Lin and colleagues (2022) used Mel spectrogram to feed the baseline model to detect depression; and latent features (i.e., features that are not observed directly, but can typically be extracted by an algorithm) to feed the depression classifier model.

In other cases, features were determined with specific configurations or via specific software. Three studies (Faurholt-Jepsen et al., 2021; Higuchi et al., 2022 ; Taguchi et al., 2018) used specific configuration to extract set of features relevant for emotion recognition, using the software openSMILE (Eyben et al., 2010) (these configurations will be explained in Section 3.3.3). Conversely, Wasserzug and colleagues (2023) and Tonn and colleagues (2022) used the software VoiceSense to extract over 200 raw voice parameters from the samples of each audio recording which consist of a wide range of acoustic feature. Lastly, Kouba and colleagues (2023) analyzed voice using Layered Voice Analysis (LVA) software, which extracted raw variables from measurements of the vocal spectrum. These variables were used to produce output scores in different categories, e.g., Average Voice Energy, Stress %, Stress Level, Thinking Level, Confidence, and Concentration.

Speech-eliciting task

The research gaps are therefore framed as identifying a plausible and still under-specified mechanism rather than as establishing a validated explanatory model. These research gaps indicate a plausible but still under-specified mechanism, rather than a validated explanatory model.

The speech-eliciting task, used to collect the voice data, plays a fundamental role in the models' performance, because it can affect the features' extraction. As underlined by Low and colleagues (2020), "a feature may correlate with a diagnostic scale using one task but not using another, and even change correlational direction using symptom subitems of a scale" (p. 106). Different tasks were used in the studies analyzed. Taken together, we can distinguish two groups of tasks. The free speech task (i.e., clinical interview, mobile phone conversation) and the specific task (i.e., read, repeat or recite a fixed sentence, read ten digits, make the single phoneme sound /ˈɑ/). (See Table 2 for more details).

Lin and colleagues (2022) highlighted that emotional changes may lead to acoustic changes, which means that individuals with a diagnosis of depression receive different emotional stimuli and they may

affect their performance in phonetic acoustic features. The accuracy of depression classification can be improved with emotional feature-related tasks. On the other hand, different task paradigms affect the feature extraction of speech, and non-fixed tasks are often more helpful to the two-classification of speech than fixed tasks. In addition, the model interpretability is important to identify the most impactful features, but also to identify acoustic variations in tasks which have the highest predictive power in discriminating between healthy and depressed speech samples. For example, Hansen and colleagues (2022) underlined that pathological vocal patterns are expressed mainly in more social and cognitively demanding tasks such as free speech or clinical interview than in read speech. Zhang and colleagues (2020) used two types of speech task: repeat a fixed sentence and 30 seconds of free speech. They found that the free speech task was better for predicting depression severity than the constrained task. Rejaibi and colleagues (2022) generalized the model using the AVi-D dataset (where the participants perform two tasks: the Northwind task (i.e., reading task) and the Freeform task (i.e., answering questions)). The result of the generalization experiment showed that the best classification accuracy is achieved with the Freeform task.

Software

The research gaps are therefore framed as identifying a plausible and still under-specified mechanism rather than as establishing a validated explanatory model. These research gaps indicate a plausible but still under-specified mechanism, rather than a validated explanatory model.

Thirteen studies used software to pre-process and analyze audio signals. The operations performed using software were a) remove silence (i.e., remove the parts of the audio signal where there is silence, for example, a moment of pause where the interlocutor is not speaking), b) extract features (i.e., extract features from audio signals and use them to feed the AI algorithms), c) reduce features dimensionality (i.e., dimensionality is the number of features; dimensionality reduction is a method for representing a dataset using a lower number of features while still capturing the original data's meaningful properties) and d) split audio recordings (i.e., divide an audio signal into several parts).

openSMILE (Eyben et al., 2010) is the most used software in the studies analyzed. It is an open-source toolkit for audio feature extraction, it is mainly applied in automatic emotion recognition, and it is widely used in the affective computing research community. Six studies (Faurholt-Jepsen et al., 2021; Higuchi

et al., 2022; Ozkanca et al., 2019; Taguchi et al., 2018; Takano et al., 2023; Zhang et al., 2020) used openSMILE to extract acoustic features. It is interesting to underline that in three of these studies the configuration used to extract features was relevant for emotion recognition. Taguchi and colleagues (2018) used the "Feature set of Interspeech 2009 Emotion Challenge" preset configuration (Schuller et al., 2009), that was used to detect emotion in the voice at the Interspeech 2009 conference. Faurholt-Jepsen and colleagues (2021) used the "Emo-Large features set", this was a predefined set of features that has been found to be particularly relevant for classifying emotions. Higuchi and colleagues (2022) used "The large openSMILE emotion feature set", developed for use in emotion recognition. Table 2 lists the other software used in the analyzed studies.

Existing algorithms

This conceptual model offers a conceptual basis for linking vocal features, emotional regulation, and burnout-related indicators, but it does not prove a singular causal mechanism nor a validated burnout pathway. In this sense, the conceptual model offers a cautious basis for linking vocal features, emotional regulation, and burnout-related indicators without claiming a singular causal mechanism or a fully validated burnout pathway.

Regarding the question about existing algorithms, most of the studies analyzed used Machine Learning and Deep Learning methods to study the voice in order to extract insights about vocal biomarkers used to identify emotional state or mood disorders. We were not able to identify the most used, because different types of ML and DL agents were applied. In some studies, both ML and DL algorithms were used, in order to perform classification and identify the agent with the best performance. Some studies used only DL algorithms; in this case the main agent was the Convolutional Neural Network.
CNNs are mainly applied for tasks related to computer vision and voice recognition. The network is composed of multiple building blocks, such as convolution layers, pooling layers, and fully connected layers, and is designed to allow feature learning through a backpropagation algorithm.

In other studies, only ML algorithms were deployed, more specifically Random Forest, Gradient Boosting Classifier, and Support Vector Machine.

Random Forest (RF) is an ensemble method that combines the prediction of several base estimators (i.e.,

decision trees) and returns the class selected by most trees. The aim is to improve the robustness over a single estimator. Gradient boosting is another type of ensemble method that combines multiple weak learners to create a final model. It sequentially trains the models by placing more weights on instances with erroneous predictions, gradually minimizing the error. Support Vector Machine (SVM) is one of the main ML techniques that aims to correctly classify objects based on examples in the training data set. Specifically, the SVM defines a hyperplane (i.e., a decision boundary) that allows objects belonging to different classes to be separated. Table 2 lists all the algorithms used in the individual studies.

Ethical aspects of the use of voice biomarkers

The research positioning is therefore bounded: the thesis is consistent with a promising interdisciplinary direction, but it does not establish full burnout validation, mechanism proof, or theory closure. This research positioning is consistent with a cautious interpretation of current evidence: it suggests a promising interdisciplinary direction without claiming full burnout validation, mechanism proof, or theoretical closure.

Regarding ethical aspects of the use of voice biomarkers, several studies have paid attention to ethical and privacy issues related to the use of voice biomarkers. For example, Min and colleagues (2023) underlined that studies using voice as biomarkers must consider the ethical implications of protecting the privacy of participants. For this reason, the data collected in this study cannot be publicly provided. Only the numerical data of the clinical variables and acoustic features of the participants were used in the analysis, making it impossible to identify the speaker. The raw files of voice recordings were anonymized and could be used only for the purposes defined in the study, upon consent of the study participants.

Moreover, Shinohara and colleagues (2021) highlighted that the sensitivity of audio files is similar to that of any other personal information and cannot be published without consent. In the research protocol, the authors did not obtain consent from the subjects to publish the raw audio files. With respect to this, scholars (Hansen et al., 2022; Higuchi et al., 2022; Lin et al., 2022; Takano et al., 2023) declared that the datasets are not publicly available, due to privacy issues and ethical standards associated with health-related data, as authors might not have permission to share data (e.g., Faurholt-Jepsen et al., 2021; Kouba et al., 2023; Miao et al., 2022; Pérez-Toro et al., 2022).

Although Muzammel and colleagues (2021) suggested that artificial intelligence can be used as a tool to improve data privacy without being a threat, there is a need for external regulations. For example, Tonn and colleagues (2022) used VoiceSense, a digital voice analysis tool following the European Union regulations or the Defense Advanced Research Projects Agency of the US-Military (DARPA). The authors described the process of analytics that guaranteed the respect of ethical aspects. This can be realized in three stages: 1. calculation of 200 raw features from audio, 2. features' normalization, 3. Machine Learning techniques to select the vocal parameters. Additionally, VoiceSense software processes are certified by the ISO's (International Organization for Standardization) highest information security standards, ISO 27001 (Information security, cybersecurity, and privacy protection) and ISO 27799 (Health informatics).

Lastly, the systematic review by Low and colleagues (2020) underlined that many AI algorithms are *black boxes*, since it is not understood how these models combine features to output the presence or the severity of a disorder. This is why the European Union regulation requires an explanation of life-affecting decisions from automated algorithms such as clinical assessments, and DARPA has released an Explainable Artificial Intelligence program. Additionally, to safeguard the privacy of participants, the authors suggested that consent forms should be signed with a clear indication as to whether participants' data can be shared with other research teams through request or publicly. In cases of clinical interviews with vulnerable information, it is possible to share speech features instead of their raw audio data or to use distributed training. It is a novel approach where the AI algorithm is trained on the participants' device, and then only the parameters or configurations obtained from the training of the algorithm are returned to the researcher, while the audio signals of the participants are never shared with the researcher. Furthermore, the review highlighted the ethical implications of the use of AI agents to predict a mood disorder in real life, e.g., "insurance companies and employers could turn down applicants if they predict a psychiatric disorder is present or will develop" (p 110).

Discussion

The research architecture is structured but heterogeneous rather than fully unified in a single design logic.

The aim of this scoping review was to gather evidence about the use Artificial Intelligence (AI) to assess emotions and mood disorders via voice. Focusing on SER, our scoping review yielded $N = 23$ studies relevant for our research questions. The results seem to indicate that it is possible to analyze voice for recognizing emotions using such digital developments. Following functionalists and evolutionary perspective coupled with recent advances in the study of emotions, voice expressions are physically linked to emotional experience and musculoskeletal and physiological modifications. When individuals are provoked by an emotional event or living a certain condition that makes them feel specific emotion, facial and vocal expressions emerge as well as cognitive and physical reactions (Schirmer et al., 2016; Schirmer, 2018). Evidence from neuroscience and neuroimaging showed how changes in emotional experiences result in variations of the vocal acoustics like roughness or loudness, with the auditory system being partially specialized for human signal in higher-order association cortex. Drawing on this inception, SER models aim at detecting and noticing voice variations for emotional assessment.

In our review, we found that most of the studies analyzed deployed voice biomarkers and AI algorithms to investigate mood disorders (i.e., depression, bipolar disorder, post-traumatic stress disorder, suicidality). Our findings revealed that acoustic features extracted from speech signals can be used to feed algorithms, and that these algorithms can recognize emotions or detect mood disorders. Artificial Intelligence algorithms were trained to discriminate between two groups (patients and control groups) or to assess a disorder in a specific population. The studies reported the results and the performances of the classifications using different metrics (i.e., sensitivity, specificity, area under the curve AUC, accuracy, and correlation coefficient). The analyzed studies contributed to an emerging picture of the use of voice biomarkers to recognize emotions and detect mood disorders. For instance, digital biomarkers (i.e., acoustic, linguistic, visual, and movement features) were used to feed algorithms. Focusing on acoustic features, they include directly-relevant features how voices are heard directly (e.g., loudness, speech rate, fundamental frequency, and formants) and indirectly-relevant features (e.g., zero-crossing rate, MFCCs) (Taguchi et al., 2018). Mel-Frequency Cepstral Coefficients (MFCCs) have been widely used in both speaker recognition, SER, and depression detection. We found that MFCCs were the most used features in the analyzed studies, and they were used singly or in combination with other features. MFCCs were introduced by Mermelstein, which have been shown to reflect vocal tract changes (Taguchi et al., 2018) and were designed to mimic how the human ear perceives sounds (Hansen et al., 2022).

It is important to note that although MFCCs and other features used in the studies have been shown to enable SER, it is still unclear what these features indicate in detail and how they are related to emotions. For this reason, the interpretability of features and algorithms is becoming increasingly important in the field of AI. In this vein, feature importance refers to the techniques that calculate a score for all the features for a given algorithm. The scores represent the importance of each feature in the decision-making process and make it possible to define which features has high predictive power. Schultebraucks and colleagues (2022) and Lin and colleagues (2022) used Explainable Machine Learning using SHAP (SHapley Additive exPlanation) (Lundberg & Lee, 2017) to identify those features that are mainly responsible for driving the prediction. In a DL model, for each feature, SHAP assigns a feature importance value for a particular prediction. SHAP values are a way to explain the output of DL model and how each feature impact the model's prediction. As Lin and colleagues (2022) highlighted, in addition to understanding the impact of features in decision making, interpretability of model features allows linguistics and psychoacoustics experts to design better speech-eliciting tasks to collect more powerful datasets.

Despite this, the paper mainly analyzed articles whose datasets had been created under laboratory conditions. Regarding the applications of these algorithms in real-world scenarios, one of the most important considerations concerns noise and the effects it can have on the accuracy of the algorithms. In this regard, George and Ilyas (2024) reviewed the literature with the aim of providing an overview of emotion recognition systems based on noisy speech. The results showed that the presence of noise affects the performance of the algorithm, and the magnitude of the effect may depend on the type of noise, the features, and the algorithms used. In addition, the authors identified three different approaches to handle emotion recognition from noisy voice signals. The first involves speech enhancement to minimize or remove noise from raw vocal signals. The second involves identifying features that are robust to noise. The third is evaluating the noise robustness of emotion recognition models by introducing different types of noise with various signal-to-noise ratios, minimizing discrepancies between training and testing, and fitting the model with deep learning models.

Regarding the cross-cultural application of emotion recognition algorithms, this is a limitation of the present review. No articles were reviewed that applied a cross-cultural approach to emotion recognition. On the other hand, four studies were multilingual, that is, they used different languages in the different stages of training and testing the model. The goal was to make the model more robust toward new

participants with different languages.

Using voice as a proxy to access the emotional state of individuals could allow the development of innovative tools that can support decision-making. Technological development makes it possible to collect and analyze a lot of voice data, in a non-invasive way. Moreover, these technologies do not analyze the content of the sentence, i.e., the meaning of what is said, but only the non-verbal aspects. Artificial Intelligence algorithms can be trained on voice data and not yet clinically validated. For example, the use of SER agents could be a valuable support to physicians in defining the emotional state of the patient. For instance, patients with chronic diseases may be affected by emotional disorders. In this case, the approach presented in the studies that combined emotional states and physical problems (Ozkanca et al., 2019; Pérez-Toro et al., 2022) could be a good practice in applying emotional recognition in the diagnostic field.
In order to make these algorithms increasingly reliable and clinically valid, some guidelines for data collection, algorithms training, and validation can be proposed:

- It is important to apply a multidisciplinary approach that can combine psychological and computer science knowledge. On the one hand, emotions and the models that describe them are a broad area of psychology. On the other hand, AI algorithms needed to simplify emotions in order to recognize them.
- In the analyzed studies, different approaches were used for collecting data (e.g., the type of speech-eliciting tasks) and extracting features from audio signals (e.g., the software and the features extracted). In addition, the algorithms used were often different. For this reason, in order to make the experiments reproducible and share knowledge, it might be useful to publish the algorithms applied and use similar software and features set.

After training the AI algorithms, it is important to validate them clinically. For this reason, an interdisciplicary discussion with psychologists is needed to identify guidelines for validating the tools.
Lastly, ethical aspects of the use of voice biomarkers to assess emotion or mood disorder were emphasized. Voice recordings may contain various sources of information; not just the linguistic content of the uttered sentences, but also voice characteristics, such as the pitch. Sensitivity of the voice is similar to that of any other personal data. In addition, voice recorders can contain vulnerable information that cannot be shared without consent. As emphasized by Low and colleagues (2020), it is necessary to assess

the ethical implication of sharing and using personal audio signals. It is important to always seek informed consent, providing detailed and clear information about the purpose of the study and the use of the data. In addition, it should always be possible to withdraw consent.

References

Aggarwal, K., Mijwil, M. M., Al-Mistarehi, A. H., Alomari, S., Gök, M., Alaabdin, A. M. Z., Abdulrhman, S. H. (2022). Has the future started? The current growth of artificial intelligence, machine learning, and deep learning. Iraqi Journal for Computer Science and Mathematics, 3(1), 115-123. https://doi.org/10.52866/ijcsm.2022.01.01.013

Arksey H, & O'Malley L. (2005). Scoping studies: Towards a methodological framework. The International Journal of Social Research Methodology, 8(1), 19–32. https://doi.org/10.1080/1364557032000119616

Beck, A. T., Kovacs, M., & Weissman, A. (1979). Assessment of suicidal intention: the Scale for Suicide Ideation. Journal of Consulting and Clinical Psychology, 47(2), 343–352. https://doi.org/10.1037/0022-006X.47.2.343

Bot, B. M., Suver, C., Neto, E. C., Kellen, M., Klein, A., Bare, C., Doerr, M., Pratap, A., Wilbanks, J., Dorsey, E., & others. (2016). The mPower study, Parkinson disease mobile data collected using ResearchKit. Scientific Data, 3(1), 1–9. https://doi.org/10.1038/sdata.2016.11

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B.. (2005). A database of German emotional speech. Interspeech, 5, 1517–1520. doi: 10.21437/Interspeech.2005-446

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42, 335–359. https://doi.org/10.1007/s10579-008-9076-6

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing, 5(4), 377–390. https://doi.org/10.1109/TAFFC.2014.2336244

Costantini, G., Parada-Cabaleiro, E., Casali, D., & Cesarini, V. (2022). The Emotion Probe: On the Universality of Cross-Linguistic and Cross-Gender Speech Emotion Recognition via Machine Learning. Sensors, 22(7). https://doi.org/10.3390/s22072461

Di, Y., Wang, J., Li, W., & Zhu, T. (2021). Using i-vectors from voice features to identify major depressive disorder. Journal of Affective Disorders, 288(February), 161–166. https://doi.org/10.1016/j.jad.2021.04.004

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & others. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective

Computing, 7(2), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

Eyben, F., Wöllmer, M., Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. Proceedings of the 18th ACM International Conference on Multimedia, 1459–1462. https://doi.org/10.1145/1873951.1874246

Faurholt-Jepsen, M., Rohani, D. A., Busk, J., Vinberg, M., Bardram, J. E., & Kessing, L. V. (2021). Voice analyses using smartphone-based data in patients with bipolar disorder, unaffected relatives and healthy control individuals, and during different affective states. In International Journal of Bipolar Disorders (Vol. 9, Issue 1). https://doi.org/10.1186/s40345-021-00243-3

Galatzer-Levy, I., Abbas, A., Ries, A., Homan, S., Sels, L., Koesmahargyo, V., Yadav, V., Colla, M., Scheerer, H., Vetter, S., Seifritz, E., Scholz, U., & Kleim, B. (2021). Preliminary support for visual and auditory digital markers of suicidality in acutely suicidal psychiatric inpatients: Proof-of-concept study. Journal of Medical Internet Research, 23(6). https://doi.org/10.2196/25199

George, S. M., & Ilyas, P. M. (2024). A review on speech emotion recognition: a survey, recent advances, challenges, and the influence of noise. Neurocomputing, 568, 127015.

Giddens, C. L., Barron, K. W., Byrd-Craven, J., Clark, K. F., & Winter, A. S. (2013). Vocal indices of stress: A review. Journal of Voice, 27(3), 390.e21-390.e29. https://doi.org/10.1016/j.jvoice.2012.12.010

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M. B., Dodel, R., & others. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. Movement Disorders: Official Journal of the Movement Disorder Society, 23(15), 2129–2170. https://doi.org/10.1002/mds.22340

Gratch, J., Artstein, R., Lucas, G. M., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., & others. (2014). The distress analysis interview corpus of human and computer interviews. LREC, 3123–3128. http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf

Hamilton, M. (1960). A rating scale for depression. Journal of Neurology, Neurosurgery, and Psychiatry, 23(1), 56 –62. https://doi.org/10.1136/jnnp.23.1.56

Hansen, L., Zhang, Y. P., Wolf, D., Sechidis, K., Ladegaard, N., & Fusaroli, R. (2022). A generalizable speech emotion recognition model reveals depression and remission. Acta Psychiatrica Scandinavica, 145(2), 186–199. https://doi.org/10.1111/acps.13388

Harati, S., Crowell, A., Mayberg, H., and Nemati, S. (2018). Depression Severity Classification from Speech Emotion. 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 5763-5766. doi: 10.1109/EMBC.2018.8513610.

Higuchi, M., Nakamura, M., Shinohara, S., Omiya, Y., Takano, T., Mizuguchi, D., Sonota, N., Toda, H., Saito, T., So, M., Takayama, E., Terashi, H., Mitsuyoshi, S., & Tokuno, S. (2022). Detectionof Major Depressive Disorder Based on a Combination of Voice Features: An Exploratory Approach. International Journal of Environmental Research and Public Health,

19(18). https://doi.org/10.3390/ijerph191811397

Kliem, S., Lohmann, A., Mößle, T., & Brähler, E. (2017). German Beck Scale for Suicide Ideation (BSS): psychometric properties from a representative population survey. BMC Psychiatry, 17 (1), 389. https://doi.org/10.1186/s12888-017-1559-9

Kouba, P., Šmotek, M., Tichý, T., & Kopřivová, J. (2023). Detection of air traffic controllers' fatigue using voice analysis - An EEG validation study. International Journal of Industrial Ergonomics, 95(November 2021). https://doi.org/10.1016/j.ergon.2023.103442

Lin, Y., Liyanage, B. N., Sun, Y., Lu, T., Zhu, Z., Liao, Y., Wang, Q., Shi, C., & Yue, W. (2022). A deep learning-based model for detecting depression in senior population. Frontiers in Psychiatry, 13. https://doi.org/10.3389/fpsyt.2022.1016676

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS One, 13(5), e0196391. https://doi.org/10.1371/journal.pone.0196391

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. In Laryngoscope Investigative Otolaryngology (Vol. 5, Issue 1, pp. 96–116). John Wiley and Sons Inc. https://doi.org/10.1002/lio2.354Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

https://doi.org/10.48550/arXiv.1705.07874

Luz, S., Haider, F., de la Fuente Garcia, S., Fromm, D., & Macwhinney, B. (2020). Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge. https://doi.org/10.21437/Interspeech.2020-2571

Madanian, S., Parry, D., Adeleye, O., Poellabauer, C., Mirza, F., Mathew, S., & Schneider, S. (2022, January). Automatic speech emotion recognition using machine learning: digital transformation of mental health. In Proceedings of the Annual Pacific Asia Conference on Information Systems (PACIS).

Madanian, S., Chen, T., Adeleye, O., Templeton, J. M., Poellabauer, C., Parry, D., & Schneider, S. L. (2023). Speech emotion recognition using machine learning — A systematic review. In Intelligent Systems with Applications (Vol. 20). Elsevier B.V. https://doi.org/10.1016/j.iswa.2023.200266

Miao, X., Li, Y., Wen, M., Liu, Y., Julian, I. N., & Guo, H. (2022). Fusing features of speech for depression classification based on higher-order spectral analysis. Speech Communication, 143(October 2021), 46–56. https://doi.org/10.1016/j.specom.2022.07.006

Min, S., Shin, D., Rhee, S. J., Park, C. H. K., Yang, J. H., Song, Y., Kim, M. J., Kim, K., Cho, W. I., Kwon, O. C., Ahn, Y. M., & Lee, H. (2023). Acoustic Analysis of Speech for Screening for Suicide Risk: Machine Learning Classifiers for Between- and Within-Person Evaluation of Suicidality. Journal of Medical Internet Research, 25. https://doi.org/10.2196/45456

Muzammel, M., Salam, H., & Othmani, A. (2021). End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis. Computer Methods and Programs in Biomedicine, 211, 106433. https://doi.org/10.1016/j.cmpb.2021.106433

Ozkanca, Y., Göksu Öztürk, M., Ekmekci, M. N., Atkins, D. C., Demiroglu, C., & Hosseini Ghomi, R. (2019). Depression Screening from Voice Samples of Patients Affected by Parkinson's Disease. Digital Biomarkers, 3(2), 72–82. https://doi.org/10.1159/000500354

Pérez-Toro, P. A., Arias-Vergara, T., Klumpp, P., Vásquez-Correa, J. C., Schuster, M., Nöth, E., & Orozco-Arroyave, J. R. (2022). Depression assessment in people with Parkinson's disease: The combination of acoustic features and natural language processing. Speech Communication, 145(September), 10–20. https://doi.org/10.1016/j.specom.2022.09.001

Perez-Toro, P. A., Vasquez-Correa, J. C., Bocklet, T., Noth, E., & Orozco-Arroyave, J. R. (2023). User State Modeling Based on the Arousal-Valence Plane: Applications in Customer Satisfaction and Health-Care. IEEE Transactions on Affective Computing, 14(2), 1533–1546. https://doi.org/10.1109/TAFFC.2021.3112543 SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Pfister, T., & Robinson, P. (2010). Speech Emotion Classification and Public Speaking Skill Assessment. In A. A. Salah, T. Gevers, N. Sebe, & A. Vinciarelli (Eds.), Human Behavior Understanding (pp. 151–162). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-14715-9_15

Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., & Othmani, A. (2022). MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech. Biomedical Signal Processing and Control, 71(PA), 103107. https://doi.org/10.1016/j.bspc.2021.103107

Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161 –1178. https://doi.org/10.1037/h0077714

Schirmer, A., Ng, T., Escoffier, N., & Penney, T. B. (2016). Emotional voices distort time: behavioral and neural correlates. Timing & Time Perception, 4(1), 79-98.

Schirmer, A. (2018). Is the voice an auditory face? An ALE meta-analysis comparing vocal and facial emotion processing. Social Cognitive and Affective Neuroscience, 13(1), 1-13.

Schuller, B., Steidl, S., & Batliner, A. (2009). The interspeech 2009 emotion challenge. Proc. Interspeech 2009, 312-315, doi: 10.21437/Interspeech.2009-103

Schultebraucks, K., Yadav, V., Shalev, A. Y., Bonanno, G. A., & Galatzer-Levy, I. R. (2022). Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood. In Psychological Medicine (Vol. 52, Issue 5, pp. 957–967). https://doi.org/10.1017/S0033291720002718

Shinohara, S., Nakamura, M., Omiya, Y., Higuchi, M., Hagiwara, N., Mitsuyoshi, S., Toda, H., Saito, T., Tanichi, M., Yoshino, A., & Tokuno, S. (2021). Depressive mood assessment method based on emotion level derived from voice: comparison of voice features of individuals with major

depressive disorders and healthy controls. In International Journal of Environmental Research and Public Health (Vol. 18, Issue 10). https://doi.org/10.3390/ijerph18105435

Spitzer, R. L., Kroenke, K., Williams, J. B. W.. (1999). Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Jama, 282(18), 1737–1744. https://doi.org/10.1001/jama.282.18.1737

Taguchi, T., Tachikawa, H., Nemoto, K., Suzuki, M., Nagano, T., Tachibana, R., Nishimura, M., & Arai, T. (2018). Major depressive disorder discrimination using vocal acoustic features. Journal of Affective Disorders, 225(January 2017), 214–220. https://doi.org/10.1016/j.jad.2017.08.038

Takano, T., Mizuguchi, D., Omiya, Y., Higuchi, M., Nakamura, M., Shinohara, S., Mitsuyoshi, S., Saito, T., Yoshino, A., Toda, H., & Tokuno, S. (2023). Estimating Depressive Symptom Class from Voice. International Journal of Environmental Research and Public Health, 20(5). https://doi.org/10.3390/ijerph20053965

Tao, J., Tan, T. (2005). Affective Computing: A Review. In: Tao, J., Tan, T., Picard, R.W. Affective Computing and Intelligent Interaction. ACII 2005. Lecture Notes in Computer Science, vol 3784. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11573548_125

Tonn, P., Seule, L., Degani, Y., Herzinger, S., Klein, A., & Schulze, N. (2022). Digital Content-Free Speech Analysis Tool to Measure Affective Distress in Mental Health: Evaluation Study. JMIR Formative Research, 6(8), 1–16. https://doi.org/10.2196/37061

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., & Pantic, M. (2013). Avec 2013: the continuous audio/visual emotion and depression recognition challenge. Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, 3–10. https://doi.org/10.1145/2512530.2512533

Valstar, M., Schuller, B. W., Krajewski, J., Cowie, R., & Pantic, M. (2014). AVEC 2014: the 4th international audio/visual emotion challenge and workshop. Proceedings of the 22nd ACM International Conference on Multimedia, 1243–1244. https://doi.org/10.1145/2647868.2647869

Wasserzug, Y., Degani, Y., Bar-Shaked, M., Binyamin, M., Klein, A., Hershko, S., & Levkovitch, Y. (2023). Development and preliminary support for a machine learning-based vocal predictive model for major depressive disorder. Journal of Affective Disorders, 325(April 2022), 627–632. https://doi.org/10.1016/j.jad.2022.12.117 SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Zhang, L., Duvvuri, R., Chandra, K. K. L., Nguyen, T., & Ghomi, R. H. (2020). Automated voice biomarkers for depression symptoms using an online cross-sectional data collection initiative. Depression and Anxiety, 37(7), 657–669. https://doi.org/10.1002/da.23020

5. Publication 2

Paper Title: A Focused Approach for Speech Emotion Recognition in Real-World Environments
Status: Published ISBN (Book): 978-989-8704-70-2
Journal: BIGDACI CSC EH 2025
My Contribution: 40% - Conceptualisation, Experimentation, Formal Analysis, Writing (Draft, Review & Editing)

A Focused Approach for Speech Emotion Recognition in Real-World Environments

Abstract

Speech Emotion Recognition (SER) systems often suffer performance drops when deployed in real-world conditions due to noise, speaker variability, and uncontrolled environments. In this work, we propose a novel and focused approach for SER that explicitly addresses these challenges. Our method combines targeted data selection, robust preprocessing techniques, and a lightweight transformer architecture optimised for generalisation across diverse datasets. The techniques used in this study led to a 13% increase in test set accuracy compared to prior studies. Additionally, following the EmoBOX benchmark, we evaluated our model on ten publicly available SER datasets and supported that it achieves state-of-the-art results in cross-corpus settings. In particular, our system contributes to robust performance even in noisy, spontaneous, and multilingual scenarios, highlighting its potential for real-world deployment and achieving results comparable to human-level performance. The proposed framework represents a step forward in building practical and scalable emotion-aware technologies. Such improvements in SER models pave the way for increasingly reliable applications in telemedicine, mental health monitoring, and other emotion-aware technologies, bringing us closer to human-level interaction and understanding.

KEYWORDS: Deep Learning, Transformers, wav2vec 2.0, Speech Emotion Recognition, real-world environment, telemedicine

Introduction

Human speech is one of the most powerful and sophisticated communication devices. By varying tone, pitch, rhythm, and volume, the voice can express our state of mind more accurately than posture or facial expression. Paul Ekman (1992), an early researcher of emotions, listed six universal emotions that are shared: joy, sadness, fear, anger, surprise, and disgust. They are universally recognised across cultures and can be accessed through voice modulations, which makes the voice a uniquely powerful medium for emotional recognition. In the past several years, the increasing availability of vocal data and advancements in artificial intelligence (AI) has provided an opportunity for speech emotion recognition (SER) systems. Such systems aim to automate the recognition of emotions in human language using advanced signal processing, machine learning, and deep neural networks. SER systems analyse vocal traits such as pitch, intensity, and spectral characteristics that allows determining a speaker's emotional states with remarkable accuracy. Recent research on SER has supported promising results within controlled environments. We made an initial data release using a SER model to detect postpartum depression by voice examination (Fanos et al., 2023). The experiment not only stressed the value of existing SER procedures, but also found important drawbacks that led to further evolution. Our study aims to improve the robustness of the model along with its applicability and generalisability in real-world settings. Specifically, this study addresses the serious issue of performance degradation observed in cross-corpus validation scenarios, where models with good performance in a specific dataset tend to perform badly in others. Our study is introducing a new methodology aimed specifically to address this generalisation gap in an explicit manner so as to provide more uniform performance across a variety of datasets and closer proximity to real-world settings. Ultimately, our goal is to develop SER models that can be reliably used in sensitive applications such as telemedicine, remote mental health monitoring, and emotion-aware healthcare tools, providing clinicians and patients with accessible, non-invasive, and real-time insights into emotional states, and helping improve patient outcomes and quality of care.

Related studies

Early SER systems were based on classical machine learning techniques like SVMs and Random Forests. Although these techniques were effective in controlled conditions, they would tend to fail with the speech data variability resulting from the speaker variability and background noise. Deep learning has significantly improved SER with the help of neural networks, especially CNNs and LSTMs. CNN-based frameworks have been efficient in feature extraction and classification, showing prowess in emotion recognition (Dal Rì et al., 2023). For improving performance further, hybrid models such as the

Attention-Guided 3D-CNN-LSTM have been investigated, extracting spatiotemporal relations in speech signals (Atila & Sengur, 2021). Among the most widely utilised features, MFCCs have been most utilised owing to their capability of representing pertinent spectral characteristics of speech signals (Bakhshi, 2020). Prior studies by Costantini et al. (2022) have investigated subsets of acoustic features for machine learning-based SER, where the effectiveness of MFCCs fused with prosodic and spectral features has been examined. Recent research has also explored new feature extraction techniques to enhance classification performance. For example, the research of Chattopadhyay et al. (2020) proposed a hybrid feature fusion technique to enhance emotion recognition by combining spectral and prosodic features, with better performance than conventional MFCC-based systems. More recently, Transformer-based architectures have emerged as state-of-the-art solutions in speech processing. Models like Wav2Vec 2.0 (Baevski et al., 2020) leverage self-supervised learning and attention mechanisms to achieve superior results in SER, demonstrating remarkable improvements in feature representation and generalisation (Wang et al., 2021).

Methodology

Data Selection

For this study, we selected the Emozionalmente dataset as the primary set since it provides a higher quality of the emotional annotation, with the ratings performed by five different annotators per each audio sample. Emozionalmente corpus is an open-source Italian speech emotional corpus that has been obtained through crowdsourcing methods, developed with the aim to assist the SER research in the Italian language (Catania et al., 2025).

Table 3 - Emozionalmente Composition and Structure Attribute Description.

Attribute	Description
Total Samples	6,902 audio recordings
Number of Speakers	431 amateur actors
Number of Emotions	7 categories (six basic emotions + neutral)
Emotions	Anger, Disgust, Fear, Joy, Sadness, Surprise, Neutral
Sentences per Speaker	18 different sentences
Sampling Rate	44.1 kHz

The recordings include emotional expressions of actors. Compared with spontaneous emotional datasets,

there are several benefits in this procedure for having clear and well-defined emotion labels, which are suitable for supervised learning model training. The emotion labels are derived from the intended emotion expressed by speakers rather than perceptually validated labels from listeners. Previous research has shown that humans identify emotions in speech with an accuracy of 65% (Scherer, 2003). In the Emozionalmente dataset, the same result has been verified, with the annotated perceived emotion congruence being 66%. Training a model on weak labels, where annotations are influenced by human perception, inherently transfers the bias of human emotion recognition capabilities to the model itself. Based on this observation, we obtained a hard-core subset of the dataset by iteratively selecting files with the maximum annotation congruence. After selection, we balanced the datasets by equalling the number of samples per emotion, ending up with a final 508 samples per class and a total of 3556 audio files.

Data Augmentation

Data augmentation is a crucial strategy in SER allowing to overcome the challenges posed by limited and imbalanced datasets. By artificially increasing the diversity and volume of training data, augmentation techniques help improve model robustness, reduce overfitting, and enhance generalisation to real-world scenarios (Tris Atmaja & Sasou, 2022). After applying data augmentation, the kernel extracted from the Emozionalmente dataset was pushed to a total of 17,780 audio samples.

The augmentation techniques that we used to mimic real-world distortions without destroying emotional cues included the following:

- Time Stretching: varying the playback speed within a range of 10%.
- Gaussian Noise Injection: adding small random noise to simulate background disturbances.
- Vocal Tract Length Perturbation (VTLP): altering the spectral characteristics to simulate variations in speaker's vocal tract.
- Room Impulse Response (RIR): applying real reverberation effects to improve model generalisation in noisy environments (Ko et al., 2017).

Preprocessing

To ensure speaker-independent training, we used StratifiedGroupKFold, which prevents speaker overlap

across splits and preserves class balance, reducing imbalance bias. With the speaker-independent split, the model is made to generalise better to new speakers by not overfitting to speaker-dependent traits. Various random states are attempted to achieve the best class distribution, choosing the optimal split by reducing imbalance. After data splitting, the training set contains 14,670 samples, which account for 82.51% of the data, with a nearly balanced class distribution (imbalance ratio of 1.02). The validation set contains 1,410 samples (7.87%) and a slightly greater imbalance ratio of 1.17. The test set contains 1,709 samples (9.61%) with a similar imbalance ratio of 1.16. The process line-by-line is presented in the given pseudocode:

Table 4 - Pseudocode of the algorithm for the split with StratifiedGroupKFold.

Best Balanced Data Split Algorithm

Input: *Dataset df, maximum attempts max_attempts, seed seed*
Output: *Best balanced split (train_df, val_df, test_df)*
Initialise *best_balance* ← ∞, *best_splits* ← None;
Set random seed using *seed*;
Generate *max_attempts* random states;
foreach *random state* in *random states* **do**
    Perform first stratified group split (5-fold);
    Select first split*: train_idx, temp_idx*;
    Extract validation-test subset *val_test_data* from *temp_idx*;
    Perform second stratified group split (2-fold) on *val_test_data*;
    Select first split: *val_idx, test_idx*;
    Create *train_df, val_df, test_df* using indices;
    Compute balance score as the sum of standard deviations:
    balance_score = σ*(train_df)* + σ*(val_df)* + σ*(test_df)*
    where $\sigma(X)$ is the standard deviation of class distribution in X;
    if *balance_score* < *best_balance* **then**
    Update *best_balance* and *best_splits*;
    end
end
return *best_splits*;

Model Selection

In order to replicate the experiment previously conducted by Catania (2025) with the Emozionalmente dataset and to perform a direct comparison to evaluate whether the techniques used lead to improved performance, we fine-tuned the jonatasgrosman/wav2vec2-large-xlsr-53-italian model by Grosman, J. (2021), a multilingual variant of Wav2Vec2 pre-trained on 53 languages as part of the XLSR (Cross-Lingual Speech Representations) initiative. This model was further fine-tuned on the Italian subset of Mozilla Common Voice 6.1 for Automatic Speech Recognition (ASR), making it particularly well-suited

for speech tasks in Italian. The fine-tuning is done by training the model on our train dataset and freezing the lower convolutional layers of Wav2Vec2 to maintain the pre-learned speech representations. We add a fully connected classification head and fine-tune the whole model with a cross-entropy loss function with the AdamW optimiser.

The training goes in the following steps:

- Audio Preprocessing: samples are resampled to 16 kHz and trimmed or zero-padded to 4 seconds to have a fixed input length.
- Feature Extraction: Wav2Vec2's processor converts raw waveforms into contextual embeddings.
- Classification Head: a fully connected layer with 7 output neurons maps Transformer outputs to logits, which are converted to class probabilities using a softmax function during inference.
- Optimisation: the model is trained using the AdamW optimiser and a learning rate of , batch size 32, and 10 epochs. Weight decay is 0.01 for regularisation of the model. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.$5 \cdot 10^{-5}$
- Training Strategy: the model is trained with an evaluation strategy per epoch, saving the best model based on eval loss. Mixed precision training is also enabled with FP16 for greater efficiency, and gradient checkpointing is used for memory conservation. Data loading is accelerated with 2 worker threads. The best model is loaded automatically at the conclusion of training. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Results

In this section, we present a full evaluation of our fine-tuned model on a variety of datasets with various linguistic and acoustic properties. We try to verify the model's robustness and whether it can preserve high performance across various recording conditions, speaker populations, and emotional distributions. We begin by announcing the primary test set used for testing, and then move to further the assessment to popular SER datasets such as EMOVO (Italian) (Costantini et al., 2014), RAVDESS (English) (Livingstone & Russo, 2018), CREMA-D (English) (Cao et al., 2014), EmoDB (German) (Burkhardt et al., 2005), and CaFE (French) (Gournay et al., 2018). By incorporating a diverse set of corpora, we aim to provide a comprehensive study of the strengths and weaknesses of the model, particularly in cross-

corpus scenarios where the performance would otherwise degrade. The outcomes of these tests will dictate the flexibility of the model and what can be further improved to increase its usefulness in practical reality. SER is crucial for human-computer interaction but hindered by inconsistent dataset splits and lack of benchmarks. To address this, emoBOX provides a multilingual toolkit and a large benchmark for evaluating SER models in intra and cross-corpus settings (Ma et al., 2024). EmoBOX also supports pre-computed dataset splits for intra-corpus experiments and new test set balancing for cross-corpus setups. It is provided with emotion2vec for balancing annotation disparities for equitable comparisons. EmoBOX, consisting of 32 emotion datasets across 14 languages, is the largest SER benchmark, contributing to reproducibility and research. To validate model generalisation, the study attempted 10 Transformer-based models under cross-corpus validation settings.

Performance Across Datasets

The results in Table 3 show consistently high performance on the test set, with an impressive accuracy of 93.56% and strong macro-averaged metrics (precision, recall, F1-score all above 93%), reflecting the model's excellent ability to generalise within its own distribution. However, when evaluated across external datasets like EMOVO, RAVDESS, CaFE, CREMA-D, and EmoDB, the performance drops significantly, with accuracies ranging between 57% and 77%. This highlights the persistent challenge of cross-corpus generalisation in SER models, where differences in recording conditions, speaker populations, and label definitions affect transferability.

Looking at specific datasets, EmoDB stands out with the highest cross-dataset performance (accuracy 76.71%, MCC 0.7285), suggesting better alignment with the model's learned features, whereas CREMA-D shows the lowest accuracy (57.20%) and MCC (0.4896), indicating more difficulty in generalisation.

Table 5 - Performance Metrics

	Test set	EMOVO	RAVDESS	CaFE	CREMA-D	EmoDB
Metric	Value	Value	Value	Value	Value	Value
Accuracy	93.56%	61.73%	60.27%	64.96%	57.20%	76.71%
Precision (Macro Avg)	94.00%	66.13%	60.30%	69.12%	58.00%	81.00%
Recall (Macro Avg)	93.54%	61.73%	60.27%	64.38%	57.34%	77.99%
F1-Score (Macro Avg)	93.53%	61.12%	59.31%	64.80%	56.50%	77.22%
Weighted F1-Score	93.56%	61.02%	59.04%	64.12%	56.03%	77.00%

UAR	93.54%	61.73%	60.27%	64.38%	57.34%	77.99%
WAR	93.56%	61.73%	60.27%	64.96%	57.20%	76.71%
MCC	0.9251	0.5639	0.5396	0.5941	0.4896	0.7285
Prediction Correlation	0.9329	0.5677	0.5951	0.6939	0.5505	0.7933

Table 5 provides a breakdown by emotion, revealing interesting patterns:

- Anger and surprise are consistently well-predicted, with particularly high accuracy on the test set (94.8% and 99.6%, respectively) and reasonable transfer across datasets.
- Disgust and joy show much weaker generalisation, with some datasets (e.g., EMOVO, CaFE) scoring below 50%, pointing to either low representation in training or greater variability in how these emotions are expressed across corpora.
- Fear and sadness also show fluctuating performance, but notably fear reaches 95.4% on EmoDB, showing that for certain datasets, the model captures these emotions well.
- Neutral tends to hold stable across sets, likely benefiting from its more distinct acoustic features.

Overall, while the model contributes to near-perfect in-distribution performance, the cross-corpus results highlight the need for further improvements in robustness and generalisability, particularly for emotions like disgust and joy. These findings reinforce the importance of designing SER models that can handle diverse and spontaneous speech conditions, critical for real-world applications such as telemedicine, where speaker diversity and recording variability are unavoidable.

Table 6 - Emotions Accuracy.

	Test set	EMOVO	RAVDESS	CaFE	CREMA-D	EmoDB
Emotion	Value	Value	Value	Value	Value	Value
Anger	94.80%	91.67%	80.21%	66.67%	51.70%	66.39%
Disgust	91.37%	44.05%	40.10%	62.50%	64.29%	70.45%
Fear	93.60%	63.10%	64.58%	81.94%	33.00%	95.38%
Joy	89.95%	46.43%	60.42%	31.25%	63.80%	90.00%
Neutral	96.41%	85.71%	38.54%	56.94%	76.50%	66.67%
Sadness	89.02%	48.81%	56.77%	64.58%	54.76%	79.03%
Surprise	99.58%	52.38%	81.25%	86.81%		

In our analysis, we have chosen a sub-group of 10 datasets from the original 32-dataset emoBOX benchmark because of access limitations on a few corpora and the unavailability of obtaining certain datasets. The selection allows fair model performance comparison across diverse datasets, though sampling a subset introduces a margin of error from the full benchmark. Margin of error (e) for a finite

population is derived by:

$$e = Z \times \sqrt{\frac{p(1-p)}{n} \times \frac{N-n}{N-1}}$$

where:

- **N** = 32 is the total number of datasets in the original benchmark,
- **n** = 10 is the number of datasets used in our evaluation,
- **Z** = 1.96 corresponds to a 95% confidence level,
- **p** = 0.5 represents the maximum variability assumption.

Using this formula, the margin of error in our study is 26.50%. So, although our results have important conclusions about model performance, the results must be taken with some caution since running the complete benchmark could result in somewhat different rankings. To mitigate this limitation, we are interested in relative model performance trends instead of absolute accuracy. Relative difference in performance attained within our chosen subset of dataset continues to be representative of overall trend and with given dataset diversity preserved.

From the comparison, we observe that: SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

- Our model (Wav2Vec 2.0 large) achieves a significant improvement over its base counterpart (+10.89% accuracy) due to its increased depth (24 layers) and parameter count (317M).
- The large models (ours, Data2Vec 2.0 large, HuBERT large and WavLM large) benefit from deeper architectures, with the latter achieving the highest accuracy (73.57%) due to speaker-aware pretraining.
- Whisper large V3 surpasses all models in accuracy (78.76%), but its 1.55B parameters make it significantly more resource-intensive, limiting its practicality for SER. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

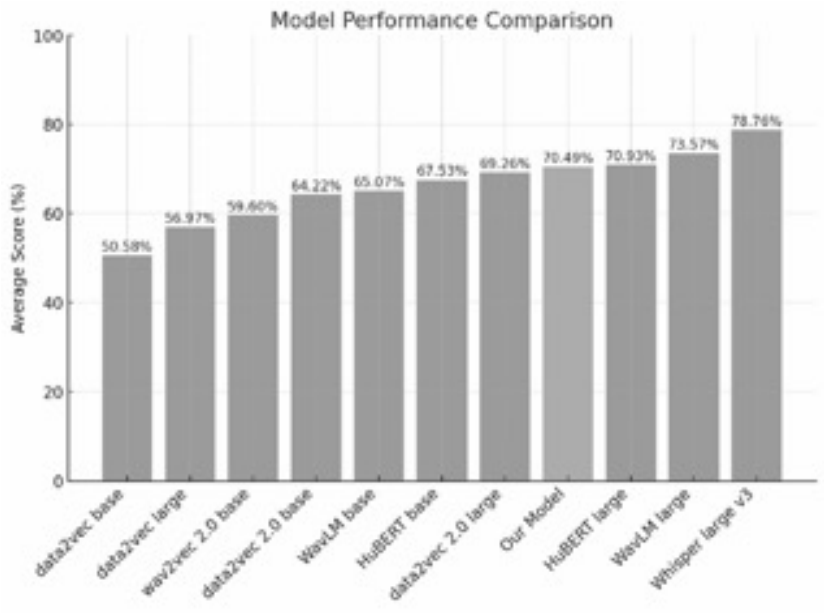

Figure 3 - emoBOX benchmark.

Discussion

In this paper, we supported that high-quality Transformer-based SER models, when tuned on high-quality Italian emotional speech corpora, can achieve high in-domain performance and maintain cross-corpus comparability. While the gains (+13% accuracy) capture the benefits of data selection, augmentation, and balanced training, the observations also reveal the residual challenges of generalisability across datasets.

Compared to previous work (e.g., Catania et al., 2025; Wang et al., 2021), our model shows stronger resistance on spontaneous and multilingual test sets, while emotions like disgust and happiness still struggle with lower generalisation, possibly due to underrepresentation or cultural divergence in expression. This is in line with larger trends in SER research, where emotions with more subtle acoustic markers fall behind in classification performance.

One of the advantages of our approach is the optimisation of performance and efficiency: although larger models (e.g., Whisper large V3) are more accurate in raw terms, they come at significant resource cost, limiting practical use. Our model achieves a more practical compromise, which is particularly attractive to sensitive, real-time applications such as telemedicine and remote mental state monitoring.

Nonetheless, there are limitations. Use of acted affective data, although beneficial for clear-cut labels, may limit ecological validity when applied to spontaneous, in-the-wild speech. Furthermore, the subset of emoBOX datasets used is within a margin of error, and results should be interpreted carefully before full-benchmark validation is obtained. We advise interpreting the current comparative results primarily as indicative rather than definitive. A full-benchmark evaluation would be essential to confirm whether the observed model rankings and performance gains generalise across the entire benchmark. Incorporating such comprehensive validation represents an important step for future work.

Looking forward, combining domain-adaptive techniques and multimodal signals (e.g., facial affect) can potentially further enhance robustness. Furthermore, exploring semi-supervised or few-shot learning methods can help models better generalise to low-resource or novel domains, a critical milestone toward scalable, emotion-sensitive healthcare technologies.

Conclusion

This study examined the performance of Transformer-based SER models on Italian data, focusing specifically on dataset preparation, model fine-tuning, and cross-corpus evaluation. Through more strict selection criteria and targeted augmentation to balance emotional classes, an improvement in performance by up to 13%, compared to the former method using the same base architecture, was attained. On our in-domain test set, the model achieved 93.56% accuracy, showing the usefulness of painstakingly collected Italian speech data. When evaluated on external datasets, however, like EMOVO, accuracy fell to 61.73%, testifying to the notorious domain mismatch problem. Cross-corpus testing on alternative datasets revealed our model to still be competitive, beating its base model and holding the performance of larger models, but without the computational expense of models such as Whisper large V3. These results underscore the importance of careful data preparation in closing the gap between experimental control and real-world deployment. By illustrating how performance varies between datasets, we have defined a framework for improving generalisability and mitigating the degradation characteristic of cross-corpus environments. In particular, the combination of data filtering, class-balanced representation, and targeted augmentations was determined to be instrumental in improving both in-domain accuracy and robustness to novel corpora. Building on this foundation, future research can combine domain-adaptive methods and multimodal data to continue to enhance the robustness of SER in various clinical and nonclinical settings, continuing to move us toward an

ultimately generalisable Speech Emotion Recognition algorithm.

References

Atila, O., & Şengür, A. (2021). Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition. Applied Acoustics, 182, 108260.

Atmaja, B. T., & Sasou, A. (2022). Effects of data augmentations on speech emotion recognition. Sensors, 22(16), 5941.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33, 12449-12460.

Bakhshi, A., Wong, A. S., & Chalup, S. (2020). End-to-end speech emotion recognition based on time and frequency information using deep neural networks. In ECAI 2020 (pp. 969-975). IOS Press.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005, September). A database of German emotional speech. In Interspeech (Vol. 5, pp. 1517-1520).

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE transactions on affective computing, 5(4), 377-390.

Catania, F., Wilke, J. W., & Garzotto, F. (2025). Emozionalmente: A Crowdsourced Corpus of Simulated Emotional Speech in Italian. IEEE Transactions on Audio, Speech and Language Processing.

Chattopadhyay, S., Dey, A., & Basak, H. (2020). Optimising speech emotion recognition using manta-ray based feature selection. arXiv preprint arXiv:2009.08909.

Costantini, G., Cesarini, V., & Casali, D. (2022). A Subset of Acoustic Features for Machine Learning-based and Statistical Approaches in Speech Emotion Recognition. In BIOSIGNALS (pp. 257-264).

Costantini, G., Iaderola, I., Paoloni, A., & Todisco, M. (2014). EMOVO corpus: an Italian emotional speech database. In Proceedings of the ninth international conference on language resources and evaluation (LREC'14) (pp. 3501-3504). European Language Resources Association (ELRA).

Dal Rí, F. A., Ciardi, F. C., & Conci, N. (2023). Speech emotion recognition and deep learning: an extensive validation using convolutional neural networks. IEEE Access, 11, 116638-116649.

Ekman, P. (1992). An argument for basic emotions. Cognition and Emotion, 6(3-4), 169–200.

Fanos, V., Dessì, A., Deledda, L., Lai, A., Ranzi, P., Avellino, I., ... & Colangelo, A. (2023). Postpartum depression screening through artificial intelligence: preliminary data through the Talking About algorithm. Journal of Paediatric and Neonatal Individualized Medicine, 12(2), 1-11.

Gournay, P., Lahaie, O., & Lefebvre, R. (2018, June). A canadian french emotional speech dataset. In Proceedings of the 9th ACM multimedia systems conference (pp. 399-402).

Grosman, J. (2021). Fine-tuned XLSR-53 large model for speech recognition in Italian. Hugging Face. https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-italian.

Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., & Khudanpur, S. (2017, March). A study on data augmentation of reverberant speech for robust speech recognition. In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5220-5224). IEEE.

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one, 13(5), e0196391.

Ma, Z., Chen, M., Zhang, H., Zheng, Z., Chen, W., Li, X., ... & Hain, T. (2024). Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark. arXiv preprint arXiv:2406.07162.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. Speech communication, 40(1-2), 227-256.

Wang, Y., Boumadane, A., & Heba, A. (2021). A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. arXiv preprint arXiv:2111.02735.

6. Publication 3

Paper Title: Effectiveness of a home-based computerised cognitive training in Parkinson's disease: a pilot randomised cross-over study

Status: Published http://dx.doi.org/10.3389/fpsyg.2024.1531688

Journal: Frontiers in Psychology

My Contribution: 30% - Conceptualisation, Experimentation, Methodology, Software, Writing (Draft, Review & Editing)

Effectiveness of a home-based computerised cognitive training in Parkinson's disease: a pilot randomised cross-over study

Abstract

Cognitive symptoms are common in Parkinson's Disease (PD), and digital interventions like telerehabilitation offer an accessible way to manage these symptoms. This study aimed to assess the effectiveness of a Home-Based Computerised Cognitive Training (HB-CCT) program in individuals with PD using a pilot randomised cross-over design. Twenty-five participants (mean age 69.32 ± 7.21 years, mean MDS-UPDRS III 33.76 ± 14.25) with PD and mild cognitive impairment were enrolled. They underwent neuropsychological assessments at three time points (5-week intervals): Baseline, after the HB-CCT, and after Standard Care. The HB-CCT consisted of the Neurotablet® platform that was used to target cognitive domains such as Attention, Memory, Perception, Executive Functioning and Language. All participants completed both the Neurotablet intervention and Standard Care blocks in a randomised order. After a Shapiro-Wilk test, non-parametric repeated measures analyse of variance (Friedman's test) and post-hoc comparisons corrected with the Bonferroni approach were performed to compare the effects on primary and secondary cognitive outcomes over experimental intervention and Standard Care. The results from the Friedman analysis revealed significant improvements in Word List Immediate Recall, Digit Span Forward and Complex Figure Recall (all $p<0.001$) following the HB-CCT, compared to the Baseline. Additionally, naming performance showed significant improvement after the HB-CCT ($p=0.02$). Significant differences were also observed when comparing the HB-CCT with Standard Care, with improved performance in TMT-A ($p=0.02$), Phonemic Fluency ($p<0.01$), and Digit Span Forward ($p<0.01$). These findings suggest that HB-CCT via Neurotablet can effectively enhance

specific cognitive abilities in PD, supporting the role of digital, home-based interventions as feasible strategies to mitigate cognitive decline.

Keywords: Rehabilitation, Telerehabilitation, Cognitive Training, Neuropsychology, PD.

Introduction

Parkinson's Disease (PD) is a progressive neurodegenerative disorder caused by a prominent loss of dopaminergic neurons in the substantia nigra pars compacta with the result of dopamine deficiency within the basal ganglia structures. The presence in the substantia nigra of aggregates of α-synuclein, known as Lewy bodies, is the neuropathological hallmark of the disease (Kalia et al., 2015). This process leads to a variety of motor symptoms (bradykinesia, muscular rigidity, rest tremor, and postural and gait impairment) (Buchman et al., 2012). However, PD is also associated with numerous non-motor symptoms (hyposmia, sleep disorders, depression, constipation), some of which precede the motor dysfunction (Schapira, Chaudhuri and Jenner, 2017).

Among non-motor symptoms, cognitive changes are frequently observed, even in the initial phases of the disease, with Mild Cognitive Impairment (MCI) affecting around 30-40% of individuals (Cosgrove, Alty and Jamieson, 2015). These cognitive changes can affect a person's independence and have a significant clinical impact, as it is related to institutionalization, mortality, and increased caregiver burden (Watson and Leverenz, 2010). Longitudinal studies have shown that approximately 50% of individuals with PD develop dementia after 10 years. Cognitive deficits may impact a person's autonomy and affect adherence to treatment due to an inability to understand the effects of medication or to follow a prescribed regimen (Bainbridge and Ruscin, 2009).

Impairment in different cognitive domains has been described in individuals with PD, affecting primarily attention and executive functions, memory, and visuo-spatial functioning (Verbaan et al., 2007; Aarsland et al., 2021; Wallace et al., 2022). Among attentional disorders, Allocock et al., (2009) showed that vigilance and reaction time, together with attentional fluctuations, are more frequently reported. Furthermore, several subcomponents of executive functioning, such as verbal fluency, planning, problem-solving, working memory, and set-shifting, are impaired in individuals with PD, reflecting frontostriatal damage. In addition, patients exhibit deficits in working memory, long-term memory, and

learning.

Interventions that target neuropsychological deficits could play a crucial role in enhancing overall quality of life. Several studies showed that individuals with PD may benefit from the verbal cue, suggesting that the new information is recorded but not readily accessible and that amnesia is mainly due to executive dysfunction (Emre, 2003), rather than a real dysfunction of the hippocampal structures. Moreover, visuo-spatial impairment involves both visuo-perceptual and visuo-motor abilities independently of cognitive decline (Girotti et al., 1988).

To date, pharmacological therapies have been the only available treatments that provide symptom relief. While pharmacological therapies are crucial in managing PD, significant limitations underscore the necessity for continued research and development of alternative treatment modalities. The current pharmacological treatments primarily alleviate motor symptoms, such as tremors, rigidity, and bradykinesia. However, they do not modify the disease's progression or significantly improve cognitive symptoms (Antonini et al., 2021). Pharmacological therapies addressing cognitive symptoms, such as cholinesterase inhibitors (e.g., rivastigmine and donepezil), have been associated with adverse drug reactions (Sun and Armstrong, 2021). Moreover, it should be noted that the response to pharmacological interventions can vary considerably between individuals with PD. Factors such as disease stage, severity, and the presence of comorbid conditions can influence the efficacy of treatment. Some of them may eventually require surgical interventions like deep brain stimulation when pharmacological treatments become less effective. However, not all patients are suitable candidates for surgery (Minafra et al., 2014; Servello et al., 2023). As of now, no treatment has been supported to halt or reverse the underlying neurodegenerative process. Consequently, non-pharmacological interventions that address the neuropsychological difficulties associated with PD may be pivotal in enhancing the overall quality of life of people living with this pathology (Sun and Armstrong, 2021).

Regarding specific interventions on cognitive functions, there is considerable inconsistency in the terminology used in the literature regarding cognitive stimulation, cognitive training, and cognitive rehabilitation for people presenting cognitive impairment. In particular, cognitive training and rehabilitation are often used interchangeably despite coming from different disciplines and having different objectives (Clare et al., 2003; Paggetti et al., 2024; Pinto et al., 2024). Cognitive rehabilitation aims to identify functional goals relevant to the person living with cognitive impairment and work

towards achieving them with the support of family members and/or caregivers. The emphasis is on improving or maintaining functioning in daily life, building on the person's strengths, and finding ways to compensate and/or sustain independence. Cognitive rehabilitation does not focus on improving cognitive function but addresses disability resulting from the impact of cognitive impairment on daily functioning and activities. Cognitive stimulation includes a series of activities and discussions (usually in a group) that aim to improve general cognitive and social functioning. Cognitive training involves guided practice on standardised paper-and-pencil or computerised cognitive tasks, with adaptable intensity and difficulty. It is based on a series of specific exercises and tasks designed to improve single or multiple cognitive functions and can be performed individually or in group sessions (Clare et al., 2003; Gavelin et al., 2020).

Despite the structural brain changes associated with the progression of neurodegenerative processes, cognitive training in PD has been shown to significantly increase functional brain connectivity and activation. This intervention resulted in improvements in cognition and functional disability, with long-term effects maintained for up to 18 months (Díez-Cirarda et al., 2018; Gavelin et al., 2022; Giustiniani et al., 2022). Growing evidence supports the benefits of cognitive intervention, yet individuals with PD still encounter many barriers to accessing rehabilitation services. Therefore their referral is suboptimal, likely due to skepticism regarding the value of intervention in the context of neurodegeneration (Battista et al., 2023; Pinto et al., 2024), the scarcity of sources in the healthcare care system of many countries (Balikuddembe and Reinhardt, 2020; Suárez-González et al., 2024), the lack of awareness regarding the role of the neuropsychological rehabilitation amongst referrers, and the geographical barriers that impede access to in-person cognitive rehabilitation services (Zaman, Ghahari and McColl, 2021). Furthermore, intensive and prolonged periods of training are emerging as crucial for chronic conditions, making it difficult to afford for all individuals (Vellata et al., 2021). These barriers may be mitigated by capitalizing upon alternative intervention modalities, such as telerehabilitation, an application of telemedicine that concerns the remote delivery of a variety of rehabilitative services through telecommunication technology (Piron et al., 2009), which has shown promising in treating individuals with PD (Vellata et al., 2021; Maggio et al., 2024).

Home-based teleneuropsychology enables individuals with comorbidities or motor disabilities to engage in cognitive training from home, enhancing their abilities and maintaining mental function through accessible, remote technology. Telerehabilitation offers flexibility in scheduling and is often more cost-

effective in terms of time and money. Compared to smartphones, tablet-based tools for teleneuropsychology are particularly accessible for older adults, thanks to user-friendly screens and clearly defined response areas. These platforms provide real-time feedback and automatic adjustments to match users' skill levels (Hammers et al., 2020; Naamanka et al., 2024). This approach places patients at the center of their rehabilitation, allowing them to view progress charts that objectively reflect their improvements. Continuous contact between patient and clinician ensures that the clinician's guidance remains a key factor in successful training outcomes. Notably, telerehabilitation has also been associated with patients' subjective perceptions of cognitive, emotional, and physical improvements (Mosca et al., 2020). Several studies have also supported the effectiveness of telerehabilitation treatments based on video games and virtual reality (Herz et al., 2013; Maggio et al., 2018). These methods have shown comparable effectiveness to face-to-face therapy in improving motor and non-motor symptoms and quality of life of individuals with PD (Cacciante et al., 2022). The integration of physical and cognitive functions stimulates the brain's reward system, increasing motivation and program adherence. In addition, telerehabilitation enables a larger group of individuals to work on a task at once, with fewer medical staff needed, and allows the clinician to monitor the progress in real-time (Vellata et al., 2021). It reduces time and costs and allows even daily intensive exercise while keeping the person in his social and physical environment (McCue, Fairman and Pramuka, 2010).

For the present study, we compared a new Home-Based Computerised Cognitive Training (HB-CCT) with Standard Care in PD patients, implementing a cross-over design. The experimental HB-CCT intervention was delivered with a platform named Neurotablet®, with the aim of evaluating its potential effectiveness in enhancing cognitive performance in individuals with PD.

Materials and Methods

Study design

We conducted a pilot cross-over randomised repeated measures study including two groups, with three testing time points (T0-T1-T2) at 5-week intervals (Figure 4). All participants underwent a neuropsychological assessment (T0) administered by expert neuropsychologists. After the first neuropsychological assessment (T0), participants were blindly allocated to Group 1 undergoing the experimental intervention or to Group 2 undergoing Standard Care. After 5 weeks, Group 1 and Group

2 patients returned to the Laboratory and underwent a second neuropsychological assessment (T1). At this point, the conditions for both groups changed. While Group 1 took Standard Care at home, Group 2 practiced the experimental intervention for 5 weeks. Finally, at the end of this period, all participants came again to the Hospital for the last neuropsychological examination (T2).

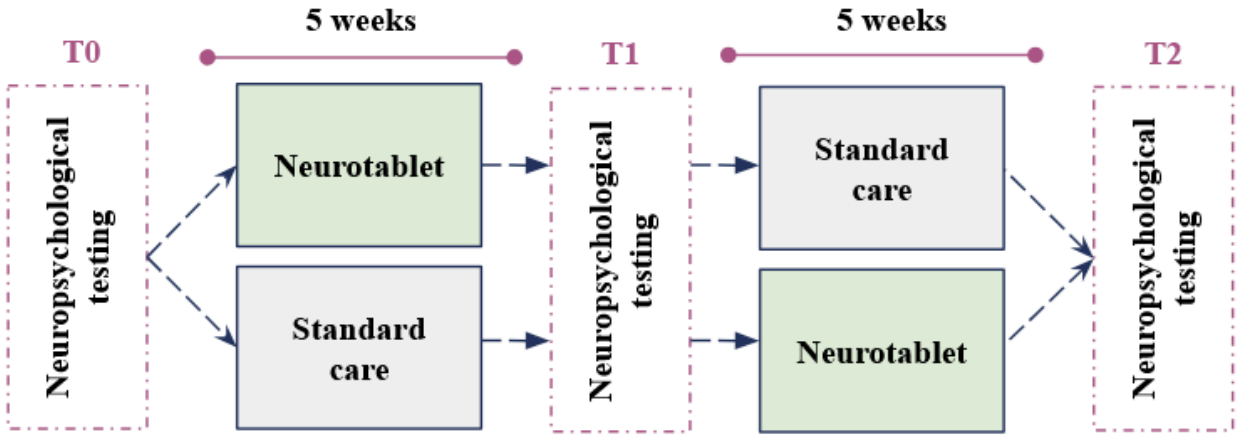

Figure 4 - The procedure of the pilot cross-over randomized study design.

Participants

Individuals who met the current clinical criteria for a PD diagnosis according to the UK Parkinson's Disease Society Brain Bank (Hughes et al., 1992) and presented with MCI diagnosed according to the Movement Disorders Society (MDS) criteria (Litvan, Goldman and Troster, 2012) were enrolled in this study upon providing their written informed consent. Other inclusion criteria were age between 40 and 85 years, at least 5 years of education (Primary School), and Hoehn and Yahr (H&Y) stage < 3. The exclusion criteria encompassed the presence of other neurological or psychiatric disorders, the presence of dementia as measured by the Montreal Cognitive Assessment (MoCA < 15.50) (Santangelo et al., 2015), a history of alcohol or drug abuse, undergoing a concomitant cognitive training treatment during the study period, changes in drug therapy during the study period, and the presence of uncorrected visual or auditory disturbances that may limit the administration of the test and/or treatment.

All of them completed the pilot randomised cross-over study design. The study protocol was approved by the Institutional Review Board of the IRCCS Giovanni Paolo II Hospital (No. Prot. 1195). Participants were recruited between March 2023 and March 2024 from the Laboratory of

Neuropsychology at the Clinical Scientific Institutes Maugeri of Bari, Italy. All participants were Italian speakers and functionally monolingual. Demographic data, including age, sex, and years of education, along with clinical data, including the disease duration, the levodopa (l-dopa) equivalent daily dosage (LEDD), the H&Y (Hoehn and Yahr, 2011) and the MDS-Unified Parkinson's Disease Rating Scale (MDS-UPDRS - part III), were collected from patients during the "ON" phase.

Thirty-eight individuals were screened throughout the study and twenty-five (7 female) participants were enrolled and completed both cross-over study blocks (attrition details are outlined in Figure 5). Participants had a mean age of 69.32 ± 7.21 years (range: 55-85) and a mean education of 13.00 ± 4.51 years. Additional information regarding demographic and clinical data is reported in Table S1 of Supplementary Materials.

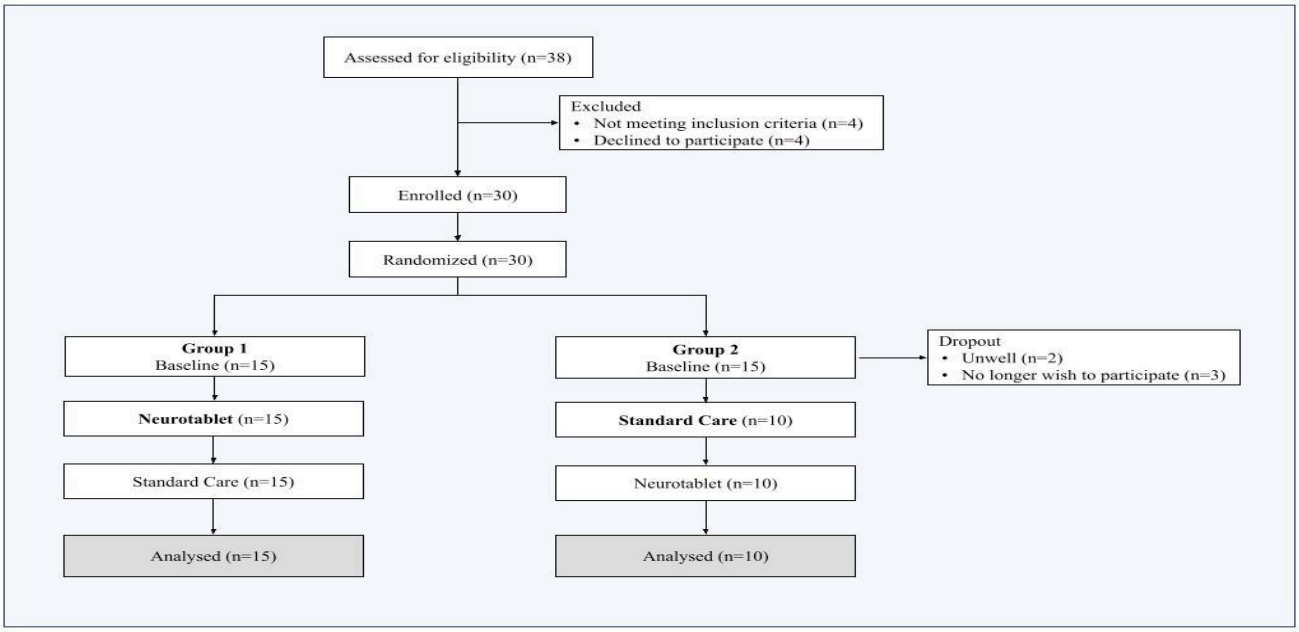

Figure 5 - Attrition details of the enrolment process.

Randomisation and blinding

A researcher blinded to participants' identities and not involved in enrolment or testing used a Randomisation minimization procedure (Altman and Bland, 2005) to allocate participants to therapy first (Group 1) or Standard Care (Group 2). To minimise potential differences between groups we took into account the severity of cognitive impairment (MoCA scores) at Baseline.

Intervention and Standard Care blocks

The intervention block consisted of 5 weeks of daily Home-Based Computerised Cognitive Training (HB-CCT), namely Neurotablet, with a target dose of 45 min/day and 2 tele-therapy sessions per week of 45 min to monitor patients' progression. The Standard Care block consisted of regular health advice. To take part in this study, all participants were provided with a web-based platform called Neurotablet, which was installed on a Samsung Galaxy tablet, as well as a stable internet connection in order to enable their participation in teletherapy sessions. Participants were instructed to register and log into a personal profile uploaded on the website platform. Neurotablet consisted of a multi-platform system which contained 40 different exercises and an amount of 10.000 customisable difficult levels. Exercises were classified according to the following cognitive domains: Attention, Memory, Perception, Executive Functioning, Language, and Neglect. The exercises employed in this study were specifically tailored to the cognitive profile exhibited by the participants at the neuropsychological examination. The level of difficulty was adapted for each patient by selecting the number of stimuli, color distractors, and trial numbers for each session. For each exercise and each level, thresholds were defined to allow progressively increasing difficulty levels. The time spent by participants on the HB-CCT was recorded by the app on a daily basis. An automated shut-down after 5 min of inactivity ensured high-fidelity data on the dose. Further details about the platform used for the intervention are reported in Supplementary Materials.

Outcome measures

The primary outcomes of the neuropsychological testing were measures of memory, attention, and executive functions. Specifically, we used the following tests to assess memory: Digit Span Forward (Monaco et al., 2015), Rey Auditory Verbal Learning Test (RAVLT) Immediate and Recall (Carlesimo et al., 1996), and the Rey–Osterrieth Complex Figure (ROCF) Recall (Caffarra, Vezzadini, Dieci, Zonato, et al., 2002). Attention and executive functions were assessed by Digit Span Backward (Monaco et al., 2015), Trail Making Test (TMT A-B; (Giovagnoli et al., 1996), Stroop test - Brief version (Caffarra, Vezzadini, Dieci and Zonato, 2002) and Phonemic fluency test (Carlesimo et al., 1996).

The secondary outcome measures included additional cognitive tests that assessed general cognitive

efficiency, as well as visuo-constructive, executive and linguistic abilities: MoCA, the Clock Drawing Test (CDT); (Caffarra et al., 2011), Rey–Osterrieth Complex Figure (ROCF) Copy (Caffarra, Vezzadini, Dieci, Zonato, et al., 2002), Category fluency test (Novelli et al., 1986) and the Screening for Aphasia in NeuroDegeneration (SAND; Catricalà et al., 2017;Battista et al., 2018).

To decrease possible learning effects, we used parallel versions of the MoCA (8.1, 8.2, 8.3). However, not all of the aforementioned cognitive tests have their respective parallel versions, therefore to minimise the learning effect we administered the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS; Randolph et al., 1998) forms A and B at T1 and T2. RBANS is a brief neuropsychological testing battery comprised of 12 subtests used to calculate five index scores: Immediate Memory Index (comprised of List Learning and Story Memory subtests), Visuo-spatial/Constructional Index (Figure Copy and Line Orientation subtests), Attention Index (Digit Span and Coding subtests), Language Index (Picture Naming and Semantic Fluency subtests), and Delayed Memory Index (List Recall, List Recognition, Story Recall, and Figure Recall subtests), and a Total scale score (comprised of all 12 subtests). Higher scores indicate better performance, both for subtests and index scores.

Table S2 provides a comprehensive overview of the tests employed in this study, describing each outcome measure.

Statistical analysis

The sociodemographic and clinical data of the enrolled patients are presented using descriptive statistics. Continuous variables are expressed as mean ± standard deviation (SD), while categorical variables as proportion (%). Neuropsychological measures are presented as median and interquartile range (IQR).

In order to ensure an accurate comparison of performance scores from disparate neuropsychological tests and batteries (including cognitive tests and the RBANS), a min-max normalization was performed. Each test may have a distinct range of scores, varying scales, and differing SD, which can lead to skewed comparisons and potentially misleading interpretations of the data. Min-max normalization is a data preprocessing technique that transforms the original score values into a standardized range, typically between 0 and 1. This approach allowed us to convert scores from different tests to a common scale.

This process ensures that all test results are standardized in terms of their range, allowing for direct comparisons. For example, if one test has a maximum score of 30 and another has a maximum score of 80, normalized scores can be expressed equally within the 0 to 1 range, making them directly comparable. Moreover, compared to the raw scores, the absolute numbers might bias analysis, as some tests may have inherently higher or lower score ranges. Normalizing the scores removes this bias, allowing the focus to be on the relative performance across tests rather than on the absolute scores. By applying min-max normalization, we prepared the data ensuring that no single test disproportionately influences the outcome. Normalized scores are also easier to interpret, as they can be considered as proportions of performance. This interpretation can help clinicians understand how an individual's scores compare to the maximum potential score of each test, providing insight into performance levels. Therefore, considering the specific domains, the following tests with different scoring ranges have been scaled: (1) Memory domain: Digit Span Forward (range: 3-9) and RBANS Digit Span (range: 0-16), RAVLT Immediate (range: 0-75) and RBANS List Learning (range: 0-40), RAVLT Recall (range: 0-15) and RBANS List Recall (range: 0-10), ROCF - Recall (range: 0-36) and RBANS Figure Recall (range: 0-20); (2) Visuo-constructive domain: ROCF - Copy (range: 0-36) and RBANS Figure Copy (range: 0-20); (3) Language domain: Naming SAND (range: 0-14) and RBANS Naming (range: 0-10).

Following the normalization and merging of selected tests and their parallel forms, the subsequent statistical analyses were conducted. To this end, tests grouped according to their respective cognitive domains have been hereinafter designated as follows: Digit Span Forward, Word List Immediate, Word List Recall, Complex Figure Recall, Complex Figure Copy, Naming. Tests not referenced herein were not subjected to standardisation, and the original nomenclature has been utilized.

The analysis involved repeated measures on the same individuals at three distinct time points: at Baseline, after HB-CCT Neurotablet intervention and after Standard Care. The neuropsychological outcomes at each assessment were treated as between-groups factors, where the groups were represented by the sample at Baseline, the sample after the experimental intervention (Neurotablet) and the sample after five weeks of Standard Care (Standard Care). A non-parametric one-way repeated measures analysis of variance, i.e. Friedman's test has been performed, to compare Baseline, Neurotablet and Standard Care. The output of the Friedman's test indicates whether there are statistically significant differences in the scores across the assessments, independently from the order of intervention delivery. Post-hoc comparisons were corrected with Bonferroni correction. If significant differences are found,

post hoc analyses can determine specifically which pairs of comparisons (e.g., Baseline vs. Neurotablet, Neurotablet vs. Standard Care, Baseline vs. Standard Care) are driving those differences. Descriptive analysis was used to evaluate Kendall's W effect size (ES) measures between time and the 95% confidence intervals (CI). The significance level adopted was 5% ($p<0.05$), with 95% confidence intervals. Data were analysed using the R Studio program version 2024.04.2.

Results

Patients were randomised to receive either the HB-CCT followed by the Standard Care (Group 1: n=15, age: 67.53 ± 7.26, 5 females, education: 13.53 ± 4.66, disease duration: 6.73 ± 6.15, LEDD: 456.53 ± 184.62, H&Y: 2.37 ± 1.09, MDS-UPDRS III: 32.27 ± 15.02), or Standard Care followed by HB-CCT (Group 2: n=10, age: 72.00 ± 6.58, 2 females, education: 12.20 ± 4.39, disease duration: 12.50 ± 9.73, LEDD: 720.00 ± 303.02, H&Y: 2.75 ± 0.95, MDS-UPDRS III: 36.00 ± 13.47). There were no significant differences between Group 1 and Group 2 for age, education, H&Y severity scale, and MoCA scores at Baseline, in line with the minimization Randomisation method ($p>0.05$). Further details are reported in Table S3.

Results from the Friedman's test on neuropsychological primary and secondary outcomes at Baseline, after the Neurotablet training and after Standard Care are displayed in Table 7.

Table 7 - Descriptive and Friedman's test analysis of neuropsychological.

	Baseline	**Neurotablet**	**Standard Care**					
Variables	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**	**p-value**	**Effect size* (95%CI)**	**p-value†**	**p-value†**	**p-value†**
						Baseline vs. Neurotablet	**Neurotablet vs. Standard Care**	**Baseline vs. Std. Care**
Primary outcomes								
Digit Span Forward	0.33 (0.33)	0.50 (0.12)	0.50 (0.17)	**< 0.001**	**0.45 (0.27 to 0.67)**	**< 0.001**	**< 0.01**	0.06
Digit Span Backward	4.00 (1.00)	4.00 (0.00)	4.00 (1.00)	**0.04**	**0.13 (0.03 to 0.37)**	0.10	0.28	1.00
Word List Immediate	0.37 (0.15)	0.55 (0.17)	0.48 (0.28)	**< 0.001**	**0.52 (0.29 to 0.76)**	**< 0.001**	0.24	**< 0.01**
Word List Recall	0.33 (0.20)	0.40 (0.40)	0.40 (0.40)	0.35	0.04 (0.003 to 0.28)	1.00	1.00	0.59
Stroop - Time	23.50 (17.50	23.00 (13.50)	28.00 (27.50)	0.17	0.047 (0.01 to 0.30)	0.69	0.15	1.00
Stroop - Errors	2.00 (7.00)	2.00 (3.00)	1.00 (5.00)	0.47	0.03. (0.003 to 0.23)	0.75	0.76	1.00

TMT-A	59.00 (45.00)	50.00 (50.00)	66.00 (83.00)	**< 0.01**	**0.20 (0.03 to 0.52)**	0.32	**0.02**	0.08
TMT-B	235.00 (378.00)	200.00 (203.00)	272.00 (350.00)	0.11	0.09 (0.01 to 0.32)	0.52	0.06	1.00
TMT B-A	174.00 (121.00)	150.00 (117.00)	185.00 (118.00)	0.72	0.01 (0.002 to 0.24)	1.00	1.00	1.00
Complex Figure Recall	0.26 (0.24)	0.50 (0.25)	0.40 (0.41)	**< 0.01**	**0.22 (0.09 to 0.45)**	**< 0.001**	0.67	0.17
Phonemic fluency test	34.00 (15.00)	37.00 (15.00)	30.00 (16.00)	**< 0.01**	**0.25 (0.05 to 0.57)**	0.13	**< 0.01**	**0.03**
Secondary outcomes								
MoCA	22.00 (6.00)	23.00 (5.00)	22.00 (4.00)	0.92	0.003 (0.001 to 0.16)	1.00	1.00	1.00
CDT	12.00 (5.00)	12.00 (7.00)	13.00 (7.00)	0.76	0.01 (0.001 to 0.17)	1.00	1.00	1.00
Complex Figure Copy	0.84 (0.28)	0.80 (0.20)	0.70 (0.30)	0.32	0.05 (0.003 to 0.26)	1.00	0.67	0.98
Category fluency test	15.00 (6.00)	11.00 (5.00)	12.00 (3.00)	**0.04**	0.13 (0.03 to 0.35)	0.19	1.00	**0.04**
Naming	1.00 (0.07)	1.00 (0.00)	1.00 (0.00)	**< 0.01**	**0.25 (0.12 to 0.46)**	**0.02**	1.00	0.20

*Kendall's W value
† Bonferroni correction

Notes. MoCA: Montreal Cognitive Assessment; CDT: Clock Drawing Test; TMT-A: Trail Making Test A; TMT-B: Trail Making Test B; TMT B-A: Trail Making Test B-A

Primary outcomes

The Friedman's test analysis revealed that, between the three assessments, individuals with PD showed significant differences in TMT-A ($p<0.01$, ES: 0.20, 95%CI 0.03 to 0.52), Phonemic fluency ($p<0.01$, ES: 0.25, 95%CI 0.05 to 0.57), Digit Span Forward ($p<0.001$, ES: 0.45, 95%CI 0.27 to 0.67), Digit Span Backward ($p=0.04$, ES: 0.13, 95%CI 0.03 to 0.37), Complex Figure Recall ($p<0.01$, ES: 0.22, 95%CI 0.09 to 0.45), and Word List Immediate ($p<0.001$, ES: 0.52, 95%CI 0.29 to 0.76).

After Bonferroni correction, significant differences between Baseline and Neurotablet for Word List Immediate ($p<0.001$), Digit Span Forward ($p<0.001$) and Complex Figure Recall ($p<0.001$) were found. Improved performances were supported after the experimental training, compared to Baseline. Bonferroni correction also showed significant differences between Neurotablet and Standard Care for TMT-A ($p=0.02$), Phonemic fluency ($p<0.01$) and Digit Span Forward ($p<0.01$). Finally, Bonferroni correction showed significant differences between Baseline and Standard Care for Word List Immediate ($p<0.01$) and Phonemic fluency ($p=0.03$).

Secondary outcomes

The Friedman's test analysis revealed that at the three assessments, individuals with PD showed significant differences in Category fluency ($p=0.04$, ES: 0.13, 95%CI 0.03 to 0.35) and Naming ($p<0.01$, ES: 0.25, 95%CI 0.12 to 0.46).

After Bonferroni correction significant differences between Baseline and Neurotablet for Naming ($p=0.02$) were found, which contributes to improved performance after the experimental training, compared to the Baseline. Concerning Neurotablet vs. Standard Care, we did not find any significant differences in other cognitive scores. It was found a significant difference in Category fluency ($p=0.04$) between Baseline and Standard Care.

Discussion

In recent years, the necessity to ensure the continuity of care at home has led to an increased emphasis on telemedicine and its potential applications in the field of neurorehabilitation. The present pilot randomised cross-over study was designed to evaluate the effectiveness of a new HB-CCT program delivered by the Neurotablet platform with respect to Standard Care in individuals with PD.

Primary outcomes

This study supported the positive effects of the HB-CCT in individuals with PD, indicating an enhancement in specific cognitive abilities. A statistically significant difference was observed between the HB-CCT Neurotablet intervention vs. Standard Care in three cognitive domains: verbal short-term memory, attentive capacities, and executive function skills. Specifically, we found an improvement in the primary outcomes Digit Span Forward, TMT-A, and Phonemic fluency tests. Few previous studies supported the effect of HB-CCT programs in individuals with PD prioritising the training of specific abilities, such as working memory (Edwards et al., 2013; Fellman et al., 2020; Ophey et al., 2020). Edwards et al., (2013) found an improvement of processing speed showing significant differences between post-training and Baseline, likewise our findings showed significant differences in processing speed between HB-CCT Neurotablet and Standard care. These studies suggest that the greater the degree

of focus on a specific cognitive domain in training, the greater the likelihood of achieving improvements in that domain (Gavelin et al., 2022).

Overall, the effects of CCT on cognitive functioning have been supported in several studies involving individuals with PD. París et al., (2011) evaluated the efficacy of a CCT on verbal short-term memory in individuals with PD, underlining a statistically significant enhancement in this ability following the completion of twelve forty-five-minutes supervised training sessions.. Promising results have also been reported for subjects with MCI who received the intervention individually at home. For example, Bahar-Fuchs et al., (2017) utilized the CogniFit software that works in an individually-tailored and adaptive way, similarly to the Neurotablet software. In line with our findings, the authors found improvements in composite measures of memory (including verbal short-term memory) immediately post-training, as well as at three-month follow-up. Moreover, a single-blinded, randomised control pilot study on community-dwelling MCI patients analysed the effects of a HB-CCT (Baik et al., 2024). It consisted of three times (around twenty-four minutes) a week sessions for eight weeks, and it was implemented with the software Neuro-World, which trained several cognitive functions including attention, visual perception, memory, and executive functions. In alignment with the findings of our study, the post-training intervention supported a notable enhancement in verbal short-term memory abilities when compared to the Baseline. We may conclude that the administration of CCT, whether in laboratory or home settings, may exert a beneficial impact on verbal short-term memory in patients with different neurodegenerative conditions.

We did not find significant differences in set-shifting neither between HB-CCT Neurotablet intervention and Standard Care, nor between HB-CCT Neurotablet intervention and Baseline. In line with our findings, previous studies investigating the effects of CCT on these domains in individuals with PD revealed no significant differences with the Baseline (Naismith et al., 2013; Ophey et al., 2020). In other studies, set shifting ability in trained individuals with PD was found to benefit from training also when compared to control participants (París et al., 2011; Alloni et al., 2018; Bernini et al., 2021). Given the difference in settings, one might speculate that performance in set shifting tests may be sensitive to the experimental setting (Guglietti, Hobbs and Collins-Praino, 2021). A controlled setting may facilitate patients in focusing on the task at hand, thereby enhancing the efficacy of the training programme.

Bernini et al., (2021) enrolled a group of PD-MCI patients who were trained with the CoRe software

These findings may support the hypothesis that attention and executive functions are the primary cognitive abilities affected in individuals with PD, which may slow down the efficacy of general domain training on various cognitive tasks in a laboratory setting. Similar to our findings, the authors did not obtain a significant result for working memory in the post-training period when compared to the Baseline. Ophey et al., (2020) employed a targeted training programme on working memory in a cohort of non-demented PD patients and observed no significant differences in working memory outcomes, either post-training or in comparison with the control group. One possible explanation is that the training was too challenging, exceeding the cognitive resources available, and therefore preventing successful performance immediately post-training.

Furthermore, when looking at the HB-CCT Neurotablet in comparison to Standard Care, a positive impact was observed with regard to Phonemic fluency. The existing literature on the effects of CCT on word production in individuals with PD has yielded inconclusive results (Alloni et al., 2018; De Luca et al., 2019; Bernini et al., 2021). Our findings suggest that Phonemic fluency is susceptible to the intervention as individuals with PD supported better performances after the HB-CCT, when compared to Standard Care. However, this gain is not evident when the post-HB-CCT scores are compared to the Baseline. These findings may support the hypothesis that attention and executive functions are the primary cognitive abilities affected in individuals with PD, which may slow down the efficacy of general domain training (Wallace et al., 2022). Nevertheless, this gain is greater when patients are exposed to a training targeting cognition. Interestingly, Phonemic fluency performance declined significantly following the administration of Standard Care, when compared to the Baseline outcome.

Concerning the comparison between HB-CCT Neurotablet intervention vs. Baseline in the primary outcomes, we found an improvement in learning ability, verbal short-term memory and visual long-term memory, namely in the following measures: Word List Immediate, Digit Span Forward and Complex Figure Recall.

In line with our findings, several studies have found significant improvements in the post-training period as compared to the Baseline period in learning trials that involved the presentation of verbal cues to memorize (Naismith et al., 2013; Petrelli et al., 2014). Instead, following the completion of the experimental intervention, we observed no significant result in long-term memory assessed by verbal tasks, in comparison to the Baseline condition. Literature showed positive effects of CCT on verbal long-

term memory only when compared to the Baseline, and not in comparison to other training (París et al., 2011; Petrelli et al., 2014). The discrepancy with the existing literature may be explained by the type of exercises implemented in Neurotablet. The training included a greater number of visual and verbal items to be recalled immediately than long-term memory exercises. Therefore, the type of exercises and of stimuli implemented in the training may influence the corresponding cognitive domain assessed in post-training. Moreover, consistently with our results, previous studies have identified post-training improvement with respect to the control condition in visual long-term memory, although none of these were HB-CCT (París et al., 2011; Alloni et al., 2018).

The present study revealed no significant difference in performance on the Digit Span Backward task (measure of working memory) between HB-CCT Neurotablet intervention vs. Baseline. Conversely, Fellman et al., (2020) supported that PD patients who underwent a five-week training programme (comprising three 30-minute sessions per week) exhibited a notable enhancement in working memory abilities, both in response to treated and untreated stimuli. One possible explanation for these results may stem from the emphasis of the training on practicing a specific cognitive domain.

Finally, when looking at the Standard Care vs. Baseline conditions we found worse performances in learning ability and executive functions, particularly in the following measures: Word List Immediate, Phonemic fluency, and Category fluency tests. Furthermore, we identified a reduction in verbal short-term memory. The implementation of an intervention that does not target cognitive abilities does not result in enhanced cognitive performance. These findings suggest that such an approach may not mitigate the neurodegenerative process. Consequently, there is a clear necessity for cognitive training to foster cognitive skills in PD.

Secondary outcomes

The analysis did not reveal significant differences in any of the secondary outcomes between HB-CCT Neurotablet intervention and Standard Care. However, when comparing HB-CCT Neurotablet intervention vs. Baseline, some significant results emerged. Interestingly, the participation in the Neurotablet training increased picture naming performance in individuals with PD. Neuropsychological testings pre- and post- CCT typically did not include the assessment of linguistic abilities. Thus far, only one study investigating the effects of a HB-CCT in PD used the Boston Naming Test to assess language

(Ophey et al., 2020). The authors revealed no improvement at the test after the training. However, a training specifically targeting working memory was used, differently from our study where a multiple domain training was included. It is noteworthy that few studies involved language exercises in the training (Clare et al., 2003; Gavelin et al., 2020). The improvement we found after Baseline may be attributed to the introduction of language training. Consequently, it may be valuable to incorporate such training into future CCT studies involving individuals with PD, given that cognitive impairments in this population frequently manifest in linguistic domains (Palmirotta et al., 2024).

A notable finding was that the analysis of global cognitive outcomes after the experimental training did not yield any statistically significant results when compared with Standard Care and with Baseline. A recent meta-analysis by Gavelin et al., (2022) empathized the effects of CCT on global cognitive efficiency, differently from prior meta-analyses (Leung et al., 2015; Orgeta et al., 2020). Multiple domain programs are more likely to be successful for global cognitive outcomes compared to programs targeting a single cognitive domain, whose effects tend to be most pronounced in the specific domains they target (Gavelin et al., 2022). The meta-analysis included both HB-CCT and CCT studies, using as global cognitive outcomes MoCA or MMSE. We may speculate that our results differ from those exposed by Gavelin et al., (2022) because they did not distinguish HB-CCT from laboratory based CCT. The setting of administration of the training has an impact on global performances (Guglietti, Hobbs and Collins-Praino, 2021). Also, it is well supported that distinct measurements for cognitive changes show different sensitivity in detecting cognitive decline (Biundo et al., 2016). The methodological differences between our study and existing literature make direct comparison of results challenging. As the number of HB-CCT studies on PD increases in future, it will become possible to make informed speculation about the effects of training on global efficiency.

In terms of the secondary outcomes, there was no evidence that visual spatial and constructive abilities benefited from the Neurotablet training. This finding is in line with previous literature (Ophey et al., 2020). Furthermore, a systematic review and meta-analysis revealed that visuo-constructive abilities supported the least benefit from CCT across all cognitive domains (Sanchez-Luengos et al., 2021).
The implementation of a multiple domain training programme, conducted in group sessions, yielded favorable outcomes in terms of Category fluency (París et al., 2011). Nevertheless, our study did not reveal any significant improvement in this ability following the experimental training, whether in comparison with the Baseline or with the post Standard Care. Similarly, Alloni et al., (2018) obtained

comparable outcomes when conducting CCT individually. It may be the case that the administration of training in groups could prove beneficial with regard to cognitive improvement. Factors such as the capacity to directly supervise participants in order to guarantee adherence and compliance, to furnish motivational support and encouragement, and to resolve IT issues as they arise, in addition to augmented social interaction for participants, are indispensable to enhance cognitive abilities, particularly Category fluency (Guglietti, Hobbs and Collins-Praino, 2021). Finally, a comparison between the Baseline and Standard Care groups revealed a decline in Category fluency. This may be attributed to the neurodegenerative process that is ongoing and not amenable to intervention through Standard Care.

Limitations and future directions

This study is not without limitations. A fundamental limitation of the cross-over study is the potential for carryover effects to obscure the impact of the training. In future studies, it would be advisable to incorporate a washout period between the two interventions, which could be useful to mitigate the risk of the carryover effect. Importantly, our findings cannot be generalized to items related to daily life. It would be crucial to explore the effects of the HB-CCT training on every-day activities. Furthermore, a limit of our findings is the lack of follow-up measures. Further research may address generalisation and aftereffects of adaptive HB-CCT in a PD population. The present work is a pilot study, therefore these results need to be supported in future research. A larger sample size would facilitate the generation of more robust results. Another issue may be addressed. All patients recruited for the study were aged 62 years or older. This cohort may be less familiar with tablets and technological platforms than the general population. It would be valuable to assess the level of satisfaction with the platform and the difficulties encountered in order to implement feedback and improve the platform's usability for an older population (Canini et al., 2014). Further research may be conducted to examine the evolution of these technologies, with the aim of developing new, tailored telerehabilitation solutions that address any significant challenges encountered with existing devices and platforms.

Conclusions

This pilot study contributes to that HB-CCT using Neurotablet may be an effective method for improving specific cognitive abilities in PD patients, including short-term verbal memory, long-term visual memory, and phonemic fluency, as compared to the standard care. Improvements were observed in

targeted areas, though results for other cognitive functions, such as processing speed and set-shifting, were mixed, with no significant differences between groups. These findings underscore the potential of HB-CCT as an innovative, accessible cognitive training tool suitable for home use, providing essential support in managing cognitive symptoms in neurodegenerative diseases.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Aarsland, D. et al. (2021) 'Author Correction: Parkinson disease-associated cognitive impairment', Nature reviews. Disease primers, 7(1), p. 53. Available at: https://doi.org/10.1038/s41572-021-00292-z.

Allcock, L.M. et al. (2009) 'Impaired attention predicts falling in Parkinson's disease', Parkinsonism & related disorders, 15(2), pp. 110–115. Available at: https://doi.org/10.1016/j.parkreldis.2008.03.010.

Alloni, A. et al. (2018) 'Evaluation of an ontology-based system for computerised cognitive rehabilitation', International journal of medical informatics, 115, pp. 64–72. Available at: https://doi.org/10.1016/j.ijmedinf.2018.04.005.

Altman, D.G. and Bland, J.M. (2005) 'Treatment allocation by minimisation', BMJ (Clinical research ed.), 330(7495), p. 843. Available at: https://doi.org/10.1136/bmj.330.7495.843.

Antonini, A. et al. (2021) 'Correction to: The TANDEM investigation: efficacy and tolerability of levodopa-carbidopa intestinal gel in (LCIG) advanced Parkinson's disease patients', Journal of neural transmission (Vienna, Austria), 128(6), pp. 863–865. Available at: https://doi.org/10.1007/s00702-020-02200-3.

Bahar-Fuchs, A. et al. (2017) 'Tailored and adaptive computerised Cognitive Training in older adults at risk for dementia: A randomised controlled trial', Journal of Alzheimer's disease: JAD, 60(3), pp. 889–911. Available at: https://doi.org/10.3233/JAD-170404.

Baik, J.S. et al. (2024) 'Effects of home-based computerised cognitive training in community-dwelling adults with mild cognitive impairment', IEEE journal of translational engineering in health and medicine, 12, pp. 97–105. Available at: https://doi.org/10.1109/JTEHM.2023.3317189.

Bainbridge, J.L. and Ruscin, J.M. (2009) 'Challenges of treatment adherence in older patients with Parkinson's disease', Drugs & aging, 26(2), pp. 145–155. Available at: https://doi.org/10.2165/0002512-200926020-00006.

Balikuddembe, J.K. and Reinhardt, J.D. (2020) 'Can digitization of health care help low-resourced countries provide better community-based rehabilitation services?', Physical therapy, 100(2), pp. 217–224. Available at: https://doi.org/10.1093/ptj/pzz162. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Battista, P. et al. (2018) 'Screening for aphasia in NeuroDegeneration for the diagnosis of patients with primary progressive aphasia: Clinical validity and psychometric properties', Dementia and geriatric cognitive disorders, 46(3-4), pp. 243–252. Available at: https://doi.org/10.1159/000492632.

Battista, P. et al. (2023) 'Access, referral, service provision and management of individuals with primary progressive aphasia: A survey of speech-language therapists in Italy', International journal of language & communication disorders, 58(4), pp. 1046–1060. Available at: https://doi.org/10.1111/1460-6984.12843. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Bernini, S. et al. (2021) 'A double-blind randomised controlled trial of the efficacy of cognitive training delivered using two different methods in mild cognitive impairment in Parkinson's disease: preliminary report of benefits associated with the use of a computerised tool', Aging clinical and experimental research, 33(6), pp. 1567–1575. Available at: https://doi.org/10.1007/s40520-020-01665-2.

Biundo, R. et al. (2016) 'MMSE and MoCA in Parkinson's disease and dementia with Lewy bodies: a multicenter 1-year follow-up study', Journal of neural transmission (Vienna, Austria), 123(4), pp. 431–438. Available at: https://doi.org/10.1007/s00702-016-1517-6.

Buchman, A.S. et al. (2012) 'Nigral pathology and parkinsonian signs in elders without Parkinson disease', Annals of neurology, 71(2), pp. 258–266. Available at: https://doi.org/10.1002/ana.22588.

Cacciante, L. et al. (2022) 'Cognitive telerehabilitation in neurological patients: systematic review and meta-analysis', Neurological sciences: official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology, 43(2), pp. 847–862. Available at: https://doi.org/10.1007/s10072-021-05770-6.

Caffarra, P., Vezzadini, G., Dieci, F., Zonato, F., et al. (2002) 'Rey-Osterrieth complex figure: normative values in an Italian population sample', Neurological sciences: official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology, 22(6), pp. 443–447. Available at: https://doi.org/10.1007/s100720200003.

Caffarra, P., Vezzadini, G., Dieci, F. and Zonato, F. (2002) 'Una versione abbreviata del test di Stroop: dati normativi nella popolazione italiana', Nuova rivista di neurologia, 12, pp. 111–115.

Caffarra, P. et al. (2011) 'Italian norms for the Freedman version of the Clock Drawing Test', Journal of clinical and experimental neuropsychology, 33(9), pp. 982–988. Available at: https://doi.org/10.1080/13803395.2011.589373.

Canini, M. et al. (2014) 'Computerised neuropsychological assessment in aging: testing efficacy and

clinical ecology of different interfaces', Computational and mathematical methods in medicine, 2014, p. 804723. Available at: https://doi.org/10.1155/2014/804723.

Carlesimo, G.A. et al. (1996) 'The mental deterioration battery: Normative data, diagnostic reliability and qualitative analyses of cognitive impairment', European neurology, 36(6), pp. 378–384. Available at: https://doi.org/10.1159/000117297.

Catricalà, E. et al. (2017) 'SAND: a Screening for Aphasia in NeuroDegeneration. Development and normative data', Neurological sciences: official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology, 38(8), pp. 1469–1483. Available at: https://doi.org/10.1007/s10072-017-3001-y.

Clare, L. et al. (2003) 'Cognitive rehabilitation and cognitive training for early-stage Alzheimer's disease and vascular dementia', Cochrane database of systematic reviews, (4), p. CD003260. Available at: https://doi.org/10.1002/14651858.CD003260.

Cosgrove, J., Alty, J.E. and Jamieson, S. (2015) 'Cognitive impairment in Parkinson's disease', Postgraduate medical journal, 91(1074), pp. 212–220. Available at: https://doi.org/10.1136/postgradmedj-2015-133247.

De Luca, R. et al. (2019) 'Computer assisted cognitive rehabilitation improves visuospatial and executive functions in Parkinson's disease: Preliminary results', NeuroRehabilitation, 45(2), pp. 285–290. Available at: https://doi.org/10.3233/NRE-192789.

Díez-Cirarda, M. et al. (2018) 'Neurorehabilitation in Parkinson's disease: A critical review of cognitive rehabilitation effects on cognition and brain', Neural plasticity, 2018, p. 2651918. Available at: https://doi.org/10.1155/2018/2651918.

Edwards, J.D. et al. (2013) 'Randomised trial of cognitive speed of processing training in Parkinson disease', Neurology, 81(15), pp. 1284–1290. Available at: https://doi.org/10.1212/WNL.0b013e3182a823ba.

Emre, M. (2003) 'What causes mental dysfunction in Parkinson's disease?', Movement disorders: official journal of the Movement Disorder Society, 18 Suppl 6(S6), pp. S63–71. Available at: https://doi.org/10.1002/mds.10565.

Fellman, D. et al. (2020) 'Training working memory updating in Parkinson's disease: A randomised controlled trial', Neuropsychological rehabilitation, 30(4), pp. 673–708. Available at: https://doi.org/10.1080/09602011.2018.1489860.

Gavelin, H.M. et al. (2020) 'Cognition-oriented treatments for older adults: A systematic overview of systematic reviews', Neuropsychology review, 30(2), pp. 167–193. Available at: https://doi.org/10.1007/s11065-020-09434-8.

Gavelin, H.M. et al. (2022) 'Computerised cognitive training in Parkinson's disease: A systematic review and meta-analysis', Ageing research reviews, 80(101671), p. 101671. Available at: https://doi.org/10.1016/j.arr.2022.101671.

Giovagnoli, A.R. et al. (1996) 'Trail making test: normative values from 287 normal adult controls',

Italian journal of neurological sciences, 17(4), pp. 305–309. Available at: https://doi.org/10.1007/bf01997792.

Girotti, F. et al. (1988) 'Dementia and cognitive impairment in Parkinson's disease', Journal of neurology, neurosurgery, and psychiatry, 51(12), pp. 1498–1502. Available at: https://doi.org/10.1136/jnnp.51.12.1498.

Giustiniani, A. et al. (2022) 'Effects of cognitive rehabilitation in Parkinson disease: a meta-analysis', Neurological sciences: official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology, 43(4), pp. 2323–2337. Available at: https://doi.org/10.1007/s10072-021-05772-4.

Guglietti, B., Hobbs, D. and Collins-Praino, L.E. (2021) 'Optimising cognitive training for the treatment of cognitive dysfunction in Parkinson's disease: Current limitations and future directions', Frontiers in aging neuroscience, 13, p. 709484. Available at: https://doi.org/10.3389/fnagi.2021.709484.

Hammers, D.B. et al. (2020) 'A survey of international clinical teleneuropsychology service provision prior to and in the context of COVID-19', The clinical neuropsychologist, 34(7-8), pp. 1267–1283. Available at: https://doi.org/10.1080/13854046.2020.1810323. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Herz, N.B. et al. (2013) 'Nintendo Wii rehabilitation ("Wii-hab") provides benefits in Parkinson's disease', Parkinsonism & related disorders, 19(11), pp. 1039–1042. Available at: https://doi.org/10.1016/j.parkreldis.2013.07.014.

Hoehn, M. and Yahr, M. (2011) 'Parkinsonism: Onset, progression, and mortality', Neurology, 77(9), pp. 874–874. Available at: https://doi.org/10.1212/01.wnl.0000405146.06300.91.

Hughes, A.J. et al. (1992) 'Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases', Journal of neurology, neurosurgery, and psychiatry, 55(3), pp. 181–184. Available at: https://doi.org/10.1136/jnnp.55.3.181.

Kalia, L.V. et al. (2015) 'Clinical correlations with Lewy body pathology in LRRK2-related Parkinson disease', JAMA neurology, 72(1), pp. 100–105. Available at: https://doi.org/10.1001/jamaneurol.2014.2704.

Leung, I.H. et al. (2015) 'Cognitive training in Parkinson disease: a systematic review and meta-analysis', Neurology, 85, pp. 1843-1851. Available at: https://doi.org/10.1212/WNL.0000000000002145.

Litvan, I., Goldman, J.G. and Troster, A.I. (2012) 'Diagnostic criteria for mild cognitive impairment in Parkinson's disease: Movement Disorder Society Task Force guidelines, Mov', Mov. Disord, 27, pp. 349–356.

Maggio, M.G. et al. (2018) 'What about the role of virtual reality in Parkinson disease's cognitive rehabilitation? Preliminary findings from a randomised clinical trial', Journal of geriatric psychiatry and neurology, 31(6), pp. 312–318. Available at: https://doi.org/10.1177/0891988718807973.

Maggio, M.G. et al. (2024) 'Effectiveness of telerehabilitation plus virtual reality (Tele-RV) in cognitive e social functioning: A randomised clinical study on Parkinson's disease', Parkinsonism & related disorders, 119, p. 105970. Available at: https://doi.org/10.1016/j.parkreldis.2023.105970.

McCue, M., Fairman, A. and Pramuka, M. (2010) 'Enhancing quality of life through telerehabilitation', Physical medicine and rehabilitation clinics of North America, 21(1), pp. 195–205. Available at: https://doi.org/10.1016/j.pmr.2009.07.005.

Minafra, B. et al. (2014) 'Eight-years failure of subthalamic stimulation rescued by globus pallidus implant', Brain stimulation, 7(2), pp. 179–181. Available at: https://doi.org/10.1016/j.brs.2013.12.011.

Monaco, M. et al. (2015) 'Erratum to: Forward and backward span for verbal and visuospatial data: standardisation and normative data from an Italian adult population', Neurol Sci, 36, pp. 345–347.

Mosca, I.E. et al. (2020) 'Analysis of feasibility, adherence, and appreciation of a newly developed Tele-rehabilitation program for people with MCI and VCI', Frontiers in neurology, 11, p. 583368. Available at: https://doi.org/10.3389/fneur.2020.583368.

Naamanka, E. et al. (2024) 'Effectiveness of teleneuropsychological rehabilitation: Systematic review of randomised controlled trials', Journal of the International Neuropsychological Society: JINS, 30(3), pp. 295–312. Available at: https://doi.org/10.1017/S1355617723000565.

Naismith, S.L. et al. (2013) 'Improving memory in Parkinson's disease: a healthy brain ageing cognitive training program', Movement disorders: official journal of the Movement Disorder Society, 28(8), pp. 1097–1103. Available at: https://doi.org/10.1002/mds.25457.

Novelli, G. et al. (1986) 'Tre test clinici di ricerca e produzione lessicale. Taratura su soggetti normali. [Three clinical tests to research and rate the lexical performance of normal subjects', Arch Psicol Neurol Psichiatr, 47, pp. 477–506.

Ophey, A. et al. (2020) 'Effects of working memory training in patients with Parkinson's disease without cognitive impairment: A randomised controlled trial', Parkinsonism & related disorders, 72, pp. 13–22. Available at: https://doi.org/10.1016/j.parkreldis.2020.02.002.

Orgeta, V. et al. (2020) 'Cognitive training interventions for dementia and mild cognitive impairment in Parkinson's disease', Cochrane database of systematic reviews, 2(2), p. CD011961. Available at: https://doi.org/10.1002/14651858.CD011961.pub2.

Paggetti, A. et al. (2024) 'The efficacy of cognitive stimulation, cognitive training, and cognitive rehabilitation for people living with dementia: a systematic review and meta-analysis', GeroScience [Preprint]. Available at: https://doi.org/10.1007/s11357-024-01400-z.

Palmirotta, C. et al. (2024) 'Unveiling the diagnostic potential of linguistic markers in identifying individuals with Parkinson's disease through artificial intelligence: A systematic review', Brain sciences, 14(2). Available at: https://doi.org/10.3390/brainsci14020137.

París, A.P. et al. (2011) 'Blind randomised controlled study of the efficacy of cognitive training in

Parkinson's disease', Movement disorders: official journal of the Movement Disorder Society, 26(7), pp. 1251–1258. Available at: https://doi.org/10.1002/mds.23688.

Petrelli, A. et al. (2014) 'Effects of cognitive training in Parkinson's disease: a randomised controlled trial', Parkinsonism & related disorders, 20(11), pp. 1196–1202. Available at: https://doi.org/10.1016/j.parkreldis.2014.08.023.

Pinto, J.O. et al. (2024) 'Ecological validity of neuropsychological interventions: A systematic review', Applied neuropsychology. Adult, pp. 1–20. Available at: https://doi.org/10.1080/23279095.2024.2328694.

Piron, L. et al. (2009) 'Exercises for paretic upper limb after stroke: a combined virtual-reality and telemedicine approach', Journal of rehabilitation medicine: official journal of the UEMS European Board of Physical and Rehabilitation Medicine, 41(12), pp. 1016–1102. Available at: https://doi.org/10.2340/16501977-0459.

Randolph, C. et al. (1998) 'The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): preliminary clinical validity', Journal of clinical and experimental neuropsychology, 20(3), pp. 310–319. Available at: https://doi.org/10.1076/jcen.20.3.310.823.

Sanchez-Luengos, I. et al. (2021) 'Effectiveness of cognitive rehabilitation in Parkinson's disease: A systematic review and meta-analysis', Journal of personalised medicine, 11(5), p. 429. Available at: https://doi.org/10.3390/jpm11050429.

Santangelo, G. et al. (2015) 'Normative data for the Montreal Cognitive Assessment in an Italian population sample', Neurological sciences: official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology, 36(4), pp. 585–591. Available at: https://doi.org/10.1007/s10072-014-1995-y.

Schapira, A.H.V., Chaudhuri, K.R. and Jenner, P. (2017) 'Non-motor features of Parkinson disease', Nature reviews. Neuroscience, 18(8), p. 509. Available at: https://doi.org/10.1038/nrn.2017.91.

Servello, D. et al. (2023) 'Complications of deep brain stimulation in Parkinson's disease: a single-center experience of 517 consecutive cases', Acta neurochirurgica, 165(11), pp. 3385–3396. Available at: https://doi.org/10.1007/s00701-023-05799-w. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Suárez-González, A. et al. (2024) 'Rehabilitation Services for Young-Onset Dementia: Examples from High-and Low-Middle-Income Countries', International Journal of Environmental Research and Public Health, 21(6). SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Sun, C. and Armstrong, M.J. (2021) 'Treatment of Parkinson's disease with cognitive impairment: Current approaches and future directions', Behaviouralsciences, 11(4), p. 54. Available at: https://doi.org/10.3390/bs11040054.

Vellata, C. et al. (2021) 'Effectiveness of telerehabilitation on motor impairments, non-motor symptoms and compliance in patients with Parkinson's disease: A systematic review', Frontiers in neurology, 12, p. 627999. Available at: https://doi.org/10.3389/fneur.2021.627999.

Verbaan, D. et al. (2007) 'Cognitive impairment in Parkinson's disease', Journal of neurology, neurosurgery, and psychiatry, 78(11), pp. 1182–1187. Available at: https://doi.org/10.1136/jnnp.2006.112367.

Wallace, E.R. et al. (2022) 'Meta-analysis of cognition in Parkinson's Disease mild cognitive impairment and dementia progression', Neuropsychology review, 32(1), pp. 149–160. Available at: https://doi.org/10.1007/s11065-021-09502-7.

Watson, G.S. and Leverenz, J.B. (2010) 'Profile of cognitive impairment in Parkinson's disease', Brain pathology (Zurich, Switzerland), 20(3), pp. 640–645. Available at: https://doi.org/10.1111/j.1750-3639.2010.00373.x.

Zaman, M.S., Ghahari, S. and McColl, M.A. (2021) 'Barriers to accessing healthcare services for people with Parkinson's disease: A scoping review', Journal of Parkinson's disease, 11(4), pp. 1537–1553. Available at: https://doi.org/10.3233/JPD-212735 SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

7. Publication 4

Paper Title: Analysing neonatal vocal expression: Methological approaches to identifying neurological and psychiatric signatures

Status: Published https://doi.org/10.56280/1703023560

Journal: Journal of Multiscale Neuroscience

My Contribution: 20% - Conceptualisation, Methodology, Investigation, Resources, Writing (Draft, Review & Editing), Supervision, Project Administration, Funding Acquisition.

Analysing neonatal vocal expression: Methodological approaches to identifying neurological and psychiatric signatures

Abstract

Analysing neonatal vocal expression provides invaluable insights into brain function and the emergence of consciousness, as early vocalisation patterns reflect neurodevelopmental trajectories and sensory integration processes. Despite progress in neonatal healthcare, identifying reliable neurological and cognitive markers from infant vocal sounds remains challenging, as it requires linking complex, multi-level brain activity with perceptual acoustic features. This paper reviews methodological approaches used to analyse neonatal vocal expressions, with a focus on techniques that bridge data-driven models with clinical applications. We examine computational methods, including signal processing, feature extraction algorithms, and machine learning models designed to capture vocal biomarkers of neurological or psychiatric disorders. Approaches include spectro-temporal analysis to detect atypical acoustic patterns, deep learning models like convolutional neural networks (CNNs) for automated feature learning, and explainable AI techniques that connect model outputs to clinically interpretable vocal features. We also explore multimodal approaches that combine vocal data with physiological and behaviouralsignals to improve diagnostic accuracy. The review addresses challenges in neonatal vocal analysis, including data scarcity, demographic variability, and the need for generalisation across different recording environments. To mitigate these issues, we highlight advances in domain adaptation, transfer learning, and data augmentation, which enable models to generalize across diverse clinical scenarios. We emphasize the need for clinical validation and interdisciplinary collaboration to ensure practical adoption of these models in healthcare. Future research should focus on refining predictive models with

larger, more diverse datasets and contributing to real-time analysis for continuous neonatal monitoring. By evaluating existing methodologies and proposing future directions, this study aims to advance neonatal vocal analysis and support early diagnosis and intervention in paediatric healthcare.

Keywords: Neonatal vocal expression, neurological and psychiatric signatures, signal processing, machine/deep learning, explainable AI, paediatric healthcare

Introduction

Analysing neonatal vocal expression represents a frontier in developmental neuroscience and paediatric healthcare, offering a unique and noninvasive pathway for understanding the early architecture of the human brain (Andonotopo et al., 2025). Neonatal vocalisations, beginning with the very first cries after, are far more common than primitive reflexive sounds; they are rather complex, biologically orchestrated signals that emerge from the intricate coordination of the neurological, respiratory, and cognitive systems (Romo et al., 2024; Shah et al., 2025). From an evolutionary perspective, these early sounds have played a crucial role in securing caregiver attention, signaling physiological needs, and promoting social bonding. Today, they are increasingly recognised as rich behaviouralbiomarkers that can offer critical insight into an infant's sensory processing capacity, emotional regulatory mechanisms, and early motor control (Filippa & Kuhn, 2024). This multidimensional nature makes them valuable proxies for assessing the integrity of the developing central nervous system, with immense potential to inform early diagnosis of neurodevelopmental and neuropsychiatric conditions.

Over the past decade, rapid advancements in computational neuroscience and biomedical engineering have fueled interest in harnessing infant vocalisations as early warning signals for atypical brain development (Onciul et al., 2025). Researchers are exploring whether subtle deviations in cry acoustics or cooing patterns could serve as preclinical indicators for conditions such as autism spectrum disorder, cerebral palsy, or language impairments well before such conditions manifest in overt behaviouralsymptoms. Early identification is particularly critical, as timely therapeutic interventions during periods of peak neuroplasticity can significantly improve cognitive and behaviouraloutcomes later in life. However, despite this exciting promise, informing these raw vocal outputs into actionable, clinically relevant information poses formidable scientific and technical hurdles (Husain et al., 2025).

One major challenge stem from the inherent variability and fleeting nature of neonatal vocalisations. Unlike adult speech, which follows structured phonetic and linguistic patterns, infant cries are brief, highly context- dependent, and easily influenced by environmental stimuli or physiological states (Kao & Zhang, 2025; Nussbaum et al., 2025). Furthermore, inconsistencies in recording environments, differences in microphone quality, and demographic variability, such as language or cultural factors, add layers of complexity, introducing confounding variables that can degrade model accuracy and generalizability across populations. Extracting meaningful patterns from these signals requires not only robust signal processing techniques but also advanced computational models capable of distinguishing noise from neurologically meaningful variation.

Moreover, bridging the gap between sophisticated computational models and real-world clinical practice necessitates methodological rigour and interpretability. Healthcare professionals need tools that not only achieve high predictive accuracy under controlled laboratory conditions but also maintain robustness and transparency in the noisy, dynamic context of neonatal intensive care units (NICUs) or home-based monitoring settings (Sheikh et al., 2025). This calls for explainable Artificial Intelligence (AI) frameworks that clarify how models reach diagnostic conclusions, fostering trust and acceptance among clinicians and caregivers alike. Additionally, integrating vocal analysis with other complementary data streams, such as physiological signals (e.g., heart rate variability), EEG patterns, or contextual behavioural observations, can significantly enhance diagnostic precision and provide a more holistic understanding of an infant's health status.

This review aims to provide a comprehensive examination of the state-of-the-art methodologies in neonatal vocal analysis, outlining how signal processing, acoustic feature extraction, and Machine Learning intersect to address the unique demands of this sensitive domain. We survey traditional techniques for denoising, segmentation, and feature extraction before delving into more recent advances such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and attention-based models that automate and refine the learning of complex acoustic patterns. A critical focus is placed on explainable AI tools that illuminate model decision pathways, as well as strategies for domain adaptation, transfer learning, and data augmentation that help mitigate data scarcity and promote generalizability. In recognition of the limitations of single-modality approaches, we further explore emerging multimodal frameworks that combine vocal signals with physiological, behavioural, and environmental data to construct a more comprehensive, context-aware portrait of neonatal

neurodevelopment. This reflects a broader trend in paediatric medicine toward systems- level analysis and personalised care pathways. Finally, we address the practical and ethical considerations that accompany the clinical translation of these tools, highlighting the importance of interdisciplinary collaboration, rigorous validation, and attention to data privacy. By synthesising these diverse threads, this review not only charts the current landscape but also identifies promising avenues for future research and clinical application. Ultimately, by advancing robust, interpretable, and contextually integrated vocal biomarkers, the field moves closer to realising the vision of early, accessible, and precise neurodevelopmental monitoring, informing how we detect, understand, and respond to risk in the earliest stages of human life.

To explore this promising domain comprehensively, Section 2 outlines the neurodevelopmental significance of neonatal vocalisations, while Section 3 discusses the key challenges inherent in analysing such delicate data. Section 4 outlines the signal processing and acoustic feature extraction methods, followed by Section 5, which explores machine learning approaches specifically designed for vocal analysis. Section 6 highlights the benefits of multimodal and integrative frameworks, and Section 7 addresses methodological challenges and proposed solutions. Section 8 examines clinical translation and validation pathways, while Section 9 explores future directions for research and application.

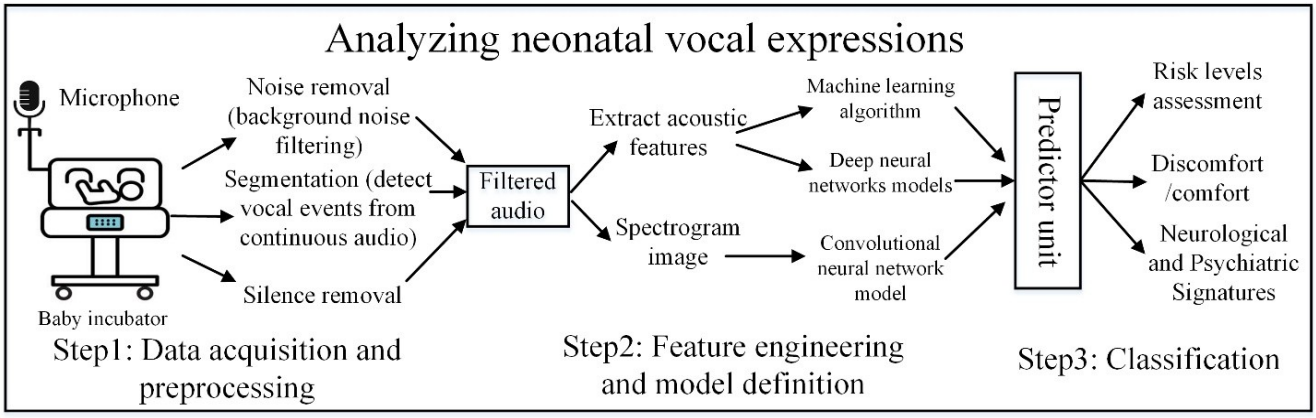

Figure 6 - A general conceptual workflow illustrating the typical stages of neonatal vocal analysis.

Finally, Section 10 provides concluding remarks, summarizing key insights and their implications for advancing neonatal neurodevelopmental care.

Neurodevelopmental significance of neonatal vocalisations

Neonatal vocalisations serve as one of the earliest and most accessible manifestations of neural activity in the developing brain (Zhang, 2025). These vocal expressions, including cries, coos, and proto-speech sounds, arise from complex coordination between the central nervous system, respiratory system, and vocal tract. Even in the absence of fully developed cognitive or linguistic abilities, infants produce vocal patterns that are shaped by underlying neural circuits involved in motor planning, auditory feedback, and affective regulation (Wang & Song, 2022). The structure, frequency, pitch, and rhythm of these early sounds are not arbitrary; they reflect maturational processes in brain regions such as the brainstem, limbic system, and auditory cortex (Cappelli & Noccetti, 2022). As such, careful analysis of neonatal vocal output provides a non-invasive proxy for assessing neurological development, offering insights into both typical and atypical trajectories of brain maturation.

In the context of recent advances, a typical framework for analysing neonatal vocal expressions can be conceptually outlined in three core stages: data acquisition and preprocessing, feature engineering with model definition, and final classification or risk assessment, as outlined in Figure 6. In this approach, audio signals are first captured in clinical environments such as neonatal intensive care units using microphones placed near incubators. Rigorous preprocessing steps, including background noise removal, silence trimming, and segmentation of relevant vocal events, are employed to ensure that only high-quality, meaningful audio segments are retained. Subsequently, acoustic features are extracted and, where appropriate, spectrogram images are generated to visually represent the temporal and spectral properties of the cry sounds. These features then serve as inputs to various computational models, ranging from traditional ML algorithms to more advanced deep neural networks and convolutional architectures. The processed outputs enable classification tasks that support clinicians in assessing discomfort levels, identifying possible neurological or psychiatric markers, and estimating overall risk. As illustrated in the diagram, this multi- step pipeline reflects best practices commonly reported in the literature, offering a structured perspective on how interdisciplinary tools can be integrated to unlock the diagnostic potential of infant vocalisations.

Early vocalisations are more reflexive acts than true indicators of an infant's ability to interact with and respond to the sensory environment (Filippa & Kuhn, 2024). The emergence of patterned, intentional vocal output signals the gradual development of conscious awareness and the infant's capacity to perceive, process, and react to external stimuli. This process involves the integration of multimodal

sensory inputs such as touch, vision, and hearing with internal motor and affective states (Sanna, 2025). The dynamic interplay between sensation and expression facilitates the infant's engagement with caregivers and surroundings, which in turn supports socioemotional bonding and cognitive stimulation (La Rosa et al., 2024). Vocalisations, therefore, represent a bridge between the internal neurological state and external behavioural expression, contributing to researchers to track the formation of consciousness and early perceptual motor coordination through acoustic analysis (Rudenko, 2023).

The clinical implications of neonatal vocal expression are profound, particularly in the early detection of neurological and psychiatric disorders. Atypical vocal characteristics such as abnormal pitch, monotonic cries, prolonged silence, or irregular prosody can be early indicators of conditions such as autism spectrum disorder (ASD), cerebral palsy, or perinatal brain injury (Filippa et al., 2021; Marschik et al., 2022a). Numerous studies (Bartl-Pokorny et al., 2022; Long et al., 2023; Marschik et al., 2022b; Wagner et al., 2025) have documented how infants with neurological impairments exhibit vocal patterns distinct from those of their typically developing peers. These differences often precede observable behavioural symptoms, positioning vocal analysis as a valuable tool for early risk assessment. In psychiatric contexts, alterations in vocal expression may signal disruptions in affective processing and social communication, domains commonly affected in disorders such as ASD or early- onset mood disorders (Ding and Zhang, 2023; Kamiloğlu and Sauter, 2021; Ribolsi et al., 2022). As such, neonatal vocal analysis not only supports early diagnosis but also holds promise for tracking developmental progress and treatment outcomes over time.

Challenges in analysing neonatal vocal data

To support robust research on infant vocalisation, cry detection, snoring recognition, and pain assessment, this study leverages a diverse collection of datasets, each contributing unique acoustic contexts and annotation standards (see Table 8 for a comprehensive summary). Among these, AudioSet (Gemmeke et al., 2017) stands out as a massive benchmark of over two million human-annotated YouTube clips, from which we specifically extracted categories relevant to infant cries and snoring events, using both weakly and strongly labelled subsets. The Baby Chillanto Database (BCD) (Reyes-Galaviz et al., 2008) provides a pathology-focused corpus of short infant cry samples categorized into clinically meaningful conditions such as asphyxia, deafness, hunger, normal, and pain cries, facilitating nuanced classification tasks. Donate A Cry (Veres, 2025) complements this by focusing on the emotional

and need-driven aspects of baby cries, covering categories such as hunger, burping, belly pain, discomfort, and tiredness, thereby contributing to models to map vocal cues to daily care needs.

The environmental and ambient noise contexts are captured through the ESC-50 dataset (Piczak, 2015), which we filter for infant crying and snoring sounds to augment the training data with real-world variability. Building on these public resources, the Infant Cry and Snoring Detection (ICSD) dataset (Liu et al., 2025) integrates samples from eight source datasets, which are systematically cleaned and balanced into weakly labelled, strongly labelled, and synthetic event clips specifically tailored for our detection models. Similarly, the DSPLab Baby Sounds challenge dataset (Alexlinander, 2022) provides additional labelled baby cry clips designed for benchmarking machine learning pipelines via MFCCs and SVMs. For pain detection, the Infant FLACC Pain Level Video Dataset (IFPaLVD) (Kristian et al., 2023) offers carefully annotated audio recordings of infants assessed for both crying and pain severity via the widely accepted FLACC scale, enriching our ability to study the acoustic correlates of distress and discomfort.

In addition to traditional audio datasets, we include rich, ego-centric multimodal corpora that capture infant and child behaviour in naturalistic contexts. BV-Home (Long et al., 2024) comprises more than 400 hours of daily life home recordings from 28 families, capturing spontaneous infant vocalisations, interactions, and environmental context alongside parent-reported language measures. BV-Preschool extends this perspective into early education settings, documenting child speech and interactions within a Montessori-inspired preschool environment. Ego-SingleChild, also from the same research initiative, provides longitudinal recordings from a single child via a wearable headband camera, whereas SAYCam (Sullivan et al., 2021) offers a similar head camera view spanning 476 to examine the relationship between early language experience and vocal development.

Table 8 - Comprehensive overview of all datasets employed in this study.

Dataset	Description	Categories Used in Research	Total Clips Used	Clip Duration	Additional Notes
AudioSet (Gemmeke et al., 2017)	2 million + human- annotated YouTube audio clips with 632 event classes; includes weakly and strongly labelled subsets	Infant Cry, Snoring	Weakly labelled: 1391 (Cry), 1713 (Snoring); Strongly labelled: 424 (Cry), 383 (Snoring)	10 sec	Strong labels include timestamps for events; hierarchical ontology

Baby Chillanto (BCD) (Reyes- Galaviz et al., 2008)	Mexican database for infant cry pathology classification; 5 pathology classes	All 5 infant cry categories	2,268 samples	1 sec	Categories: asphyxia, deaf, hunger, normal, pain
Donate A Cry (Veres, 2025)	Collected from 0–2-year- old babies; designed for infant need recognition	Hunger, Burping, Belly Pain, Discomfort, Tiredness	457 files	7 sec	Cleaned and categorized for infant need recognition
ESC-50 (Piczak, 2015)	2,000 environmental sounds across 50 classes and 5 main categories	Infant Cry, Snoring (extracted only)	Subset extracted	5 sec	Includes 10 classes per category across 5 main sound groups
ICSD Dataset (Liu et al., 2025)	Proposed Infant Cry and Snoring Detection (ICSD) dataset built from 8 unified source datasets; used for event detection task	Infant Cry, Snoring	8,000 (train), 1,000 (validation/test); plus: 1,699 Cry & 1,577 Snoring weakly labelled (train); 338 Cry & 305 Snoring real strongly labelled (train)	10 sec	Includes weakly labelled, real strongly labelled, and synthetic strongly labelled clips; test set excludes weak labels; fully cleaned and standardized
DSPLab: Detecting Baby Sounds (Alexlinan der, 2022)	Kaggle competition dataset to classify baby sounds using MFCCs and SVM (or other models)	Baby sound types (not explicitly listed)	3,996 labelled clips (train) + dev set	Not specifi ed	F1 score used for evaluation
IFPaLVD (Infant FLACC Pain Level Video Dataset) (Kristian et al., 2023)	Pain and cry assessment dataset collected at Dr. Soetomo General Hospital using FLACC scale	Cry/No Cry, Pain levels: Neutral, Discomfort, Mild, Moderate, Severe	253 audio recordings	Not specifi ed	23 infants (<1 year); pain labels based on tuple (pain level, cry); 5 pain categories derived from tuple combinations
BV-Home (Long et al., 2024)	Home recordings of infant-toddler daily life collected from 28 families (avg. child age 11 months); includes ego-centric video, audio, transcripts, and motion data	Infant vocalisations in naturalistic contexts	Not explicitly clips, but ~433 hours of recordings	Varies	Ego-centric, long-form recordings; parent-reported language development; audio, transcript, motion data available
BV-Preschool (Long et al., 2024)	Egocentric preschool recordings in Montessori-like setting	Child vocalisations and interactions	Not explicit clips; ~63 hours	Varies	39 children (2.11–5.11 years); play-based learning
Ego- SingleChil d (Long et al., 2024)	Frequent recordings of a single infant with headband camera	Single child daily vocalisations	Not explicit clips; 47 hours	Varies	High continuity; alternative camera; lower resolution
SAYCam (Sullivan et al., 2021)	Longitudinal head- camera recordings of daily infant life	Infant language & visual experience	Not explicit clips; 476 hours	Varies	3 infants; focus on language acquisition context

Despite the breadth and depth of these datasets, analysing neonatal vocal data poses persistent challenges. First, the limited availability of high- quality, consistently annotated infant recordings

remains a bottleneck owing to ethical restrictions, the fragile nature of neonatal subjects, and practical constraints in clinical environments such as NICUs (Keles & Bagci, 2023). Consequently, many studies rely on relatively small or demographically narrow samples, limiting the statistical power and generalizability of the resulting models (Frank, 2020). Furthermore, inconsistent recording protocols, varying microphone placements, and environmental noise, such as hospital equipment hums or household disturbances, introduce significant acoustic variability, which can degrade feature extraction and classification accuracy (Mallegni et al., 2022).

Moreover, neonatal vocalisations are inherently influenced by biological and social factors such as age, gestational maturity, cultural background, and health status (Hou et al., 2024). Premature infants, for example, may vocalize differently than their full-term peers because of developmental differences in respiratory control or neurological function. Without careful inclusion of diverse demographic groups, models risk embedding bias and may fail to detect anomalies accurately across varied populations (Meissen et al., 2024). Finally, even well-tuned models

often struggle to maintain their performance when transferred from controlled research conditions to real- world settings, such as busy hospitals or remote home monitoring. Variations in recording devices, background activity, and infant states (e.g., feeding, sleeping, or interacting) can shift the acoustic profile of vocalisations, demanding robust techniques such as domain adaptation and transfer learning to bridge this gap.

Signal processing and acoustic feature extraction

The foundation of computational infant cry analysis lies in informing raw audio into mathematically tractable representations that capture both the spectral (frequency-related) and temporal (time-varying) properties of vocalisations (Fu et al., 2025). A range of signal processing techniques, including cepstral analysis, wavelet transforms, zero-crossing rates, energy measures, and image-based time series encoding, are used to extract distinctive features that encode the subtle dynamics of neonatal cries, coos, and atypical vocal behaviour.

Table 9 summarizes how these methods have been systematically applied in recent studies, each employing specialised tools to ensure precise and consistent feature extraction.

Cepstral features: MFCC and GFCC

Mel-frequency cepstral coefficients (MFCCs) (Ali et al., 2021) are among the most widely adopted acoustic features for cry analysis. They approximate how the human cochlea perceives sound by mapping the power spectrum of short overlapping frames onto the Mel scale, a scale that reflects human auditory sensitivity to frequency. MFCCs are computed by applying a Fourier transform to windowed frames, mapping the result through triangular Mel filter banks, taking the logarithm of the power at each filter, and then performing a discrete cosine transform (DCT) to decorrelate the coefficients.

In Dey et al. (2025), MFCCs were extracted from the Baby Chillanto dataset (1049 normal and 340 asphyxia cries). The audio data were resampled to 8 kHz, segmented into 1-second windows, denoised, balanced with random oversampling, and framed before MFCC extraction. They visualized MFCCs and structured the data as Pandas DataFrames for downstream processing via Librosa, a popular Python library for audio analysis. Similarly, Ozcan and Gungor (Ozcan & Gungor, 2025) used 13-dimensional MFCCs on the Donate Cry dataset, applying robust data augmentation to diversify training data for their structure-tuned artificial neural network. Kumar Nukala et al. (2024) and Hammoud et al. (2024) combined MFCCs with other metrics in a multidomain framework to encode short-term spectral envelopes alongside other cues.

Expanding beyond MFCCs, Zayed et al. (2023) employed gammatone frequency cepstral coefficients (GFCCs). Unlike MFCCs, GFCCs use a gammatone filter bank, which models human auditory filters more accurately than triangular Mel filters do, especially in noisy conditions. GFCCs were extracted via MATLAB scripts, complementing prosodic and image-based features to construct a robust, fused representation.

Prosodic features: harmonic ratio

The harmonic ratio (HR) (Bellanca et al., 2013) quantifies the proportion of periodic (harmonic) energy relative to total signal energy, serving as a measure of voice periodicity and phonatory control. In infants, HR helps capture breath support and vocal fold vibration stability. Zayed et al. (2023) extracted HR using MATLAB, integrating it with GFCCs and spectrogram-derived features for multidomain fusion.

Spectrograms and spectro-temporal representations

A spectrogram visualizes how signal energy is distributed over frequency and time, offering a time–frequency representation useful for highlighting transitions, pitch modulations, and noise bursts. It is typically computed via the short-time Fourier transform (STFT), which divides the signal into short overlapping frames and computes a frequency spectrum for each.

Zayed et al. (2023) generated spectrogram images in Python and extracted deep features via a pretrained VGG16 CNN. This allowed the capture of intricate spectral patterns that simpler features might overlook. In multidomain setups, the spectrogram serves as a high-dimensional, image-based descriptor fused with GFCC and HR. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Kumar Nukala et al. (2024) and Hammoud et al. (2024) extended this by using Mel-Spectrograms, which compress frequency bins onto the Mel scale, further aligning them with human perception. This is especially beneficial when analysing cries, where high- pitched harmonics and formants convey critical diagnostic information.

Zero-crossing rate (ZCR) and root mean square energy (RMS)

The zero-crossing rate (ZCR) (Joo et al., 2021) is the rate at which the signal waveform crosses the zero-amplitude axis. This indicates the noisiness or tonal quality of a signal: voiced sounds tend to have a lower ZCR, whereas noisy or unvoiced segments have a higher ZCR. The RMS energy quantifies the signal's average power, reflecting the loudness and respiratory effort in cries. Kumar Nukala et al. (2024) and Hammoud et al. (2024) extracted ZCRs and RMSs via Python libraries, integrating them as simple yet informative descriptors of time-domain dynamics.

Autocorrelation-based Features

Narayanan et al. (2024) used autocorrelation function (ACF) (Hassani et al., 2024)-based features, such as FZCP (fractional zero-crossing periodicity), rmax (maximum autocorrelation value), kmax (lag at

maximum autocorrelation), ZCP12 (zero-crossing periodicity at lag 12), and DR (decorrelation ratio). These features capture periodicity and pitch consistency by analysing the correlation between a signal and its delayed version. Their block-wise detection approach enabled the precise localisation of sound events within longer cry or community audio segments.

Table 9 - Datasets and extracted features used for infant cry and community sound classification.

Study	Dataset	Features Used	Feature Extraction Details
(Xu et al., 2025)	Custom CED Sound Dataset	FF-Orbital Patterns, UTMDWT	Orbital + wavelet band features (5632); INCA selection
(Dey et al., 2025)	Baby Chillanto dataset: 1049 normal + 340 asphyxia cries; audio sampled to 8 kHz, 1-second segments	MFCC	MFCCs extracted per frame (time & frequency domain); preprocessing: noise removal, outlier handling, label encoding, Random Oversampling; MFCC plots and Pandas DataFrame conversion for model input
(Ozcan & Gungor, 2025)	Donate a Cry	MFCC + Data Augmentation	MFCC (13), robust DA, structure-tuned ANN
(Narayanan et al., 2024)	ESC, Donate a Cry, YouTube	ACF	Five ACF-based features (FZCP, rmax, kmax, ZCP12, DR); blockwise detection
(Kumar Nukala et al., 2024)	Donate a Cry	ZCR, RMS, MFCC, Mel-spectrogram, TSI	457 features; 5 sec audio; multidomain; TSI converts MFCCs to images
(Hammoud et al., 2024)	Donate a Cry	ZCR, RMS, Mel- spectrogram, MFCC, TSI	5-sec audio segments; time (ZCR, RMS), frequency (Mel-spectrogram), time-frequency (MFCC); MFCC transformed with multiple TSI algorithms (GADF, GASF, MTF, RP, RGB-GAF)
(Zayed et al., 2023)	Cry audio recordings from newborns with neonatal RDS, sepsis, and healthy cries; collected in hospitals, segmented into expiratory segments, balanced to 3396 samples	GFCC (cepstral), Harmonic Ratio (prosodic), Spectrogram (image-based)	GFCCs and HR extracted via MATLAB; spectrograms generated via Python; VGG16 CNN used for feature extraction; fusion applied both by simple concatenation and through the deep learning process
(Lee et al., 2020)	Audio from video recordings of 39 infants (ASD & TD); collected at Seoul National University Bundang Hospital; ages 6–24 months	Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS, 88 features); AutoEncoder bottleneck features	eGeMAPS features extracted using OpenSMILE (frame size: 25 ms, overlap: 10 ms); normalized; AutoEncoder compresses to latent dimension (54); joint optimization with BLSTM; feature map evaluated with t-SNE

Time series imaging (TSI)-based techniques SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

To benefit from the strengths of Deep Learning (DL) models designed for visual input, several studies

have converted sequential acoustic features into 2D images. This includes techniques such as the following:

- Gramian angular summation field (GASF) and Gramian angular difference field (GADF) (Alsalemi et al., 2023) transform a time series into a polar coordinate system and compute the Gramian matrix on the basis of the angular cosine (summation) or sine (difference), encoding temporal correlations as textures. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.
- The Markov transition field (MTF) (Zhao et al., 2022) encodes transition probabilities between discrete quantile bins of the time series, capturing dynamics as spatial maps. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.
- Recurrence plot (RP) (Marwan et al., 2007) plots when states of a system recur in phase space, visualizing similarity patterns
- RGB-GAF (Chai et al., 2025) stacks multiple GAF matrices into RGB channels, enriching feature diversity.

Kumar Nukala et al. (2024) and Hammoud et al. (2024) used these TSI methods to convert MFCC sequences into image-like representations, contributing to convolutional neural networks to learn spatial correlations in temporal data.

Autoencoder-based dimensionality reduction

To address the high dimensionality and redundancy of large acoustic feature sets, Lee et al. (2020) implemented an autoencoder (AE), a deep neural network, trained to compress input features into a compact latent representation and then reconstruct the original data. They used the extended Geneva minimalistic acoustic parameter set (eGeMAPS), comprising 88 prosodic and spectral features, extracted with OpenSMILE, a robust open-source toolkit for speech analysis. The AE reduced these vectors to a 54- dimensional latent vector, which was then fed to a bidirectional long short-term memory (BLSTM) network for sequential modelling.

Machine learning approaches in neonatal vocal analysis

Machine Learning has emerged as a transformative tool in the analysis of neonatal vocalisations, empowering researchers to uncover subtle, clinically meaningful patterns hidden within complex, high-dimensional acoustic signals. As summarized in Table 10, various supervised ML techniques, including support vector machines (SVMs), random forests (RFs), logistic regression (LR), decision trees (DTs), k-nearest neighbors (KNNs), and gradient boosting methods such as XGBoost, are commonly deployed to classify infant vocal samples into diagnostically relevant categories, such as typical versus atypical neurodevelopment, respiratory distress, sepsis, or signs of birth asphyxia. Supervised learning relies on datasets annotated with clear ground-truth labels, contributing to these algorithms to learn optimal boundaries and rules for differentiating pathological cries from healthy cries with remarkable precision.

For example, Xu et al. (2025) coupled advanced feature selection (INCA) with a Bayesian-optimized SVM, achieving an impressive accuracy of 98.81% for cry classification. Similarly, Dey et al. (2025) systematically compared traditional ML classifiers (such as LR, RF, SVM, KNN, and NB) and reported that logistic regression achieved a near-perfect accuracy of 99.16% for asphyxia detection, outperforming several deep neural variants in the same study (Table 10). Compared with these conventional models, ensemble techniques have supported significant gains in prediction robustness. Hammoud et al. (2024) and Kumar Nukala et al. (2024) reported how combining multiple learners, such as RF, XGBoost, and bagging, can exploit the strengths of individual algorithms while mitigating their weaknesses, increasing the classification accuracy above 98% in some cases. Narrayanan et al. (2024) applied a diverse suite of models, including decision trees, naive Bayes, multilayer perceptron (MLP), light gradient boosting machine (LGBM), and KNN, and achieved standout results for rapid and lightweight prediction pipelines suitable for real-time applications in resource- constrained settings.

In addition to classic ML, Deep Learning architectures have become indispensable in neonatal vocal research because of their ability to automatically discover and hierarchically encode complex acoustic patterns without the need for extensive manual feature crafting. Convolutional neural networks (CNNs) excel at extracting localized spectral features from spectrogram images of cry signals, capturing fine-grained details such as pitch modulations, formant transitions, and transient noise bursts that can distinguish healthy from pathological cries. For example, Ozcan & Gungor (2025) leveraged a structure-tuned ANN optimized via GridSearch to enhance the performance of cry classification, whereas Zayed et al. (2023) fused spectrogram-derived CNN features with handcrafted

Table 10 - Machine learning (ML) and deep learning (DL) algorithms with performance metrics.

Study	ML/DL Type	Algorithms Used	Performance Measures
(Xu et al., 2025)	ML	INCA (feature selection) + Bayesian-optimized SVM	Accuracy: 98.81%; high Recall, Precision, F1-score (~98.8%)
(Dey et al., 2025)	Combined ML & DL	ML: Logistic Regression (LR), SVM, RF, KNN, DT, NB; DL: custom ANN, ANN1, CNN, CNN1, CNN2 with hidden layers	ML: Best Logistic Regression: 99.16% accuracy, 0.008% error; DL: Best ANN1: 98.20% accuracy, 0.018% error; evaluated using Precision, Recall, F1-score, Confusion Matrix, ROC
(Ozcan & Gungor, 2025)	DL	Structure-Tuned Artificial Neural Network (ANN) with GridSearch + Data Augmentation	Accuracy: 90%, F1-score: 90%
(Narayanan et al., 2024)	ML	DT, RF, NB, MLP, LGBM, KNN	Best SE: NB 99.29% (fast, smallest model); Best overall accuracy: RF 93.77% (higher cost); DT: fast (1.07 ms); KNN: balanced SE 92.82%, SP 94.08%
(Kumar Nukala et al., 2024)	ML (Ensemble focus)	Random Forest, XGBoost, SVM, DT, KNN, LR	Best: RF & XGBoost with 98.03% accuracy (10-fold CV); strong feature importance on MFCCs, ZCR, RMS; visualized via confusion matrices
(Hammoud et al., 2024)	ML Ensemble	RF, SVM, DT, KNN, Bagging	MFCC-RF: Accuracy 96.39% (outperforms prior SOTA 95.17%); multiple TSI variants show robust F1 & precision
(Zayed et al., 2023)	Hybrid ML and DL with feature fusion	Support Vector Machine (SVM), Random Forest (RF), Deep Neural Network (DNN) with VGG16 for spectrogram; GridSearchCV and Keras Tuner for hyperparameter optimization	Achieved highest accuracy of 97.50% (spectrogram + GFCC + HR fused through learning); evaluated using accuracy, precision, recall, F1-score, confusion matrix, ROC curves
(Lee et al., 2020)	Hybrid DL with AutoEncoder feature compression and BLSTM	SVM with linear kernel; vanilla BLSTM; BLSTM jointly optimized with AutoEncoder bottleneck features	Joint optimized BLSTM showed improved ASD detection over vanilla BLSTM; results scored with Unweighted Average Recall (UAR) & Weighted Average Recall (WAR); t- SNE shows clearer feature separability

GFCC and prosodic measures, yielding a peak accuracy of 97.50% when deep neural networks with joint feature learning were used (Table 10).

Recurrent neural networks (RNNs), especially bidirectional long short-term memory (BLSTM) networks, are equally pivotal for modelling the temporal dependencies inherent in cry sequences. Lee et al. (2020) innovatively combined an AE for feature compression with a BLSTM classifier, refining the latent representation of high-dimensional speech parameters such as eGeMAPS and improving the discrimination of autism spectrum disorder vocal signatures. Their approach supported that compressing features into a meaningful bottleneck and jointly optimising it with a sequence model can enhance both detection accuracy and interpretability when dealing with sparse and noisy infant vocal data.

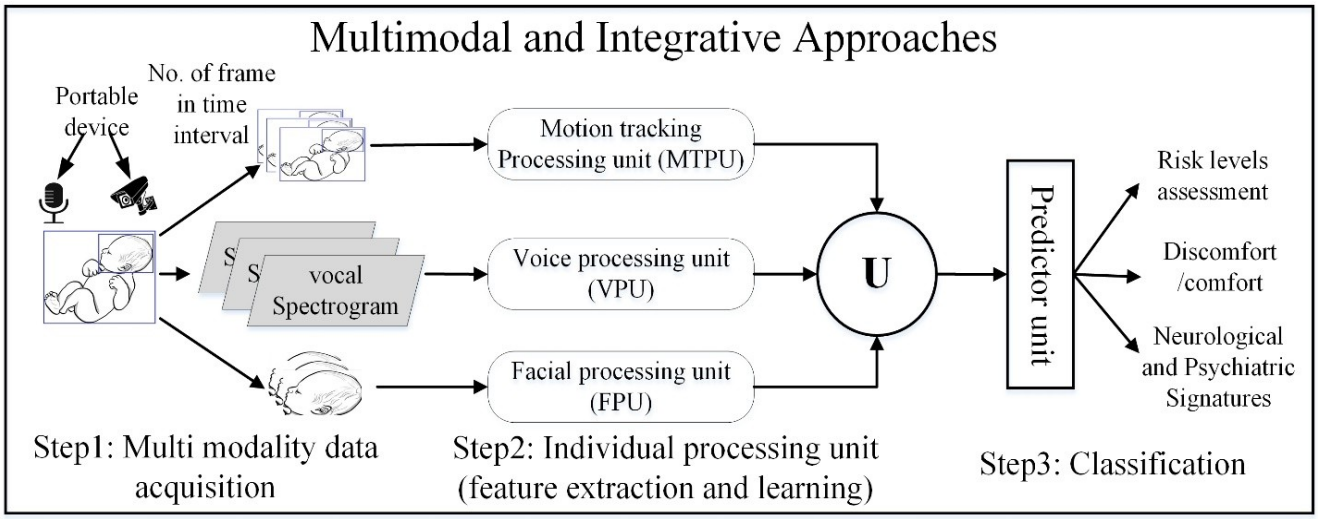

Figure 7 - A conceptual multimodal workflow integrating motion, voice, and facial data.

A critical consideration across studies is the balance between manual feature engineering and automated feature learning. Traditional pipelines often depend on carefully selected prosodic (pitch, intonation), cepstral (MFCC, GFCC), and spectral (harmonics-to-noise ratio, formant frequencies) features, which are grounded in phonetic and neurological knowledge. These handcrafted features are transparent and interpretable but may overlook nuanced, nonlinear relationships. In contrast, DL's automated feature learning, which is evident in CNNs and autoencoders, allows models to discover latent signal representations directly from raw waveforms or time-frequency images. While powerful, this strategy demands larger datasets and can reduce interpretability, a challenge partially mitigated by hybrid approaches such as those of Zayed et al. (2023) and Lee et al. (2020), who combine handcrafted and learned features to exploit the best of both worlds (Table 10).

As models grow more sophisticated and are increasingly entrusted with supporting early diagnosis and clinical decision-making, explainability becomes paramount. Explainable AI (XAI) techniques such as SHAP, LIME, and neural attention mechanisms are now being integrated into neonatal vocal analysis pipelines. These tools clarify which features or time segments drive a model's predictions, bolster clinician

confidence and facilitate trust in automated systems. This transparency not only helps validate the biological plausibility of discovered vocal biomarkers but also aids in refining models by exposing biases or misclassifications. Ultimately, embedding XAI transforms black-box predictors into intelligible,

clinically actionable tools, closing the gap between advanced computation and practical neonatal healthcare.

Multimodal and integrative approaches

In neonatology and early developmental neuroscience, the interpretation of infant vocalisations is increasingly recognised as a valuable window into the developing brain and nervous system (Narayanan et al., 2022). However, the diagnostic and predictive utility of infant cry and vocal sound analysis is substantially amplified when these acoustic features are interpreted within a multimodal framework that includes concurrent physiological and behavioural signals (Pigueiras-del-Real et al., 2024b). Modern research and clinical practice emphasize that no single data stream, whether audio, visual, or physiological, can fully capture the complexity of an infant's internal state or developmental trajectory.

As depicted in Figure 7, a multimodal and integrative pipeline typically begins with synchronised data acquisition from portable sensors, such as microphones and cameras, capturing voice, facial expressions, and body movements within defined time intervals. Each modality is then processed through specialised units such as motion tracking, vocal spectrogram analysis, and facial feature extraction before being unified in a central fusion module. This combined information feeds into a predictive unit that classifies the infant's state, contributing to risk assessment, discomfort detection, and the identification of early neurological or psychiatric signatures. Such architectures highlight how leveraging diverse and complementary data streams can significantly improve the sensitivity and reliability of neonatal monitoring systems.

Combining vocal features with physiological measures such as heart rate variability, respiration, or cortical signals (EEGs) can reveal how the autonomic and central nervous systems coordinate during stress, pain, or social engagement (Shah et al., 2025), as described in Table 11. For example, synchronising cry acoustics with heart rate or oxygen saturation can clarify whether a seemingly normal cry actually masks distress or autonomic dysregulation. Similarly, coupling vocalisations with EEG signals helps identify the neural circuits involved in phonation control, which can be disrupted in certain neurological conditions. Video-based monitoring adds further depth by capturing facial expressions and body movements that contextualize vocal sounds, distinguishing a pain cry from a hunger cry or a cry accompanied by unusual posturing that may suggest a neuromotor problem.

To achieve this integration, researchers deploy multisensory platforms that combine microphones, high-definition cameras for pose or facial key point tracking, wearable biosensors for heart rate and oxygen saturation, and sometimes bedside devices for cortical or hemodynamic measurements (e.g., near-infrared spectroscopy, NIRS) (Pigueiras-del-Real et al., 2024a). These sensors collect time-synchronised data, creating a detailed temporal map of behaviours and physiological states. This is crucial because temporal cooccurrences such as a cry immediately following a noxious stimulus and accompanied by a spike in heart rate can provide robust evidence of pain perception and the effectiveness of analgesic interventions.

Recent studies have supported the ability of ML and DL to process rich, multimodal data. For example, Natraj et al. (2024) combined OpenPose (Cao et al., 2021) with a deep feature extractor (VGG16) and a temporal sequence model (LSTM) to extract body keypoints from video and interpret body movement patterns over time. In parallel, audio streams are processed using signal features such as Mel-frequency cepstral coefficients (MFCCs) and spectrograms and then classified via 1D-CNN architectures designed for sequential audio data. These modality-specific predictions are then fused via ensemble decision rules (logical OR/AND) or trained decision tree classifiers, increasing accuracy by leveraging complementary strengths: movement cues can disambiguate ambiguous cries, and vice versa.

Similarly, Salekin (2022) used a bilinear VGG16 network, a variant that models fine-grained interactions between features extracted from different modalities (e.g., face and body), followed by an LSTM to capture the temporal progression of pain-related facial expressions and body movements in neonates in the NICU. For the audio channel (crying), spectrogram images were analysed with another VGG16 network. The outputs of these unimodal branches were fused via decision fusion, where predictions from each channel were combined to produce a final estimate of pain or behaviour. The results showed that this integrated approach outperformed single-channel models, with the multimodal system achieving an area under the curve (AUC) of 0.90 compared with 0.78--0.87 for individual modalities.

Additionally, Shah (2025) presented a unified, explainable framework for neonatal health monitoring that integrates four complementary data streams: facial features, vocal expressions, electrocardiograms (ECGs), and motion keypoints. For facial analysis, high-quality facial regions are automatically cropped via MediaPipe and then processed via a hybrid CNN architecture (FRAI) that combines GoogleNet and

AlexNet, which achieves precision, recall, and F1 scores of approximately 92-95% on benchmark datasets. Vocal expressions are converted to spectrograms and analysed with modified XceptionNet, a computationally efficient variant of the Xception network, which delivers robust emotion and pain detection with precision and recall ranging from 90% to 98% in both the adult and infant datasets. ECG signals from out-of-hospital cardiac arrest cases are transformed into visual descriptors via SIFT and bag- of-visual-words, fused with patient metadata, and classified with an ensemble machine learning model, attaining a balanced accuracy of 0.82 and an AUC- ROC of 0.90. For motion tracking, advanced angular and inertia-based interpolation combined with LSTM networks accurately imputes missing trajectory data, significantly improving error metrics such as the MSE, MAE, RMSE, cosine, and Huber losses. This multisensor, hybrid-fusion approach contributes to strong real-time performance and interpretability through SHAP, LIME, and Grad-CAM, offering clinicians a transparent and holistic tool for the early detection of neurodevelopmental issues and distress in neonates.

Table 11 - Overview of recent studies on multimodal automatic pain and behaviour monitoring.

Study	Multimodal Data	Features Used	ML/DL Techniques	Performance Measures
(Natraj et al., 2024)	Video (pose estimation) Audio (ADOS sessions)	Video: OpenPose keypoints Audio: MFCC, Mel spectrogram, tonal & spectral features	Video: VGG16-LSTM Audio: 1D-CNN Fusion: Ensemble (OR, AND, Decision tree)	Video: 80% acc Audio: 78.8% acc Ensemble: 82.5% acc, F1: 0.816 OR: Sensitivity 90% AND: Specificity 92.5%
(Salekin, 2022)	Neonates in NICU (USF-MNPAD-I & II)	Video (face/body), Audio (crying), Vital signs, NIRS	Facial/Body: Bilinear VGG16 + LSTM Sound: Spectrogram VGG16 Multimodal: Decision fusion	Unimodal AUC: Face (0.82), Body (0.78), Sound (0.87) Multimodal AUC: 0.90)
(Shah, 2025)	Facial features, Vocal expressions, ECG, Motion keypoints	Facial: Cropped face regions via MediaPipe; Vocal: Spectrograms processed with lightweight Xception (VocalXpressNet); ECG: Visual SIFT + BoVW; Motion: Angular & inertia interpolation + LSTM for missing data reconstruction	Facial: GoogleNet + AlexNet; Vocal: Modified Xception; ECG: Ensemble ML with visual features + demographic metadata; Motion: LSTM; fusion via hybrid early-late fusion	Facial: M3B Diseases: Acc 0.89, Bal. Acc 0.94; Vocal (Adult/Infant): Precision 90–97%, Recall 90–98%, F1 >90%; ECG (OHCA): Balanced Accuracy 0.82, AUC-ROC 0.90; Motion: Improved MSE, MAE, RMSE, Cosine & Huber Loss; reliable real- time imputation

A crucial advantage of these frameworks is that they provide contextualization: an isolated abnormal

vocal pattern may have low specificity but becomes clinically meaningful when corroborated by co-occurring abnormal motor or physiological signs. This mitigates false positives and adapts more flexibly to interindividual variability, which is a major challenge in neonatal care, where rapid developmental changes are the norm. Multimodal systems can thus support personalised baselines and help clinicians distinguish pathology-driven anomalies from benign situational variations, enhancing both sensitivity and specificity.

Addressing methodological challenges

As neonatal vocal analysis becomes increasingly reliant on data-driven models, several methodological challenges must be overcome to ensure the reliability, robustness, and clinical applicability of these technologies (Gómez-Vilda et al., 2022). One of the foremost obstacles is the variability introduced by differences in recording environments, devices, and population characteristics across datasets. Models trained on a single dataset often fail to generalize effectively to new clinical settings or demographic groups, thereby limiting their scalability and practical utility. To mitigate this, domain adaptation and transfer learning have emerged as powerful strategies (Gichoya et al., 2023). These techniques enable models to utilize knowledge learned from one domain, such as a well- annotated dataset from a specific hospital, and apply it to new, unseen domains with minimal retraining (S. T. H. Shah et al., 2024). Through mechanisms such as fine-tuning, adversarial learning, and feature alignment, models can adapt to new acoustic conditions or patient demographics, thereby enhancing cross-context reliability without requiring extensive new annotations.

Another critical tool for addressing methodological constraints is data augmentation. Given the scarcity and heterogeneity of high-quality neonatal vocal datasets, augmentation techniques play a vital role in expanding the effective size and diversity of training data (Fayaz et al., 2024). Traditional methods such as pitch shifting, time stretching, background noise injection, and waveform perturbation help simulate real-world variability and improve model robustness (Wen et al., 2025). More advanced techniques, including generative adversarial networks (GANs) and synthetic speech generation, are now being explored to produce realistic infant vocalisations that preserve meaningful acoustic features while introducing novel examples. These synthetic datasets not only increase model performance but also help reduce overfitting, especially for small or imbalanced datasets (Ma et al., 2025).

Finally, rigorous cross-dataset and cross-population validation is essential to establish the generalizability of predictive models. Too often, algorithms are evaluated only on the dataset used for training or within a narrow clinical context, which can result in overestimated performance and limited real-world applicability. By testing models across multiple datasets collected from diverse institutions, geographic regions, and patient subgroups, researchers can detect performance inconsistencies, potential biases, and applicability constraints. This level of validation is particularly critical in neonatal care, where interindividual variability is high and the clinical stakes are significant. Ensuring that models maintain accuracy across diverse populations not only builds clinical confidence but also advances the field toward the deployment of reliable, inclusive tools for early neurodevelopmental assessment.

Clinical translation and validation

The successful deployment of neonatal vocal analysis technologies in clinical practice hinges on the seamless translation of computational models into practical, reliable tools that meet the rigorous demands of real-world healthcare settings (Ganti, 2025). While laboratory-based models often demonstrate promising accuracy in detecting vocal biomarkers, transitioning these models to bedside and remote applications requires addressing critical factors such as clinical workflow compatibility, regulatory approval, interpretability, and ease of use (Arya et al., 2023). This means that models must not only perform well under controlled conditions but also maintain their diagnostic accuracy and consistency amid the noisy, unpredictable nature of clinical environments. To gain acceptance among health care professionals, these tools must produce outputs that are understandable, actionable, and clinically meaningful, which often necessitates the integration of explainable AI techniques and intuitive, user-friendly interfaces.

Clinical translation is inherently interdisciplinary. It demands active, sustained collaboration between data scientists, engineers, clinicians, speech-language pathologists, neonatologists, and healthcare administrators (Dalwai, 2021; Murphy et al., 2025). This collaboration ensures that the development of neonatal vocal analysis tools is firmly grounded in real-world needs, constraints, and practical opportunities. Clinicians contribute crucial domain knowledge about symptom presentation, patient variability, and the clinical relevance of vocal and physiological features, whereas engineers and data scientists design robust algorithms and ensure seamless system integration. These partnerships are also vital for addressing the ethical, privacy, and logistical complexities of data collection, especially when

working with vulnerable populations such as neonates.

In addition to hospital environments, interest in the use of neonatal vocal analysis for remote and continuous monitoring is increasing. Real-world applications include home-based tracking of high-risk infants post-discharge, early screening in rural or underserved communities, and integration with telehealth platforms (Chiang et al., 2021). Wearable or ambient sensors, combined with cloud-based processing and mobile interfaces, can empower caregivers and clinicians to monitor developmental signals in real time and intervene promptly when anomalies arise. These applications represent a paradigm shift in paediatric healthcare, moving from reactive treatment to proactive, data-driven early intervention (Shah et al., 2022, 2025).

However, for such systems to be clinically viable, they must undergo rigorous validation in diverse, uncontrolled environments and be designed to safeguard privacy, minimise caregiver burden, and function with minimal calibration or technical oversight (Al-Worafi, 2024). When effectively translated, these innovations have the potential to transform neonatal care, contributing to earlier diagnosis, personalised treatment plans, and improved neurodevelopmental outcomes for at-risk infants.

Future directions

Future directions must prioritise advancing and refining multimodal integration approaches to fully utilize the diverse streams of information available from neonates. While significant progress has been made in analysing vocalisations (Jeong & Ha, 2025), physiological signals (Gentile et al., 2023), motion patterns (Bruschetta et al., 2025), and facial features (Shah et al., 2023) independently, the greatest clinical benefit will come from combining these data sources in an intelligent, synchronised manner. The development of flexible and robust fusion strategies, including hybrid architectures that blend early, intermediate, and late fusion methods, will enable systems to capture subtle interactions across modalities, increasing accuracy and reliability in real-world settings (Guarrasi et al., 2025). Advanced techniques such as attention mechanisms, transformer-based fusion, and graph- based relational models hold great promise for modelling complex cross-modal relationships.

Equally crucial is the establishment of standardized, high-quality multimodal datasets that represent diverse clinical contexts, populations, and recording conditions. Current datasets often vary widely in

quality and completeness, hindering generalizability (Krones et al., 2025). Collaborative efforts to build large-scale, open-access repositories can support the training and preliminary support for models that are resilient to variations and bias and can better reflect the full spectrum of neonatal health states. Additionally, research should investigate innovative data augmentation and synthetic data generation methods to address data imbalance and scarcity, particularly for rare conditions.

To ensure that these technologies can be deployed effectively at the bedside and beyond, efficiency and scalability must be prioritised. Future work should focus on compressing complex multimodal models without sacrificing interpretability or diagnostic performance. Techniques (Violos et al., 2025) such as model pruning, quantization, and edge computing adaptations can enable real-time monitoring of portable devices or smart incubators, extending continuous care into both hospital and home environments.

Explainability and user interaction will remain central to the clinical adoption of these systems. As models become increasingly complex with multimodal inputs, SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

new explainable AI frameworks must be designed to clearly communicate how each data type contributes to predictions and enable clinicians to understand the rationale behind alerts or risk scores (S. A. H. Shah et al., 2024). User-friendly interfaces and visualization tools support this goal, fostering trust among caregivers and healthcare professionals. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Finally, longitudinal and adaptive modelling should be a key research direction. Rather than relying solely on single-timepoint predictions, future systems should track developmental changes over time, detect subtle deviations from expected growth patterns, and adapt recommendations dynamically as more data becomes available (Kraus et al., 2023). This shift toward continuous, personalised monitoring can empower clinicians and families to intervene earlier, tailor care plans more precisely, and ultimately improve neurodevelopmental outcomes for infants at risk.

Conclusion

The analysis of neonatal vocal expression stands at the intersection of neuroscience, signal processing, and artificial intelligence, offering a promising avenue for the early detection of neurological and developmental conditions. This review outlines key methodologies, ranging from traditional spectro-temporal analysis and acoustic feature extraction to advanced machine learning models capable of capturing subtle vocal biomarkers. Foundational preprocessing steps, such as noise reduction and feature extraction, transform raw infant cries into informative representations for robust computational analysis. Both classical algorithms and modern deep learning architectures, including convolutional and recurrent neural networks, have shown substantial potential for detecting early signs of neurocognitive risk. Importantly, the integration of explainable AI techniques and the growing shift toward multimodal data fusion, which combines vocal cues with complementary physiological signals, motion patterns, and facial information, has enhanced the robustness and interpretability of these systems, thereby increasing their clinical utility. The implications for early diagnosis and timely intervention are profound. By capturing and contextualising vocal signals within the broader framework of an infant's physiological and behavioural state, clinicians can detect risk factors for conditions such as autism spectrum disorder, cerebral palsy, or language impairments well before overt symptoms emerge. Early identification paves the way for targeted interventions during critical windows of neuroplasticity, improving long-term developmental outcomes. Moreover, non-invasive and cost-effective voice analysis methods are ideally suited for widespread screening, including in low-resource settings where access to advanced diagnostic tools may be limited.

As this field evolves, the integration of wearable and ambient sensors, mobile platforms, and cloud-based analytics promises to enable continuous, personalised monitoring for high-risk infants, both in hospital and home settings. In the future, the field must converge on a unified framework for developing and validating neonatal cognitive and behavioural biomarkers. This should prioritise standardized data collection protocols, inclusive and diverse datasets, rigorous cross- population validation, and close interdisciplinary collaboration to ensure both technical excellence and clinical relevance. Equally, ethical safeguards, including data privacy, informed consent, and equitable access, must remain central as these tools transition into practice. By advancing toward this vision, neonatal vocal analysis and multimodal monitoring can transition from promising research prototypes to practical, trustworthy tools for neurodevelopmental health, laying the groundwork for a new era of precision medicine for newborns and providing every child with the best possible start in life.

Conflict of Interest Statement

The authors declare that they have no conflict of interest.

Acknowledgement

The authors would like to express their sincere gratitude to Politecnico di Torino, particularly the PolitoBioMed Lab under the Department of Mechanical and Aerospace Engineering, for their valuable support and collaboration. Special thanks also go to GPI SpA, Department of Research and Development, for their technical insights and contributions, and to 7HC SRL for their partnership and continued encouragement throughout this work.

References

Alexlinander (2022) 2022DSPLab: Detecting baby sounds. https://kaggle.com/competitions/2022dsplab-detecting-baby- sounds, 2022. Kaggle.

Ali, S., Tanweer, S., Khalid, S.& Rao, N. (2021) Mel frequency cepstral coefficient: a review. In, Proceedings of the 2nd International Conference on ICT for Digital, Smart, and Sustainable Development, ICIDSSD 2020, 27-28 February 2020, Jamia Hamdard, New Delhi, India.

Alsalemi, A., Amira, A., Malekmohamadi, H. & Diao, K. (2023) Novel domestic building energy consumption dataset: 1D timeseries and 2D Gramian angular fields representation. Data Brief 47, 108985. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Al-Worafi, Y.M. (2024) Patient care related issues in the developing countries: monitoring parameters. In, Handbook of Medical and Health Sciences in Developing Countries. Springer, Cham, pp. 1–23.

Andonotopo, W., Bachnas, M.A., Dewantiningrum, J., Pramono, M.B.A., Stanojevic, M. & Kurjak, A. (2025) AI and early diagnostics: mapping fetal facial expressions through development, evolution, and 4D ultrasound. Journal of Perinatal Medicine 53, 263–285.

Arya, S.S., Dias, S.B., Jelinek, H.F., Hadjileontiadis, L.J. & Pappa, A.-M. (2023) The convergence of traditional and digital biomarkers through AI-assisted biosensing: a new era in translational diagnostics? Biosensors and Bioelectronics 235, 115387.

Bartl-Pokorny, K.D., Pokorny, F.B., Garrido, D., Schuller, B.W., Zhang, D. & Marschik, P.B. (2022) Vocalisation repertoire at the end of the first year of life: an exploratory comparison of Rett syndrome and typical development. Journal of Developmental and Physical Disabilities 34, 1053–1069.

Bellanca, J.L., Lowry, K.A., VanSwearingen, J.M., Brach, J.S. & Redfern, M.S. (2013) Harmonic ratios: a quantification of step to step symmetry. Journal of Biomechanics 46, 828–831.

Bruschetta, R., Caruso, A., Micai, M., Campisi, S., Tartarisco, G., Pioggia, G. & Scattoni, M.L. (2025) Marker-less video analysis of infant movements for early identification of neurodevelopmental disorders. Diagnostics 15, 136.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E. & Sheikh, Y. (2021) OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 43, 172-186.

Cappelli, G. & Noccetti, S. (2022) A Linguistic Approach to the Study of Dyslexia. Multilingual Matters, Bristol, UK.

Chai, Y., Deng, L., Shao, R., Zhang, J., Xing, L., Zhang, H. & Liu,

Y. (2025) GAF: Gaussian action field as a dynamic world model for robotic manipulation. arXiv.org. Preprint.

Chiang, M.F., Starren, J.B. & Demiris, G. (2021) Telemedicine and telehealth. In, Shortliffe, E.H., Cimino, J.J. (Eds.), Biomedical Informatics: Computer Applications in Health Care and Biomedicine. Springer, Cham, pp. 667–692.

Dalwai, S.H. (2021) IAP Handbook of Developmental and BehaviouralPaediatrics. Jaypee Brothers Medical Publishers.

Dey, S.K., Mohi Uddin, K.M., Howlader, A., Mahbubur Rahman, Md., Babu, H.Md.H., Biswas, N., Siddiqi, U.R. & Mazumder, B., (2025) Analysing infant cry to detect birth asphyxia. Neuroscience Informatics 5, 100193.

Ding, H. & Zhang, Y. (2023) Speech prosody in mental disorders. Annual Review of Linguistics 9, 335–355.

Fayaz, S., Shah, S.Z.A., Din, N.M. ud, Gul, N. & Assad, A. (2024) Advancements in data augmentation and transfer learning: a comprehensive survey to address data scarcity challenges. Recent

Advances in Computer Science and Communications 17, 14–35.

Filippa, M., Della Casa, E., D'amico, R., Picciolini, O., Lunardi, C., Sansavini, A. & Ferrari, F. (2021) Effects of early vocal contact in the neonatal intensive care unit: study protocol for a multi-centre, randomised clinical trial. International Journal of Environmental Research and Public Health 18, 3915.

Filippa, M. & Kuhn, P. (2024) Early parental vocal contact in neonatal units: rationale and clinical guidelines for implementation. Frontiers in Neurology 15, 1441576

Frank, M.C. (2020) Quantifying sources of variability in infancy research using the infant-directed-speech preference. Advances in Methods and Practices in Psychological Science 3, 24–52.

Fu, M., Li, D., Gadhiya, A., Lambright, B., Alowais, M., Bahnassy, M., Elletter, S.E.D., Toyin, H.O., Jiang, H., Zhang, K., Aldarmaki, H. (2025) Infant cry detection using causal temporal representation. In, ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Ganti, V.K.A.T. (2025) Beyond the stethoscope: how artificial intelligence is redefining diagnosis, treatment, and patient care in the 21st century. Deep Science Publishing.

Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M. & Ritter, M. (2017) Audio set: an ontology and human-labelled dataset for audio events. In, 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 776–780.

Gentile, F.R., Shah, S.T.H., Sperti, M., Panagiotopoulos, K., Primi, R., Bendotti, S., Currao, A., Compagnoni, S., Baldi, E., Lopiano, C., Vicini Scajola, L., Marconi, G., Deriu, M.A. & Savastano, S. (2023) An innovative medical decision support tool for neurological outcome prediction from post-resuscitation electrocardiograms (MILESTONE). European Heart Journal 44, ehad655.650.

Gichoya, J.W., Thomas, K., Celi, L.A., Safdar, N., Banerjee, I., Banja, J.D., Seyyed-Kalantari, L., Trivedi, H. & Purkayastha, S. (2023) AI pitfalls and what not to do: mitigating bias in AI. British Journal of Radiology 96, 20230023.

Gómez-Vilda, P., Gómez-Rodellar, A., Palacios-Alonso, D., Rodellar-Biarge, V. & Álvarez-Marquina, A. (2022) The role of data analytics in the assessment of pathological speech—a critical appraisal. Applied Sciences 12, 11095.

Guarrasi, V., Aksu, F., Caruso, C.M., Di Feola, F., Rofena, A., Ruffini, F. & Soda, P. (2025) A systematic review of intermediate fusion in multimodal deep learning for biomedical applications. Image and Vision Computing 158, 105509.

Hammoud, M., Getahun, M.N., Baldycheva, A. & Somov, A. (2024) Machine learning-based infant crying interpretation. Frontiers in Artificial Intelligence 7,1337356.

Hassani, H., Royer-Carenzi, M., Mashhad, L.M., Yarmohammadi, M. & Yeganegi, M.R. (2024) Exploring the depths of the autocorrelation function: its departure from normality. Information 15, 449.

Hou, X., Zhang, P., Mo, L., Peng, C. & Zhang, D. (2024) Neonatal sensitivity to vocal emotions: a milestone at 37 weeks of gestational age. eLife 13, RP95393.

Husain, A., Knake, L., Sullivan, B., Barry, J., Beam, K., Holmes, E., Hooven, T., McAdams, R., Moreira, A., Shalish, W. & Vesoulis, Z. (2025) AI models in clinical neonatology: a review of modelling approaches and a consensus proposal for standardized reporting of modelperformance. Paediatric Research https://doi.org/10.1038/s41390-025-04207-6

Jeong, Y. & Ha, S. (2025) Early developmental changes in infants' vocal responses in interactions with caregivers. Infant Behaviour and Development 78, 102022.

Joo, S., Choi, J., Kim, N. & Lee, M.C. (2021) Zero-crossing rate method as an efficient tool for combustion instability diagnosis. Experimental Thermal and Fluid Science 123, 110340.

Kamiloğlu, R.G. & Sauter, D.A. (2021) Voice production and perception. In: Oxford Research Encyclopedia of Psychology. Oxford University Press, Oxford.

Kao, C. & Zhang, Y. (2025) Age and sex differences in infants' neural sensitivity to emotional prosodies in spoken words: a multifeature oddball study. Journal of Speech, Language, and Hearing Research 68, 332–348.

Keles, E. & Bagci, U. (2023) The past, current, and future of neonatal intensive care units with artificial intelligence: a systematic review. NPJ Digital Medicine 6, 220.

Kraus, B., Zinbarg, R., Braga, R.M., Nusslock, R., Mittal, V.A. & Gratton, C. (2023) Insights from personalised models of brain and behaviour for identifying biomarkers in psychiatry. Neuroscience and BiobehaviouralReviews 152, 105259.

Kristian, Y., Simogiarto, N., Sampurna, M.T.A., Hanindito, E. & Visuddho, V. (2023) Ensemble of

multimodal deep learning autoencoder for infant cry and pain detection. F1000Research 11, 359.

Krones, F., Marikkar, U., Parsons, G., Szmul, A. & Mahdi, A. (2025) Review of multimodal machine learning approaches in healthcare. Information Fusion 114, 102690.

Kumar Nukala, V., Reddy Motheline, S., Wesley Kolasanakoti, J., Vankayalapati, S., Velupula, V. & Reddy Dodda, V. (2024) Advanced machine learning approaches for infant cry classification using audio feature extraction. In, 2024 International Conference on Integrated Intelligence and Communication Systems (ICIICS), pp.1–7.

La Rosa, V.L., Geraci, A., Iacono, A. & Commodari, E. (2024) Affective touch in preterm infant development: neurobiological mechanisms and implications for child–caregiver attachment and neonatal care. Children 11, 1407.

Lee, J.H., Lee, G.W., Bong, G., Yoo, H.J. & Kim, H.K. (2020) Deep-learning-based detection of infants with autism spectrum disorder using auto-encoder feature representation. Sensors 20, 6762.

Liu, Q., Song, L., Xu, D. & Long, Y. (2025) ICSD: an open-source dataset for infant cry and snoring detection.arXiv.org. Preprint.

Long, B., Xiang, V., Stojanov, S., Sparks, R.Z., Yin, Z., Keene, G.E., Tan, A.W.M., Feng, S.Y., Zhuang, C., Marchman, V.A., Yamins, D.L.K. & Frank, M.C.(2024) The BabyView dataset: high-resolution egocentric videos of infants' and young children's everyday experiences.arXiv.org. Preprint.

Long, H.L., Eichorn, N. & Oller, D.K. (2023) A probe study on vocal development in two infants at risk for cerebral palsy. Developmental Neurorehabilitation 26, 44–51.

Ma, F., Li, Y., Xie, Y., He, Y., Zhang, Y., Ren, H., Liu, Z., Yao, W., Ren, F., Yu, F.R. & Ni, S. (2024) A review of human emotion synthesis based on generative technology.IEEE Transactions on Affective Computing (Preprint).

Mallegni, N., Molinari, G., Ricci, C., Lazzeri, A., La Rosa, D., Crivello, A. & Milazzo, M. (2022) Sensing devices for detecting and processing acoustic signals in healthcare. Biosensors 12, 835.

Marschik, P.B., Widmann, C.A.A., Lang, S., Kulvicius, T., Boterberg, S., Nielsen-Saines, K., Bölte, S., Esposito, G., Nordahl- Hansen, A., Roeyers, H., Wörgötter, F., Einspieler, C., Poustka, L. & Zhang, D. (2022a) Emerging verbal functions in early infancy: lessons from observational and computational approaches on typical development and neurodevelopmental disorders. Advances

in Neurodevelopmental Disorders 6, 369–388.

Marschik, P.B., Widmann, C.A.A., Lang, S., Kulvicius, T., Boterberg, S., Nielsen-Saines, K., Bölte, S., Esposito, G., Nordahl-

Hansen, A., Roeyers, H., Wörgötter, F., Einspieler, C., Poustka, L. & Zhang, D. (2022b) Emerging verbal functions in early infancy: lessons from observational and computational approaches on typical development and neurodevelopmental disorders. Advances in Neurodevelopmental Disorders 6, 369–388. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Marwan, N., Romano, M.C., Thiel, M. & Kurths, J. (2007) Recurrence plots for the analysis of complex systems. Physics Reports 438, 237-329

Meissen, F., Breuer, S., Knolle, M., Buyx, A., Müller, R., Kaissis, G., Wiestler, B. & Rückert, D. (2024) (Predictable) performance bias in unsupervised anomaly detection. eBioMedicine 101,105002.

Murphy, M.M., Colquitt, G.T., Ryals, P.S., Shin, K., Kjeldsen, W.C., McIntyre, A., Whitten, S.V.W., Modlesky, C.M. & Maitre, N.L. (2025) Synergies, discrepancies, and action priorities: a statewide engagement study to strengthen clinical research in cerebral palsy. Health Expectations 28, e70257.

Narayanan, D.Z., Takahashi, D.Y., Kelly, L.M., Hlavaty, S.I., Huang, J. & Ghazanfar, A.A. (2022) Prenatal development of neonatal vocalisations. eLife 11, e78485.

Narayanan, S.P., Manikandan, M.S. & Cenkeramaddi, L.R. (2024) Fast autocorrelation feature-based infant cry detector for resource- efficient affordable edge cry sound analysis systems. In, 2024 IEEE 19th Conference on Industrial Electronics and Applications (ICIEA), pp.1–6.

Natraj, S., Kojovic, N., Maillart, T. & Schaer, M. (2024) Video- audio neural network ensemble for comprehensive screening of autism spectrum disorder in young children. PLOS ONE 19, e0308388.

Nussbaum, C., Frühholz, S. & Schweinberger, S.R. (2025) Understanding voice naturalness. Trends in Cognitive Sciences 29, 467–480.

Onciul, R., Tataru, C.-I., Dumitru, A.V., Crivoi, C., Serban, M., Covache-Busuioc, R.-A., Radoi, M.P. & Toader, C. (2025) Artificial intelligence and neuroscience: transformative synergies in brain research and clinical applications. Journal of Clinical Medicine 14, 550. SER is therefore

presented as a promising assessment direction rather than a clinically validated tool.

Ozcan, T. & Gungor, H. (2025) Baby cry classification using structure-tuned artificial neural networks with data augmentation and MFCC features. Applied Sciences 15, 2648.

Piczak, K.J. (2015) ESC: dataset for environmental sound classification. In, Proceedings of the 23rd ACM International Conference on Multimedia, MM '15. Association for Computing Machinery, New York, NY, USA, pp.1015–1018.

Pigueiras-del-Real, J., Gontard, L.C., Benavente-Fernández, I., Lubián-López, S.P., Gallero-Rebollo, E. & Ruiz-Zafra, A. (2024a) NRP: a multi-source, heterogeneous, automatic data collection system for infants in neonatal intensive care units. IEEE Journal of Biomedical and Health Informatics 28, 678–689.

Pigueiras-del-Real, J., Ruiz-Zafra, A., Benavente-Fernández, I., Lubián-López, S.P., Shah, S.A.H., Shah, S.T.H. & Gontard, L.C. (2024b) NeoVault: empowering neonatal research through a neonate data hub. BMC Paediatrics 24, 787.

Reyes-Galaviz, O.F., Cano-Ortiz, S.D. & Reyes-García, C.A. (2008) Evolutionary-neural system to classify infant cry units for pathologies identification in recently born babies. In, 2008 Seventh Mexican International Conference on Artificial Intelligence, pp.330–335.

Ribolsi, M., Fiori Nastro, F., Pelle, M., Medici, C., Sacchetto, S., Lisi, G., Riccioni, A., Siracusano, M., Mazzone, L. & Di Lorenzo, G. (2022) Recognizing psychosis in autism spectrum disorder. Frontiers in Psychiatry 13, 768586.

Romo, N., Robb, M.P., Lee, J. & Wermke, K. (2024) Noise phenomena in distress cries of term and very preterm infants at term-equivalent age. Logopedics Phoniatrics Vocology 50, 48-54.

Rudenko, Y. (2023) Neurophysiological and neuropsychological mechanisms of vocalisation contrasting with music perception. SSRN Blog.

Salekin, M.S. (2022) Generative spatio-temporal and multimodal analysis of neonatal pain. Ph.D.thesis University of South Florida, United States – Florida.

Sanna, M. (2025) Proprioceptive resonance and multimodal semiotics: readiness to act, embodied cognition, and the dynamics of meaning. NeuroSci 6, 42.

Shah, S.A.H., di Terlizzi, A. & Deriu, M.A. (2022) Intelligent system development to monitor neonatal behaviour: a review. In, Conference: International Workshop in Neurodevelopmental

Impairments in Preterm Children - Computational Advancements (DETERMINED 2022). Ljubljana, Slovenia.

Shah, S.A.H., Shah, S.T.H., Khaled, R., Buccoliero, A., Shah, S.B.H., Di Terlizzi, A., Di Benedetto, G. & Deriu, M.A. (2024) Explainable AI-based skin cancer detection using CNN, particle swarm optimization and machine learning. Journal of Imaging 10, 332.

Shah, S.T.H., 2025. Multimodal AI tools for predicting neurological and neurodevelopmental trajectories. PhD thesis. Politecnico di Torino, Italy – Torino.

Shah, S.T.H., Shah, S.A.H., Khan, I.I., Imran, A., Shah, S.B.H., Mehmood, A., Qureshi, S.A., Raza, M., Di Terlizzi, A., Cavaglià, M. and Deriu, M.A., 2024. Data-driven classification and explainable-AI in the field of lung imaging. Frontiers in Big Data, 7, 1393758.

Shah, S.T.H., Shah, S.A.H., Panagiotopoulos, K., Pigueiras-del- Real, J., Qayyum, K., Shah, S.B.H., Qureshi, S.A., Di Terlizzi, A., Di Benedetto, G. & Deriu, M.A. (2025) Artificial intelligence coupled with the Internet of Things targeting neurodevelopmental challenges in preterm neonates. Journal of Multiscale Neuroscience 4, 32–56.

Shah, S.T.H., Shah, S.A.H., Qureshi, S.A., Di Terlizzi, A. & Deriu, M.A. (2023) Automated facial characterization and image retrieval by convolutional neural networks. Frontiers in Artificial Intelligence 6, 1230383.

Sheikh, S.A., Sahidullah, M. & Kodrasi, I. (2025) Deep learning for pathological speech: A survey. arXiv preprint.

Sullivan, J., Mei, M., Perfors, A., Wojcik, E. and Frank, M.C., 2021. SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. Open Mind 5, 20–29. Veres, G. (2025) gveres/donateacry-corpus. [online] GitHub, Inc.

Violos, J., Diamanti, K.-C., Kompatsiaris, I. & Papadopoulos, S. (2025) Frugal machine learning for energy-efficient, and resource- aware artificial intelligence. arXiv preprint.

Wagner, L., Banchik, M., Tsang, T., Okada, N.J., Altshuler, R., McDonald, N., Bookheimer, S.Y., Jeste, S.S., Green, S. & Dapretto, M. (2025) Atypical early neural responses to native and non-native language in infants at high likelihood for developing autism. Molecular Autism 16, 6.

Wang, T.V. & Song, P.C. (2022) Neurological voice disorders: A review. International Journal of Head and Neck Surgery 13, 32–40.

Wen, Y., Innuganti, A., Ramos, A.B., Guo, H., & Yan, Q. (2025). SoK: How robust is audio watermarking in generative AI models? arXiv preprint.

Xu, L., Yildiz, A.M., Tuncer, I., Ozyurt, F., Dogan, , S. & Tuncer, T. (202.5) Detection of community emotions through sound: An investigation using the FF-Orbital chaos-based feature extraction model. Ain Shams Engineering Journal 16, 103248.

Zayed, Y., Hasasneh, A., & Tadj, C. (2023). Infant cry signal diagnostic system using deep learning and fused features. Diagnostics 13, 2107.

Zhang, E.Q. (2025). The influence of prenatal auditory input on newborn vocalisations. SSRN Blog.

Zhao, X., Sun, H., Lin, B., Zhao, H., Niu, Y., Zhong, X., Wang, Y., Zhao, Y., Meng, F., Ding, J., Zhang, X., Dong, L., & Liang, S. (2022). Markov transition fields and deep learning-based event-classification and vibration-frequency measurement for φ-OTDR. IEEE Sensors Journal 22, 3348–3357.

8. Synthesis & Discussion

8.1 Integration of findings across papers

This section interprets the thesis as a structured thesis-by-publication programme in which the overall contribution emerges not from the simple accumulation of individual studies, but from their cumulative interpretation within a common research trajectory. In this format, coherence cannot be assumed solely on the basis of thematic proximity or chronological sequencing. Rather, it must be supported by clarifying how each publication contributes to the broader intellectual purpose of the dissertation and by distinguishing between central, supporting, and adjacent levels of contribution.

The present thesis addresses a common field of inquiry located at the intersection of speech emotion recognition, voice-based assessment, psychological interpretation, and digital health applications. However, the way in which this field is explored is intentionally heterogeneous, combining literature synthesis, technical development, empirical application, and cross-domain methodological reflection.

Seen in this light, the thesis does not claim that all included studies address the same question in the same way, nor that they provide identical forms of evidence. Its coherence lies instead in a layered architecture. At its core is the attempt to examine whether and how voice-based analytical approaches, and in particular speech emotion recognition, may contribute to psychologically relevant assessment contexts, including burnout, treated here as a focal but bounded construct-related assessment. Around this core, the thesis also considers the methodological conditions, translational possibilities, and cross-domain implications that surround such approaches. The result is not a single linear validation sequence, but a cumulative programme in which different studies illuminate different dimensions of the same broader problem space.

This interpretation is especially important in a thesis-by-publication format, where individual papers are necessarily shaped by their own research questions, publication contexts, and methodological boundaries. The doctoral contribution therefore resides not only in the content of each paper taken separately, but also in the integrative work of showing how they can be read as parts of a broader scholarly project. In the present case, that project concerns the development and interpretation of voice-based approaches to emotional and psychological assessment, with burnout functioning as a focal line

of inquiry, but not as the sole evidentiary anchor for every included study.

To make this architecture explicit, the thesis can be read through three interrelated lines of contribution.

The first is the core line, which concerns the **role of speech emotion recognition and voice-based indicators in psychologically relevant assessment**, and their possible relevance for burnout, treated here as a focal but bounded construct. This line is most directly supported by the conceptual framing developed in the first part of the thesis and by the technical and empirical work that examines the feasibility of linking vocal and emotional information to structured psychological interpretation. Within this line, the central doctoral question is not whether burnout detection through voice biomarkers has already been conclusively supported, but whether the thesis provides a rigorous basis for understanding the potential, conditions, and limits of this line of research.

The second is a supporting line of contribution. This concerns the **methodological and translational conditions** that make such research meaningful. It includes the systematic mapping of the field, the technical refinement of SER approaches, and the effort to situate these tools within broader discussions of digital assessment, interpretability, and applied use. This supporting line is essential because the doctoral contribution does not depend on a single isolated empirical result. It depends instead on whether the thesis constructs a sufficiently robust foundation for arguing that voice-based assessment deserves serious consideration as part of an emerging interdisciplinary research agenda.

The third is an adjacent line of contribution. This includes studies that do not provide direct relevance to burnout but nevertheless contribute to the broader interpretive and translational architecture of the thesis. Their value lies in showing that **voice-based or digitally mediated approaches can acquire relevance across different applied contexts**, populations, and methodological settings. In this sense, these studies should not be interpreted as equivalent demonstrations of the core claim, but rather as adjacent contributions that broaden the thesis's understanding of transferability, contextual adaptation, and the wider research landscape in which burnout-related voice assessment must ultimately be positioned.

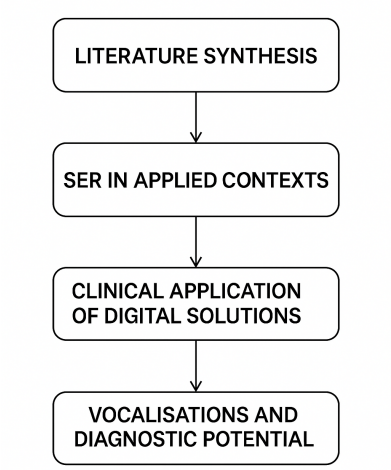

Figure 8 - Flowchart of research trajectory from literature synthesis to empirical validation.

This three-level reading helps avoid two opposite distortions. On the one hand, it prevents the thesis from being reduced to a set of disconnected papers. On the other hand, it avoids overstating unity by pretending that all included works provide the same kind of support for the same conclusion. The contribution of the thesis is strongest precisely when these differences are made explicit and interpreted productively rather than obscured.

The four publications included in this thesis do not carry identical evidentiary weight because they address different objects, populations, and levels of analysis. This differentiation is not a weakness in itself; it becomes problematic only when heterogeneous studies are read as if they all directly validate the same thesis claim.

Publication 1 provides the foundational basis of the thesis. Its principal role is to map the state of the art on voice biomarkers and emotional assessment, thereby establishing the scientific and methodological landscape within which the rest of the thesis is situated. It does not directly validate burnout detection, but it performs an indispensable architectural function: it clarifies why voice-based approaches deserve investigation, which lines of prior research are most relevant, and where important opportunities and limitations lie.

Publication 2 occupies the most central methodological position in the thesis. It contributes direct technical value by examining speech emotion recognition in real-world environments and by refining the conditions under which voice-based analysis can be performed beyond highly controlled settings.

Although this publication does not in itself constitute definitive burnout validation, it is the study that most directly supports the thesis's focal interest in the feasibility and applied relevance of SER-based approaches for psychologically meaningful assessment.

Publications 3 and 4 make a different kind of contribution. They should not be interpreted as direct confirmations of the thesis's focal burnout. Rather, they broaden the translational and methodological frame within which the thesis is understood:

- Publication 3, through its application in a digital intervention context, contributes adjacent translational relevance by showing how digitally mediated assessment and intervention environments may create conditions in which emotionally and cognitively sensitive technologies become meaningful. Their role is framed as indirect rather than as direct evidence for burnout.
- Publication 4 extends the methodological horizon further by demonstrating the broader interpretive reach of vocal analysis in another domain of health-related research. Its contribution is therefore best understood as cross-domain methodological extension rather than as indirectly relevant to burnout.

Making these distinctions explicit is essential for the integrity of the thesis. Once the publication set is interpreted according to differentiated evidentiary roles, the coherence no longer depends on treating all four studies as equal proof of a single claim. Instead, coherence emerges from the structured relationship among foundational mapping, methodological core development, and adjacent translational or cross-domain extension.

8.2 Collective contribution to knowledge

Publication 1 supports the foundational knowledge base from which the rest of the thesis proceeds. Its principal contribution lies in systematically mapping the use of voice biomarkers for emotional assessment and in clarifying the scientific terrain within which speech emotion recognition must be situated. By synthesising the existing literature, it identifies both the promise and the limitations of voice-based approaches, showing that the field is sufficiently developed to justify further investigation while also remaining fragmented in terms of populations, outcomes, analytical approaches, and validation

standards. This is important because it situates the thesis within an supported yet still evolving research area rather than presenting its topic as entirely novel or methodologically ungrounded.

From the perspective of the cumulative thesis argument, Publication 1 does not provide indirectly relevant to burnout nor does it independently support claims about digital care or patient empowerment. Its significance is of a different kind. It defines the broader evidentiary landscape in which those more specific claims must be located. In particular, it contributes to that voice-based emotional assessment already constitutes a meaningful line of inquiry in the wider literature and that speech-derived indicators can plausibly be considered in relation to psychological and health-relevant states. In this sense, Publication 1 performs a foundational rather than confirmatory role. It identifies the conditions under which a doctoral investigation into voice-based emotional assessment becomes scientifically legitimate.

Its contribution is therefore cumulative in an architectural sense. It provides the conceptual and empirical map that allows the later studies to be interpreted as part of a broader programme rather than as isolated technical exercises. It also makes clear that any thesis-level claim concerning burnout must be developed cautiously, in dialogue with a field that remains methodologically heterogeneous and still in need of stronger validation across contexts. As such, Publication 1 does not answer the thesis question directly, but it supports why that question is worth asking and how it can be approached in a disciplined way.

Publication 2 adds the most direct technical and methodological support to the thesis's focal line of inquiry. Whereas Publication 1 maps the broader field, Publication 2 moves into the active development and evaluation of speech emotion recognition in real-world environments. Its significance lies in demonstrating that SER can be implemented under conditions that are closer to practical application than highly constrained laboratory settings, and that the analysis of vocal and emotional information can be pursued in a way that is methodologically robust, operationally relevant, and attentive to ecological complexity.

Within the cumulative structure of the thesis, this study is particularly important because it narrows the distance between conceptual plausibility and applied methodological execution. It does not by itself establish a complete pathway to burnout, nor does it resolve all questions concerning validity across populations and contexts. However, it provides bounded evidence that the technical infrastructure necessary for psychologically relevant voice-based assessment is not merely hypothetical. It shows that

operationalising SER in realistic settings is feasible and that such work can produce meaningful performance and interpretive insights. In that respect, Publication 2 represents the methodological core of the dissertation.

Its contribution is cumulative in two senses. First, it extends the foundational literature logic supported in Publication 1 by moving from field mapping to implementation-oriented methodological development. Second, it establishes the strongest link within the thesis between voice-based analysis and the possibility of burnout. This bridge remains bounded: the study should be interpreted as supporting the feasibility and methodological seriousness of the research line, rather than as offering definitive clinical validation for burnout detection. Yet it is precisely through this bounded contribution that Publication 2 becomes central to the doctoral argument. It provides the strongest basis for claiming that speech emotion recognition may constitute a meaningful component of future psychologically informed digital assessment frameworks.

Publications 3 and 4 contribute indirectly to the thesis by extending its interpretive and translational horizons rather than by directly validating its focal burnout-related. Their role in the dissertation is therefore not to serve as equivalent evidence for the central claim, but to show how the broader methodological and applied relevance of voice and digitally-mediated approaches may be understood across different settings. This distinction is crucial because these studies lose coherence when they are read as if they were expected to provide the same kind of evidentiary support as the more directly relevant components of the thesis.

Publication 3 contributes to adjacent translational relevance. Its importance lies in showing how digital intervention frameworks can be examined in relation to psychologically meaningful outcomes in a clinical context. Although the study does not address burnout directly and its population and immediate aims differ from the dissertation's central thesis line, it nonetheless expands the broader applied horizon. It indicates that digitally mediated approaches may be integrated into care-related contexts in ways that invite further reflection on assessment, monitoring, and support. Its contribution should therefore be interpreted as adjacent and translational rather than central and confirmatory. Their role is framed as indirect rather than as direct evidence for burnout.

Publication 4 contributes a different form of indirect value. By focusing on neonatal vocal expression

and methodological approaches to identifying neurological and psychiatric signatures, it extends the methodological reach of vocal analysis beyond the occupational and burnout focus of the thesis. This study does not provide direct evidence for the thesis's focal claim, nor should it be read as doing so. Its contribution instead lies in demonstrating that the interpretation of vocal signals has broader scientific relevance across health-related domains. In cumulative terms, it helps situate the thesis within a wider methodological ecosystem in which vocal analysis is not limited to a single population or diagnostic target but participates in a broader family of research efforts concerned with extracting meaningful information from voice and behaviour.

Taken together, the significance of Publications 3 and 4 is indirect and bounded. They show that the broader logic of technologically mediated psychological or health-related interpretation can travel across contexts, but they also underline the importance of distinguishing between the transferability of methodological reasoning and direct preliminary support for a specific clinical or occupational claim. This distinction allows the studies to retain a meaningful place in the thesis without requiring them to bear a burden of proof they were not designed to carry.

When the four publications are interpreted through differentiated evidentiary roles, the cumulative learning of the thesis becomes clearer. The thesis does not provide a single, uniform line of proof leading from one study to the next. Instead, it constructs a layered contribution. At the foundational level, it supports that voice biomarkers and speech emotion recognition constitute a scientifically credible and methodologically active field of inquiry for emotional and psychologically relevant assessment. At the technical level, it contributes to that SER can be developed and assessed in realistic settings with methodological seriousness and practical relevance. At the translational and cross-domain level, it indicates that voice-based and digitally mediated approaches may acquire broader significance across health-related contexts, although such significance does not automatically amount to direct confirmation of burnout claims.

The thesis's cumulative lesson is therefore not that burnout detection through voice biomarkers has already been conclusively supported across all included studies. Rather, the thesis shows that there is a defensible interdisciplinary basis for investigating this possibility, that the methodological tools required for such investigation can be meaningfully developed, and that the broader research trajectory has relevance beyond a single narrow application domain. In this sense, the doctoral contribution lies in

connecting foundational mapping, methodological refinement, and bounded translational extension into a coherent argument about the potential and limits of voice-based psychological assessment.

This cumulative interpretation also helps clarify what kind of originality the thesis offers. Its originality does not derive from a single decisive empirical breakthrough or from the introduction of a fully new theoretical model. It derives instead from the way the thesis brings together heterogeneous but related contributions into a structured doctoral argument that links affective computing, psychological assessment, and digital health reasoning. The thesis thus contributes less as a definitive endpoint and more as a carefully articulated research position: one that supports the relevance of SER and voice-based approaches in burnout and psychologically meaningful assessment contexts, while also making explicit the boundaries within which that relevance should currently be understood.

8.3 Theoretical implications

The central contribution of this thesis should therefore be interpreted in a bounded and differentiated manner. It does not rest on the assumption that all four publications provide indirectly relevant to burnout detection, nor does it claim that voice biomarkers have already achieved definitive clinical validity for this purpose. Rather, the contribution lies in showing that speech emotion recognition and related voice-based approaches constitute a scientifically credible and methodologically promising line of inquiry for psychologically relevant assessment, within which burnout-related assessment represents a focal, but not uniformly supported, application domain.

Under this interpretation, burnout remains an important organising line of the thesis, but it should not be treated as the single evidentiary anchor through which every publication must be justified. The more defensible reading is that the thesis develops a cumulative argument about the relevance, conditions, and limits of applying SER and voice-based analysis to emotional and psychological assessment. Within that broader contribution, burnout occupies a privileged position because it motivates the thesis conceptually, frames several of its interpretive ambitions, and helps define the translational significance of the research. At the same time, the empirical basis for burnout-specific validation remains partial and should therefore be presented with appropriate caution.

The strongest direct support for the thesis's focal line comes from the combination of the foundational

literature mapping and the methodological work on SER in real-world conditions. Together, these elements provide a basis for arguing that voice-based approaches deserve serious consideration for burnout assessment and broader psychologically informed digital health applications. However, that basis should be understood as enabling and suggestive rather than conclusive. The thesis contributes to the understanding of how such approaches may be positioned, interpreted, and further developed; it does not offer a final demonstration that burnout can already be reliably detected through vocal analysis across contexts.

This distinction between direct and indirect contribution is essential for the doctoral coherence of the work. Publication 1 and Publication 2 are most closely aligned with the central contribution because they define the field and develop the methodological core of the thesis. Publications 3 and 4, by contrast, should be understood as broadening the translational and methodological horizon of the dissertation rather than confirming its focal claim. When read in this way, the thesis gains coherence not by treating heterogeneous studies as equivalent, but by recognising that different forms of evidence can still contribute to a common interdisciplinary research programme.

The doctoral contribution of the thesis is therefore best characterised as cumulative, bounded, and interdisciplinary. It is cumulative because it connects literature synthesis, methodological development, and broader application-oriented reflection into a coherent interpretive sequence. It is bounded because it makes the strongest claims where the evidence is most direct and more cautious claims where the contribution is indirect or exploratory. It is interdisciplinary because it links psychological theory, affective computing, digital assessment, and health-related interpretation without reducing the thesis to any one of these domains alone.

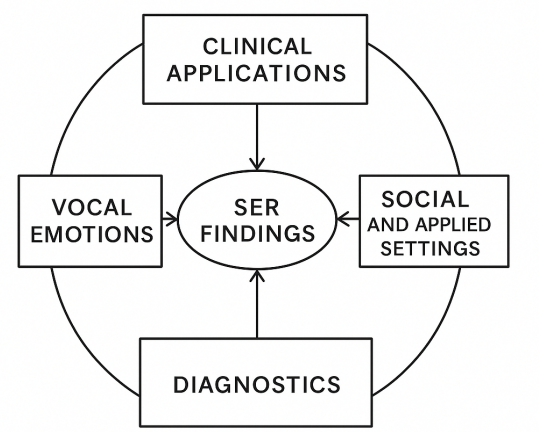

Figure 9 - Conceptual framework integrating SER findings across domains.9

8.4 Practical implications (conditional and prospective)

This differentiated interpretation is particularly important when considering the role of Publications 3 and 4 within the overall thesis architecture. Both studies contribute meaningfully to the dissertation, but they do so through forms of relevance that are adjacent to, rather than directly constitutive of, the thesis's focal burnout line. Their inclusion is therefore justified not because they independently validate burnout detection, but because they expand the translational and methodological frame within which the central contribution can be understood.

Publication 3 should be interpreted as providing adjacent translational relevance. Its focus on a digital intervention in Parkinson's disease places it outside the direct empirical domain of burnout assessment, and it would be misleading to present it as direct evidence for the thesis's central claim. Its value lies elsewhere: it illustrates how digitally mediated tools can be embedded in clinically meaningful contexts and how technology-supported approaches may interact with psychologically relevant outcomes in real-world care environments. In this sense, the study contributes to the broader translational horizon of the thesis by showing that the movement from assessment technology to applied context is neither abstract nor purely speculative, even if the specific application studied is not burnout. Their role is framed as indirect rather than as direct evidence for burnout.

Publication 4 should be interpreted as a cross-domain methodological extension. By examining neonatal vocal expression and the identification of neurological and psychiatric signatures, it enters a domain that is clearly distinct from occupational burnout, treated here as a focal but bounded construct, and from the

adult populations more directly relevant to the thesis's focal line. Its contribution should therefore not be framed as thematic confirmation. Instead, it contributes to the analysis of vocal signals can sustain scientific relevance across highly different health-related contexts. Within the thesis architecture, this broadens the methodological horizon of voice-based analysis and shows that the interpretation of vocal markers can be pursued beyond a single application domain. Their role is framed as indirect rather than as direct evidence for burnout.

At the same time, it is important to state clearly what Publications 3 and 4 do not demonstrate. They do not provide direct evidence that burnout can be identified through voice biomarkers. They do not establish patient empowerment as a supported outcome of the thesis as a whole. They do not validate occupational health claims in a direct empirical sense. Their relevance is therefore bounded. It consists in strengthening the broader translational and methodological landscape of the dissertation, not in carrying the central evidentiary burden of the thesis.

Once these boundaries are made explicit, the inclusion of Publications 3 and 4 becomes more intelligible and more defensible. They remain meaningful because a doctoral thesis in this area is not only required to show a narrowly defined empirical result, but also to demonstrate awareness of how a research line may travel across contexts, where its limits lie, and how its broader scientific significance can be interpreted. In this respect, Publications 3 and 4 help the thesis move beyond a narrow technical focus and situate voice-based analysis within a wider translational and cross-domain research conversation. Their role is therefore secondary but legitimate: not central confirmation, but bounded extension that enriches the interpretive architecture of the thesis.

8.5 Theoretical implications

The theoretical implications of this thesis should be interpreted with caution, particularly for the link between emotional processes, vocal features, and burnout states. Taken together, the studies included in the dissertation support the view that speech carries psychologically meaningful information and that voice-based analysis can contribute to the interpretation of affective and stress-related conditions. At the same time, the thesis does not provide grounds for claiming that a single validated mechanism linking burnout to specific vocal markers has been conclusively supported. Rather, the work supports a plausible interdisciplinary framework in which emotional processes, psychophysiological activation, and changes

in vocal expression may be meaningfully related under certain conditions.

This theoretical contribution is therefore best understood as one of clarification and structured positioning. The thesis brings together psychological models of emotion, emerging work on speech emotion recognition, and the use of psychometric instruments in order to suggest that burnout states may be reflected, at least in part, in measurable vocal and emotional patterns. This suggestion is consistent with the broader literature on affective expression and with the methodological direction of the empirical work presented here. However, consistency should not be mistaken for full theoretical confirmation. The dissertation provides a basis for interpreting burnout-related vocal assessment as a credible research avenue, but it does not resolve all questions concerning specificity, causal pathways, or stability across contexts.

A further implication concerns the level at which the thesis contributes theoretically. Its value does not lie in proposing an entirely new theory of burnout, nor in offering definitive proof that specific acoustic parameters map directly onto burnout as a discrete diagnostic construct. Instead, the thesis contributes to an intermediate theoretical level. It supports the idea that burnout may be approached through a layered interpretive model in which psychological states, behavioural manifestations, and voice-based indicators can be examined in relation to one another. In this sense, the work contributes to the conceptual development of voice-based psychological assessment by articulating a plausible space of connection between affective computing and clinically or occupationally relevant constructs.

This contribution is strengthened by the fact that the thesis does not rely on a single type of evidence. The literature synthesis helps define the conceptual territory, while the methodological work on SER in real-world settings provides bounded empirical support for the feasibility of this line of inquiry. The adjacent studies do not validate the focal theoretical link, but they reinforce the broader proposition that vocal analysis can sustain meaningful interpretation across diverse applied and health-related domains. The theoretical significance of the thesis, therefore, lies less in mechanism validation than in demonstrating that a structured and interdisciplinary interpretive framework is both possible and scientifically worthwhile.

At the same time, several theoretical questions remain open. The conditions under which burnout manifests reliably in vocal expression, the extent to which such manifestations can be distinguished from

other forms of emotional or cognitive strain, and the role of contextual variables in shaping vocal output all require further clarification. For this reason, the thesis should be read as providing a basis for more precise theory development rather than as delivering a definitive theoretical settlement. Its contribution is to support, refine, and delimit an emerging line of inquiry, thereby making future work on the burnout-related voice assessment conceptually more grounded and methodologically more plausible.

8.6 Practical and translational implications

The practical relevance of the thesis also requires careful calibration. The work provides the strongest support for the idea that voice-based analysis and speech emotion recognition may be practically relevant in assessment-oriented contexts. In particular, the thesis suggests that such approaches may contribute to the development of tools that complement supported psychological measures by adding a non-invasive, potentially scalable, and context-sensitive source of information. This implication is most credible when framed in terms of assessment potential rather than as evidence of fully validated clinical deployment.

Within this bounded interpretation, the thesis has translational relevance for digital environments in which emotionally meaningful signals can be collected, analysed, and interpreted in support of broader assessment frameworks. The methodological work presented in the dissertation indicates that SER can be moved closer to ecologically valid conditions and that voice-based indicators may be considered alongside more traditional psychometric approaches. This has potential relevance for digital health and technology-supported monitoring contexts. However, such relevance should be interpreted cautiously. The thesis does not demonstrate that these approaches are already sufficient for routine implementation in clinical or occupational settings, nor does it establish that they can independently sustain diagnostic or decision-making functions.

The same caution applies to occupational health. The dissertation supports the view that SER and related voice-based methods may inform future work on emotionally relevant patterns in workplace or burnout contexts, particularly where stress, affective strain, or behavioural change are of concern. Yet the empirical basis for direct occupational application remains limited. The thesis, therefore, contributes more convincingly to the recognition and assessment side of this field than to the demonstration of workplace outcomes or occupational interventions. In practical terms, its main value lies in opening a

credible pathway for future applied work rather than in showing that occupational benefits have already been empirically supported.

Digital care and patient empowerment should also be interpreted within carefully bounded limits. The thesis suggests that voice-based and digitally mediated approaches may be relevant for more responsive, personalised, or continuous forms of assessment in health-related settings. Nevertheless, it does not demonstrate that patient empowerment, understood as increased involvement in decision-making and condition management, has been directly achieved across the empirical corpus. Nor does it show that digital care has been transformed by the approaches examined here. These broader implications remain prospective and conditional. They are best understood as translational possibilities supported by the research trajectory rather than as outcomes conclusively supported by the thesis itself.

Accordingly, the practical significance of the dissertation lies in its contribution to a realistic translational horizon. It provides a basis for considering how voice-based approaches might in the future inform about burnout, psychologically sensitive monitoring, and digital health applications, while making clear that these implications remain contingent on further validation, contextual adaptation, and stronger evidence across targeted populations. This more cautious reading does not diminish the practical value of the thesis; rather, it strengthens its credibility by aligning implications with the evidentiary limits supported in the preceding sections.

8.7 Limits of the cumulative thesis contribution

The cumulative contribution of the thesis is subject to several important limitations, and these need to be stated explicitly to maintain the integrity of the overall argument. The most significant limitation concerns the level of direct burnout validation. Although the dissertation provides a credible interdisciplinary basis for considering the relevance of speech emotion recognition and voice-based analysis in burnout-related assessment, it does not offer definitive empirical confirmation that burnout can be reliably identified through vocal markers across populations and contexts. The strongest contribution of the thesis is therefore not a final validation claim, but a structured and bounded argument for why such a line of inquiry deserves further development.

A second limitation derives from the heterogeneity of the publication set. The studies differ substantially

in design, population, immediate objectives, and level of direct connection to the thesis's focal line. This heterogeneity does not negate the value of the dissertation, but it constrains the extent to which the thesis can be read as a single linear validation programme. The cumulative contribution must therefore be interpreted through differentiated evidentiary roles rather than through the assumption of full thematic and methodological uniformity. This is particularly relevant in relation to Publications 3 and 4, whose contributions are meaningful but indirect, and whose place in the dissertation depends on bounded interpretation rather than direct alignment with the burnout focus of the thesis.

A further limitation concerns generalisability. The empirical and methodological elements presented in the thesis do not yet establish that the findings can be transferred straightforwardly across occupational groups, languages, clinical settings, or technological environments. Voice-based indicators are likely to be shaped by multiple contextual variables, including recording conditions, linguistic differences, population characteristics, and the specific constructs under investigation. For this reason, the thesis supports the plausibility of the research line more strongly than it supports broad general claims about stable applicability across domains.

There are also conceptual limits that must be acknowledged. While the dissertation supports a theoretically plausible connection between emotional processes, vocal expression, and burnout, it does not fully resolve questions of specificity. Burnout may overlap with other forms of affective strain, cognitive fatigue, distress, or context-dependent emotional variation, and the extent to which vocal signals can reliably distinguish remains uncertain. The thesis should therefore not be read as having closed the conceptual debate around burnout-related voice assessment, but rather as having contributed to a more credible basis for pursuing it.

These limitations are not peripheral to the argument; they are integral to its proper interpretation. The value of the work lies in showing what can reasonably be claimed at this stage and in resisting the temptation to present methodological promise as completed validation. This is particularly important in a thesis-by-publication format, where coherence must be earned through disciplined synthesis. The present dissertation is strongest when read as a bounded, cumulative, and interdisciplinary contribution whose significance depends on respecting the evidentiary and conceptual limits within which it operates.

8.8 Future research and validation pathway

The limitations identified above point directly to a structured agenda for future research. If the line of inquiry developed in this thesis is to mature into a more robust framework for burnout, the next step will involve targeted empirical validation in populations and settings that are directly relevant to burnout. This includes the need for studies specifically designed to examine burnout-related constructs through voice-based indicators, ideally using larger and more diverse samples, repeated measurements, and clearer links between psychometric benchmarks, behavioural interpretation, and acoustic analysis.

A second priority concerns contextual and translational validation. Future research should test whether the methodological promise shown in the current thesis can be sustained in occupational, clinical, and digitally mediated environments where real-world variability is unavoidable. This requires closer attention to language, culture, recording conditions, contextual stressors, and the practical demands of implementation. It also requires validation designs capable of distinguishing between proof of technical feasibility and proof of applied usefulness. Without this step, the translational significance of voice-based assessment will remain suggestive rather than operationally grounded.

Further work is also needed to clarify the relationship between burnout and neighbouring constructs. Because burnout may share features with other forms of emotional distress, fatigue, or psychological strain, future studies should aim to determine under what conditions voice-based markers are specific, sensitive, and interpretable in relation to burnout rather than to broader affective or cognitive states. This is not a marginal issue: it is central to the possibility of building a more rigorous conceptual and empirical framework for burnout-related vocal assessment.

In addition, the broader implications of the thesis for occupational health, digital care, and patient empowerment require more direct testing. The current dissertation provides a basis for considering these domains as relevant areas of application, but it does not establish their outcomes empirically. Future research will therefore examine whether and how voice-based approaches can be meaningfully integrated into workplace well-being initiatives, digital assessment pathways, or patient-facing systems without overstating their role or bypassing ethical, interpretive, and practical constraints. This would allow the broader significance of the thesis to be developed on firmer empirical grounds.

The future agenda implied by this thesis is therefore neither optional nor merely aspirational. It is the

natural continuation of a bounded contribution that has clarified why this research line matters, how it may be approached, and where its current limits lie. In that sense, the dissertation should be read not as the completion of the field's central questions, but as a disciplined platform for more targeted validation, more precise theory building, and more context-sensitive translational development in the years ahead.

As mentioned, although not included among the main studies of this thesis due to timing constraints, it is important to acknowledge the relevance of the study by Bassi et al. (2025), which was only recently finalised, submitted, and formally registered as a protocol on Open Science Framework. This stady, whose working team I coordinate, represents a significant advancement in the research trajectory developed throughout the thesis, as it moves more directly toward the empirical validation of AI-based models for detecting burnout risk from vocal biomarkers. While the present thesis primarily establishes the conceptual, methodological, and translational foundations of voice-based assessment, this study will extend that line of inquiry by operationalizing and testing a concrete predictive framework. Its exclusion from the main body of the dissertation is therefore not indicative of a lack of relevance but rather reflects the temporal boundaries of the doctoral process. In this sense, the study can be interpreted as a continuation and consolidation of the thesis contributions, providing a more targeted and prospective step toward the type of validation that the thesis itself frames as necessary for future research.

9. Conclusion

9.1 Summary of key findings

This thesis provides a structured examination of the role of speech emotion recognition (SER) and voice-based analytical approaches within psychologically relevant assessment and burnout contexts. Rather than producing a single linear finding, the research develops a layered set of insights that collectively clarify the potential, conditions, and limits of using vocal features as indicators of emotional and psychological states.

First, the thesis establishes that voice biomarkers and SER constitute a scientifically grounded and methodologically active field of inquiry. The synthesis of existing literature demonstrates that vocal signals can encode emotionally relevant information in ways that are increasingly accessible through computational analysis, while also highlighting the heterogeneity and fragmentation that still

characterize the field. This foundation is essential in positioning voice-based assessment not as a speculative idea, but as an emerging research domain requiring further consolidation and validation.

Second, the thesis shows that SER can be implemented and evaluated in conditions that extend beyond controlled experimental settings. The empirical and methodological work presented indicates that voice-based analysis can operate in more ecologically valid environments, supporting the feasibility of integrating such approaches into applied contexts. At the same time, these findings underline that feasibility does not equate to full validation, and that performance, interpretability, and contextual sensitivity remain critical dimensions requiring careful consideration.

Third, the thesis identifies burnout as a relevant but bounded application domain for voice-based assessment. The results support the view that burnout-related states may be meaningfully explored through vocal and emotional indicators, particularly when considered in conjunction with established psychometric frameworks. However, the evidence does not justify claims of definitive detection or clinical diagnostic validity. Instead, the thesis positions burnout-related assessment as a promising line of inquiry that benefits from, but is not exhausted by, the current empirical contributions.

Fourth, the inclusion of cross-domain and translational studies extends the interpretive horizon of the thesis. These studies demonstrate that voice-based and digitally mediated approaches can acquire relevance across different populations and application contexts. Their contribution is not to directly validate the central burnout-related claim, but to show that the underlying methodological logic of vocal analysis has broader applicability. This reinforces the idea that the value of SER lies not only in specific use cases, but in its capacity to operate within a wider ecosystem of digital health and psychological research.

Taken together, these findings support a cumulative interpretation of the thesis. The contribution does not consist in a single definitive result, but in the articulation of a coherent research position: one that recognises voice-based approaches as methodologically viable, theoretically meaningful, and practically promising, while also explicitly acknowledging the limits of current evidence. This balance between advancement and constraint is central to the scientific value of the work, as it enables the thesis to contribute to the field without overstating what has been empirically demonstrated.

9.2 Unified contribution statement

The unified contribution of this thesis lies in the integration of interdisciplinary evidence into a coherent and carefully delimited research framework addressing the role of speech emotion recognition (SER) and voice-based approaches in psychologically relevant assessment contexts. Rather than advancing a single definitive empirical claim, the thesis establishes a structured basis for understanding how voice-derived indicators may be meaningfully interpreted within the broader landscape of psychological and digital health research.

At its core, the thesis contributes by bridging three traditionally separate domains: affective computing, psychological assessment, and digital health. Through this integration, it demonstrates that voice-based analytical approaches can be positioned as methodologically viable components of emerging assessment paradigms, while also clarifying the conditions under which such approaches can be considered informative, interpretable, and contextually appropriate.

Within this framework, burnout is addressed as a focal but bounded construct. The thesis does not present burnout detection as a resolved outcome; instead, it advances a disciplined argument for why burnout-related assessment represents a relevant and promising application domain for SER and related technologies. In doing so, it contributes to reframing burnout not only as a clinical or occupational construct, but as a complex psychological state that may be approached through multimodal and technologically mediated forms of observation.

The originality of the thesis resides in this integrative positioning. Its contribution is not reducible to any single study, methodological innovation, or theoretical proposition. Rather, it emerges from the cumulative interpretation of a heterogeneous publication set, in which foundational mapping, methodological development, and cross-domain exploration are brought into a coherent doctoral argument. This structure allows the thesis to articulate both the potential and the current limits of voice-based psychological assessment without conflating exploratory findings with definitive validation.

In this sense, the thesis advances the field by providing a conceptually grounded and methodologically informed framework within which future research can operate. It clarifies what can be reasonably inferred from current evidence, identifies the conditions required for stronger validation, and situates

voice-based approaches within a broader trajectory of digitally mediated psychological assessment. The result is a contribution that is not conclusive in a narrow empirical sense, but that is structurally significant in defining how this line of research can develop in a rigorous and scientifically credible manner.

9.3 Implications for the field

The findings of this thesis carry implications that are best understood as methodological, conceptual, and translational, rather than as direct evidence of established clinical or occupational outcomes. Interpreted within these boundaries, the work contributes to ongoing developments at the intersection of affective computing, psychological assessment, and digital health.

From a methodological perspective, the thesis supports the viability of integrating speech emotion recognition into broader assessment frameworks. It demonstrates that voice-based analysis can be implemented in ecologically relevant conditions and can generate information that is potentially meaningful for understanding emotional states. However, this implication remains conditional: the reliability, interpretability, and contextual sensitivity of such systems require further refinement before they can be considered robust assessment tools across diverse populations and settings.

From a conceptual perspective, the thesis contributes to clarifying how voice-based indicators may be situated within models of psychological assessment. In particular, it supports a view of vocal features as complementary signals that can enrich, but not replace, established psychometric approaches. This is especially relevant in relation to burnout, which is treated in this work as a focal but bounded construct. The thesis therefore encourages a reframing of burnout-related assessment as a multidimensional process in which digital signals, including vocal features, may play an informative but not yet definitive role.

From a translational perspective, the research suggests that voice-based and SER-driven approaches may hold relevance within emerging digital health ecosystems. Their potential lies in enabling more continuous, low-burden, and context-sensitive forms of data collection, which could inform monitoring and early-stage assessment practices. At the same time, these implications should not be interpreted as evidence of established improvements in occupational health outcomes or patient empowerment. Rather, they indicate directions in which such outcomes might be explored through future, more targeted

research.

Importantly, the thesis also highlights the ethical and practical conditions under which these implications can be meaningfully pursued. Issues such as privacy, transparency, algorithmic bias, and user trust are not peripheral considerations but central requirements for the responsible deployment of voice-based technologies. As such, any future application of SER in psychological or health-related contexts must be developed within frameworks that prioritise explainability, accountability, and user-centred design.

Overall, the implications of this thesis lie in strengthening the plausibility and framing of a research trajectory, rather than in demonstrating its full realisation. The work supports the view that voice-based approaches can become a meaningful component of interdisciplinary research on psychological assessment and digital health, while also making clear that their integration into validated, practice-ready systems remains an open and necessary step for future investigation.

9.4 Final reflections

This thesis has examined the role of voice-based and speech emotion recognition approaches within psychologically relevant assessment contexts, with a particular focus on burnout as a focal but bounded construct. Rather than claiming definitive validation, the work has aimed to clarify the conditions under which such approaches may be meaningfully developed and interpreted.

The contribution of the thesis lies in articulating a coherent research position: one that recognises the methodological viability and interdisciplinary relevance of voice-based analysis, while explicitly acknowledging its current limitations. In doing so, the thesis positions itself not as a point of conclusion, but as a structured step within an evolving research trajectory.

This trajectory requires further empirical validation, conceptual refinement, and careful ethical consideration before voice-based systems can be integrated into robust assessment practices. At the same time, the results presented here support the view that such integration is both plausible and worthy of continued investigation.

Ultimately, the value of this work lies in balancing advancement with restraint: contributing to the

development of an emerging field while preserving the clarity and rigour necessary for its future consolidation.

10. References/Bibliography

Aarsland, D. et al. (2021) 'Author Correction: Parkinson disease-associated cognitive impairment', Nature Reviews. Disease primers, 7(1), p. 53. Available at: https://doi.org/10.1038/s41572-021-00292-z.

Acuña Mora, M., Sparud-Lundin, C., Moons, P., & Bratt, E. L. (2022). Definitions, instruments and correlates of patient empowerment: A descriptive review. Patient Education and Counselling, 105(2), 346–355. https://doi.org/10.1016/j.pec.2021.06.014

Advances in Facial Expression Recognition: A Survey of Methods, Benchmarks, Models, and Datasets. Accessed: Jan. 30, 2025. [Online]. Available: https://www.mdpi.com/2078-2489/15/3/135

Aggarwal, K., Mijwil, M. M., Al-Mistarehi, A. H., Alomari, S., Gök, M., Alaabdin, A. M. Z., Abdulrhman, S. H. (2022). Has the future started? The current growth of artificial intelligence, machine learning, and deep learning. Iraqi Journal for Computer Science and Mathematics, 3(1), 115-123. https://doi.org/10.52866/ijcsm.2022.01.01.013

Ahn, S.; Springer, K.; Gibson, J.S. Social withdrawal in Parkinson's disease: A scoping review. Geriatr. Nurs. 2022, 48, 258–268.

Al-Worafi, Y.M. (2024). Patient care-related issues in the developing countries: monitoring parameters. I, Handbook of Medical and Health Sciences in Developing Countries. Springer, Cham, pp. 1–23.

Alalayah, K.M.; Senan, E.M.; Atlam, H.F.; Ahmed, I.A.; Shatnawi, H.S.A. Automatic and Early Detection of Parkinson's Disease by Analysing Acoustic Signals Using Classification Algorithms Based on Recursive Feature Elimination Method. Diagn. 2023, 13, 1924.

Alexlinander (2022) 2022DSPLab: Detecting baby sounds. https://kaggle.com/competitions/2022dsplab-detecting-baby-sounds, 2022. Kaggle.

Ali, S., Tanweer, S., Khalid, S.& Rao, N. (2021). Mel frequency cepstral coefficient: a review. I, Proceedings of the 2nd International Conference on ICT for Digital, Smart, and Sustainable Development, ICIDSSD 2020, 27-28 February 2020, Jamia Hamdard, New Delhi, India.

Allcock, L.M. et al. (2009) 'Impaired attention predicts falling in Parkinson's disease', Parkinsonism & related disorders, 15(2), pp. 110–115. Available at: https://doi.org/10.1016/j.parkreldis.2008.03.010.

Alloni, A. et al. (2018) 'Evaluation of an ontology-based system for computerised cognitive rehabilitation', International Journal of Medical Informatics, 115, pp. 64–72. Available at: https://doi.org/10.1016/j.ijmedinf.2018.04.005.

Almutairi, N., Vlahu-Gjorgievska, E., & Win, K. Than. (2023). Persuasive features for patient engagement through mHealth applications in managing chronic conditions: A systematic literature review and meta-analysis. Informatics for Health and Social Care, 48(3), 267–291. https://doi.org/10.1080/17538157.2023.2165083

Alsalemi, A., Amira, A., Malekmohamadi, H. & Diao, K. (2023) Novel domestic building energy consumption dataset: 1D timeseries and 2D Gramian angular fields representation. Data Brief 47, 108985. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Altman, D.G. and Bland, J.M. (2005) 'Treatment allocation by minimisation', BMJ (Clinical research ed.), 330(7495), p. 843. Available at: https://doi.org/10.1136/bmj.330.7495.843.

Analysing neonatal vocal expression/ methodological approaches to identifying neurological and psychiatric signatures

Andonotopo, W., Bachnas, M.A., Dewantiningrum, J., Pramono, M.B.A., Stanojevic, M., & Kurjak, A. (2025). AI and early diagnostics: mapping fetal facial expressions through development, evolution, and 4D ultrasound. Journal of Perinatal Medicine 53, 263–285.

Angel, S., & Frederiksen, K. N. (2015). Challenges in achieving patient participation: A review of how patient participation is addressed in empirical studies. International Journal of Nursing Studies, 52(9), 1525–1538. https://doi.org/10.1016/j.ijnurstu.2015.04.008

Ankomah, S. E., Fusheini, A., Ballard, C., Kumah, E., Gurung, G., & Derrett, S. (2021). Patient-public engagement strategies for health system improvement in sub-Saharan Africa: A systematic scoping review. BMC Health Services Research, 21(1). https://doi.org/10.1186/s12913-021-07085-w SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

AnnalesD997Jokinen. Available online: https://www.utupub.fi/bitstream/handle/10024/72567/AnnalesD997Jokinen.pdf?sequence=1&isAllowed=y (accessed on 7 July 2025).

Ansari, K.A.; Johnson, A. Olfactory Function in Patients with Parkinson's Disease. J. Cbron Dis. 1975, 28, 493–497.

Antonini, A. et al. (2021) 'Correction to: The TANDEM investigation: efficacy and tolerability of levodopa-carbidopa intestinal gel in (LCIG) advanced Parkinson's disease patients', Journal of

neural transmission (Vienna, Austria), 128(6), pp. 863–865. Available at: https://doi.org/10.1007/s00702-020-02200-3.

Arksey H, & O'Malley L. (2005). Scoping studies: Towards a methodological framework. The International Journal of Social Research Methodology, 8(1), 19–32. https://doi.org/10.1080/1364557032000119616

Aromataris, E., Fernandez, R. S., Godfrey, C., Holly, C., & Khalil, H. (2014). Methodology for JBI umbrella reviews. https://ro.uow.edu.au/smhpapers/3344

Arya, S.S., Dias, S.B., Jelinek, H.F., Hadjileontiadis, L.J. & Pappa, A.-M. (2023) The convergence of traditional and digital biomarkers through AI-assisted biosensing: a new era in translational diagnostics? Biosensors and Bioelectronics 235, 115387.

Atila, O., & Şengür, A. (2021). Attention-guided 3D CNN-LSTM model for accurate speed-based emotion recognition. Applied Acoustics, 182, 108260.

Atmaja, B. T., & Sasou, A. (2022). Effects of Data Augmentations on Speech Emotion Recognition. Sensors, 22(16), 5941.

Aurélien, G. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Unsupervised Learning Techniques; O'Reilly Media, Incorporated: Sebastopol, CA, USA, 2019.

Auwal, F. I., Copeland, C., Clark, E. J., Naraynassamy, C., & McClelland, G. R. (2023). A systematic review of models of patient engagement in the development and life cycle management of medicines. Drug Discovery Today, 28(9). https://doi.org/10.1016/j.drudis.2023.103702

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449-12460.

Bahar-Fuchs, A. et al. (2017) 'Tailored and adaptive computerised Cognitive Training in older adults at risk for dementia: A randomised controlled trial', Journal of Alzheimer's disease: JAD, 60(3), pp. 889–911. Available at: https://doi.org/10.3233/JAD-170404.

Baik, J.S. et al. (2024) 'Effects of home-based computerised cognitive training in community-dwelling adults with mild cognitive impairment', IEEE journal of translational engineering in health and medicine, 12, pp. 97–105. Available at: https://doi.org/10.1109/JTEHM.2023.3317189.

Bainbridge, J.L. and Ruscin, J.M. (2009) 'Challenges of treatment adherence in older patients with Parkinson's disease', Drugs & ageing, 26(2), pp. 145–155. Available at: https://doi.org/10.2165/0002512-200926020-00006.

Bakhshi, A., Wong, A. S., & Chalup, S. (2020). End-to-end speech emotion recognition based on time

and frequency information using deep neural networks. In ECAI 2020 (pp. 969-975). IOS Press.

Balikuddembe, J.K. and Reinhardt, J.D. (2020) 'Can digitisation of health care help low-resourced countries provide better community-based rehabilitation services?', Physical therapy, 100(2), pp. 217–224. Available at: https://doi.org/10.1093/ptj/pzz162. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Barello, S., Triberti, S., Graffigna, G., Libreri, C., Serino, S., Hibbard, J., & Riva, G. (2016). eHealth for patient engagement: A systematic review. Frontiers in Psychology, 6, 2013. https://doi.org/10.3389/fpsyg.2015.02013 SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Barnes, M. (2005). The same old process? Older people, participation and deliberation. Ageing and Society, 25(2), 245–259. https://doi.org/10.1017/S0144686X04002508

Bartl-Pokorny, K.D., Pokorny, F.B., Garrido, D., Schuller, B.W., Zhang, D. & Marschik, P.B. (2022). Vocalisation repertoire at the end of the first year of life: an exploratory comparison of Rett syndrome and typical development. Journal of Developmental and Physical Disabilities 34, 1053–1069.

Bassi, C., Marinaro, F., Anarbaeva, A., Buccoliero, A., Scandola, M., Sartori, R., Ceschi, A. (2025, September 3). Developing and Validating an AI Model to Detect Burnout Risk from Vocal Biomarkers. https://doi.org/10.17605/OSF.IO/HMK4F

Battista, P. et al. (2018) 'Screening for aphasia in Neuro Degeneration for the diagnosis of patients with primary progressive aphasia: Clinical validity and psychometric properties', Dementia and geriatric cognitive disorders, 46(3-4), pp. 243–252. Available at: https://doi.org/10.1159/000492632.

Battista, P. et al. (2023) 'Access, referral, service provision and management of individuals with primary progressive aphasia: A survey of speech-language therapists in Italy', International Journal of Language & Communication Disorders, 58(4), pp. 1046–1060. Available at: https://doi.org/10.1111/1460-6984.12843. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Beck, A. T., Kovacs, M., & Weissman, A. (1979). Assessment of suicidal intention: the Scale for Suicide Ideation. Journal of Consulting and Clinical Psychology, 47(2), 343–352. https://doi.org/10.1037/0022-006X.47.2.343

Bellanca, J.L., Lowry, K.A., VanSwearingen, J.M., Brach, J.S., & Redfern, M.S. (2013). Harmonic ratios: a quantification of step-to-step symmetry. Journal of Biomechanics 46, 828–831.

Bernini, S. et al. (2021) 'A double-blind randomised controlled trial of the efficacy of cognitive training delivered using two different methods in mild cognitive impairment in Parkinson's disease: preliminary report of benefits associated with the use of a computerised tool', Ageing clinical and experimental research, 33(6), pp. 1567–1575. Available at: https://doi.org/10.1007/s40520-020-01665-2.

Berrar, D. Cross-validation. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics. 2018, 1–3, 542–545.

Bethell, J., Commisso, E., Rostad, H. M., Puts, M., Babineau, J., Grinbergs-Saull, A., ... & McGilton, K. S. (2018). Patient engagement in research related to dementia: A scoping review. Dementia, 17(8), 944–975. https://doi.org/10.1177/1471301218789292

Biundo, R. et al. (2016) 'MMSE and MoCA in Parkinson's disease and dementia with Lewy bodies: a multicenter 1-year follow-up study', Journal of neural transmission (Vienna, Austria), 123(4), pp. 431–438. Available at: https://doi.org/10.1007/s00702-016-1517-6.

Boersma, P.; van Heuven, V. Speak and unSpeak with Praat. Glot International. 2001, 5, 341–347.

Bombard, Y., Baker, G. R., Orlando, E., Fancott, C., Bhatia, P., Casalino, S., ... & Pomey, M. P. (2018). Engaging patients to improve quality of care: A systematic review. Implementation Science, 13(1). https://doi.org/10.1186/s13012-018-0784-z

Bonetti, L., Tolotti, A., Anderson, G., Nania, T., Vignaduzzo, C., Sari, D., & Barello, S. (2022). Nursing interventions to promote patient engagement in cancer care: A systematic review. International Journal of Nursing Studies, 133. https://doi.org/10.1016/j.ijnurstu.2022.104289

Bosch, S. J., & Lorusso, L. N. (2019). Promoting patient and family engagement through healthcare facility design: A systematic literature review. Journal of Environmental Psychology, 62, 74–83. https://doi.org/10.1016/j.jenvp.2019.02.003

Bot, B. M., Suver, C., Neto, E. C., Kellen, M., Klein, A., Bare, C., Doerr, M., Pratap, A., Wilbanks, J., Dorsey, E., & others. (2016). The mPower study, Parkinson's disease mobile data collected using ResearchKit. Scientific Data, 3(1), 1–9. https://doi.org/10.1038/sdata.2016.11

Brabenec, L.; Mekyska, J.; Galaz, Z.; Rektorova, I. Speech disorders in Parkinson's disease: Early diagnostics and effects of medication and brain stimulation. J. Neural Transm. 2017, 124, 303–334.

Breiman, L. Random Forests. Mach. Learn. 2001, 45, 5–32.

Bruschetta, R., Caruso, A., Micai, M., Campisi, S., Tartarisco, G., Pioggia, G., & Scattoni, M.L. (2025). Marker-less video analysis of infant movements for early identification of neurodevelopmental

disorders. Diagnostics 15, 136.

Buchman, A.S. et al. (2012) 'Nigral pathology and Parkinsonian signs in elders without Parkinson disease', Annals of neurology, 71(2), pp. 258–266. Available at: https://doi.org/10.1002/ana.22588.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005, September). A database of German emotional speech. In Interspeech (Vol. 5, pp. 1517-1520).

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B. (2005). A database of German emotional speech. Interspeech, 5, 1517–1520. doi: 10.21437/Interspeech. 005-446

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42, 335–359. https://doi.org/10.1007/s10579-008-9076-6

C. Lugaresi et al., "MediaPipe: A Framework for Building Perception Pipelines," Jun. 14, 2019, arXiv: arXiv:1906.08172. doi: 10.48550/arXiv.1906.08172.

C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. M. Montero, and F. Fernández-Martínez, "Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning," Sensors, vol. 21, no. 22, Art. no. 22, Jan. 2021, doi: 10.3390/s21227665.

Cacciante, L. et al. (2022) 'Cognitive telerehabilitation in neurological patients: systematic review and meta-analysis', Neurological sciences: official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology, 43(2), pp. 847–862. Available at: https://doi.org/10.1007/s10072-021-05770-6.

Cadel, L., Marcinow, M., Sandercock, J., Dowedoff, P., Guilcher, S. J. T., Maybee, A., ... & Kuluski, K. (2021). A scoping review of patient engagement activities during COVID-19: More consultation, less partnership. PLoS ONE, 16(9), e0257880. https://doi.org/10.1371/journal.pone.0257880

Caffarra, P. et al. (2011) 'Italian norms for the Freedman version of the Clock Drawing Test', Journal of clinical and experimental neuropsychology, 33(9), pp. 982–988. Available at: https://doi.org/10.1080/13803395.2011.589373.

Caffarra, P., Vezzadini, G., Dieci, F. and Zonato, F. (2002) 'Una versione abbreviata del test di Stroop: dati normativi nella popolazione italiana', Nuova Rivista di Neurologia, 12, pp. 111–115.

Caffarra, P., Vezzadini, G., Dieci, F., Zonato, F., et al. (2002) 'Rey-Osterrieth complex figure: normative values in an Italian population sample', Neurological sciences: official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology, 22(6), pp. 443–447. Available at: https://doi.org/10.1007/s100720200003.

Canini, M. et al. (2014) 'Computerised neuropsychological assessment in ageing: testing efficacy and clinical ecology of different interfaces', Computational and mathematical methods in medicine, 2014, p. 804723. Available at: https://doi.org/10.1155/2014/804723.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE transactions on affective computing, 5(4), 377-390.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing, 5(4), 377–390. https://doi.org/10.1109/TAFFC.2014.2336244

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E. & Sheikh, Y. (2021) OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 43, 172-186.

Cappelli, G. & Noccetti, S. (2022) A Linguistic Approach to the Study of Dyslexia. Multilingual Matters, Bristol, UK.

Carlesimo, G.A. et al. (1996) 'The mental deterioration battery: Normative data, diagnostic reliability and qualitative analyses of cognitive impairment', European neurology, 36(6), pp. 378–384. Available at: https://doi.org/10.1159/000117297.

Carman, K. L., Dardess, P., Maurer, M., Sofaer, S., Adams, K., Bechtel, C., & Sweeney, J. (2013). Patient and family engagement: A framework for understanding the elements and developing interventions and policies. Health Affairs, 32(2), 223–231. https://doi.org/10.1377/hlthaff.2012.1133

Catania, F., Wilke, J. W., & Garzotto, F. (2025). Emozionalmente: A Crowdsourced Corpus of Simulated Emotional Speech in Italian. IEEE Transactions on Audio, Speech and Language Processing.

Catricalà, E. et al. (2017) 'SAND: a Screening for Aphasia in NeuroDegeneration. Development and normative data', Neurological sciences: official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology, 38(8), pp. 1469–1483. Available at: https://doi.org/10.1007/s10072-017-3001-y.

Cené, C. W., Johnson, B. H., Wells, N., Baker, B., Davis, R., & Turchi, R. (2016). A narrative review of patient and family engagement. Medical Care. Retrieved from https://www.lww-medicalcare.com

Chai, Y., Deng, L., Shao, R., Zhang, J., Xing, L., Zhang, H., & Liu, Y. (2025). GAF: Gaussian action

field as a dynamic world model for robotic manipulation. arXiv.org. Preprint.

Chattopadhyay, S., Dey, A., & Basak, H. (2020). Optimising speech emotion recognition using manta-ray based feature selection. arXiv preprint arXiv:2009.08909.

Chaudhuri, K.R.; Azulay, J.P.; Odin, P.; Lindvall, S.; Domingos, J.; Alobaidi, A.; Kandukuri, P.L.; Chaudhari, V.S.; Parra, J.C.; Yamazaki, T.; et al. Economic Burden of Parkinson's Disease: A Multinational, Real-World, Cost-of-Illness Study. Drugs Real World Outcomes 2024, 11, 1–11.

Chegini, Z., Arab-Zozani, M., Shariful Islam, S. M., Tobiano, G., & Abbasgholizadeh Rahimi, S. (2021). Barriers and facilitators to patient engagement in patient safety from patients and healthcare professionals' perspectives: A systematic review and meta-synthesis. Nursing Forum, 56(4), 938–949. https://doi.org/10.1111/nuf.12635

Chiang, M.F., Starren, J.B., & Demiris, G. (2021). Telemedicine and telehealth. In Shortliffe, E.H., Cimino, J.J. (Eds.), Biomedical Informatics: Computer Applications in Health Care and Biomedicine. Springer, Cham, pp. 667–692.

Christen, P.; Hand, D.J.; Kirielle, N. A review of the F-measure: Its history, properties, criticism, and alternatives. ACM Comput. Surv. 2023, 56, 1–24.

Chudyk, A. M., Horrill, T., Waldman, C., Demczuk, L., Shimmin, C., Stoddard, R., ... & Schultz, A. S. H. (2022). Scoping review of models and frameworks of patient engagement in health services research. BMJ Open, 12(8), e063507. https://doi.org/10.1136/bmjopen-2022-063507 SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Clare, L. et al. (2003) 'Cognitive rehabilitation and cognitive training for early-stage Alzheimer's disease and vascular dementia', Cochrane database of systematic reviews, (4), p. CD003260. Available at: https://doi.org/10.1002/14651858.CD003260.

Clarke, V., & Braun, V. (2017). Thematic analysis. The Journal of Positive Psychology, 12(3), 297–298. https://doi.org/10.1080/17439760.2016.1262613

Convey, R.B.; Laukkanen, A.M.; Ylinen, S.; Penttilä, N. Analysis of Voice in Parkinson's Disease Utilising the Acoustic Voice Quality Index. J. Voice 2024, 1–10. https://doi.org/10.1016/j.jvoice.2023.12.025.

Cosgrove, J., Alty, J.E. and Jamieson, S. (2015) 'Cognitive impairment in Parkinson's disease', Postgraduate medical journal, 91(1074), pp. 212–220. Available at: https://doi.org/10.1136/postgradmedj-2015-133247.

Costantini, G., Cesarini, V., & Casali, D. (2022). A Subset of Acoustic Features for Machine Learning-based and Statistical Approaches in Speech Emotion Recognition. In BIOSIGNALS (pp. 257-

264).

Costantini, G., Iaderola, I., Paoloni, A., & Todisco, M. (2014). EMOVO corpus: an Italian emotional speech database. In Proceedings of the ninth international conference on language resources and evaluation (LREC'14) (pp. 3501-3504). European Language Resources Association (ELRA).

Costantini, G., Parada-Cabaleiro, E., Casali, D., & Cesarini, V. (2022). The Emotion Probe: On the Universality of Cross-Linguistic and Cross-Gender Speech Emotion Recognition via Machine Learning. Sensors, 22(7). https://doi.org/10.3390/s22072461

Cozad, M. J., Crum, M., Tyson, H., Fleming, P. R., Stratton, J., Kennedy, A. B., ... & Horner, R. D. (2022). Mobile health apps for patient-centred care: Review of United States rheumatoid arthritis apps for engagement and activation. JMIR mHealth and uHealth, 10(12), e39881. https://doi.org/10.2196/39881

Crawford, M. J., Rutter, D., Manley, C., Weaver, T., Bhui, K., Fulop, N., & Tyrer, P. (2002). Systematic review of involving patients in the planning and development of health care. British Medical Journal, 325(7375), 1263–1265. https://doi.org/10.1136/bmj.325.7375.1263

Csipke, E., Flach, C., McCrone, P., Rose, D., Tilley, J., Wykes, T., & Craig, T. (2014). Inpatient care 50 years after the process of deinstitutionalisation. Social Psychiatry and Psychiatric Epidemiology, 48, 639–648. https://doi.org/10.1007/s00127-013-0788-6

D. Ortiz-Perez, M. Benavent-Lledo, J. Garcia-Rodriguez, D. Tomás, and M. F. Vizcaya-Moreno, "Deep Insights into Cognitive Decline: A Survey of Leveraging Non-Intrusive Modalities with Deep Learning Techniques," Oct. 24, 2024, arXiv: arXiv:2410.18972. doi: 10.48550/arXiv.2410.18972.

Dal Rí, F. A., Ciardi, F. C., & Conci, N. (2023). Speech emotion recognition and deep learning: an extensive validation using convolutional neural networks. IEEE Access, 11, 116638-116649.

Dalwai, S.H. (2021) IAP Handbook of Developmental and Behavioural Paediatrics. Jaypee Brothers Medical Publishers.

Danhof-Pont, M. B., van Veen, T., & Zitman, F. G. (2011). Biomarkers in burnout, as a focal but bounded construct,: A systematic review. Journal of Psychosomatic Research, 70(6), 505–524. https://doi.org/10.1016/j.jpsychores.2010.10.012

Dao, S.V.; Yu, Z.; Tran, L.V.; Phan, P.N.; Huynh, T.T.; Le, T.M. An Analysis of Vocal Features for Parkinson's Disease Classification Using Evolutionary Algorithms. Diagnostics 2022, 12. 1980.

Das, R. A comparison of multiple classification methods for diagnosis of Parkinson's disease. Expert Syst Appl. 2010, 37, 1568–1572.

Davis, S.; Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition

in continuously spoken sentences. IEEE Trans Acoust. 1980, 28, 357–366.

De Luca, R. et al. (2019) 'Computer-assisted cognitive rehabilitation improves visuospatial and executive functions in Parkinson's disease: Preliminary results', NeuroRehabilitation, 45(2), pp. 285–290. Available at: https://doi.org/10.3233/NRE-192789.

Dejonckere, P.H.; Bradley, P.; Clemente, P.; Cornut, G.; Friedrich, G.; Van De Heyning, P. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques: Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS). Eur. Arch. Oto-Rhino-Laryngol. 2001, 258, 77–82.

Dejonckheere, M., & Vaughn, L. M. (2019). Semistructured interviewing in primary care research: A balance of relationship and rigour. Family Medicine and Community Health, 7, e000057. https://doi.org/10.1136/fmch-2018-000057

Demiroglu, C.; Si, D.; Atkins, D.C.; Ghomi, R.H.; Wroge, T.J.; Ozkanca, Y. Parkinson's Disease Diagnosis Using Machine Learning and Voice. In Proceedings of the 2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), Philadelphia, PA, USA, 1 December 2018; IEEE: New York, NY, USA, 2018; pp. 1–7.

Dey, S.K., Mohi Uddin, K.M., Howlader, A., Mahbubur Rahman, Md., Babu, H.Md.H., Biswas, N., Siddiqi, U.R. & Mazumder, B., (2025). Analysing infant cry to detect birth asphyxia. Neuroscience Informatics 5, 100193.

Di Cesare, M.G.; Perpetuini, D.; Cardone, D.; Merla, A. Assessment of Voice Disorders Using Machine Learning and Vocal Analysis of Voice Samples Recorded through Smartphones. BioMedInformatics 2024, 4, 549–565.

Di, Y., Wang, J., Li, W., & Zhu, T. (2021). Using i-vectors from voice features to identify major depressive disorder. Journal of Affective Disorders, 288(February), 161–166. https://doi.org/10.1016/j.jad.2021.04.004

Díez-Cirarda, M. et al. (2018) 'Neurorehabilitation in Parkinson's disease: A critical review of cognitive rehabilitation effects on cognition and brain', Neural plasticity, 2018, p. 2651918. Available at: https://doi.org/10.1155/2018/2651918.

Ding, H. & Zhang, Y. (2023.) Speech prosody in mental disorders. Annual Review of Linguistics 9, 335–355.

Domecq, J. P., Prutsky, G., Elraiyah, T., Wang, Z., Nabhan, M., Shippee, N., ... & Murad, M. H. (2014). Patient engagement in research: A systematic review. BMC Health Services Research, 14, 89.

https://doi.org/10.1186/1472-6963-14-89 SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Dorsey, E.R.; Sherer, T.; Okun, M.S.; Bloem, B.R. The emerging evidence of the Parkinson pandemic. J. Park. Dis. 2018, 8, S3–S8.

Doty, R.L. Olfactory dysfunction in neurodegenerative diseases: Is there a common pathological substrate? Lancet Neurol. 2017, 16, 478–488.

Edwards, J.D. et al. (2013) 'Randomised trial of cognitive speed of processing training in Parkinson's disease', Neurology, 81(15), pp. 1284–1290. Available at: https://doi.org/10.1212/WNL.0b013e3182a823ba.

Effectiveness of a Home-Based Computerised Cognitive Training in Parkinson's Disease/ A Pilot Randomised Cross-Over Study

Ekman, P. (1992). An argument for basic emotions. Cognition and Emotion, 6(3-4), 169–200.

Emre, M. (2003) 'What causes mental dysfunction in Parkinson's disease?', Movement disorders: official journal of the Movement Disorder Society, 18 Suppl 6(S6), pp. S63–71. Available at: https://doi.org/10.1002/mds.10565.

Etoom, M.; Alwardat, M.; Aburub, A.S.; Lena, F.; Fabbrizo, R.; Modugno, N.; Centonze, D. Therapeutic interventions for Pisa syndrome in idiopathic Parkinson's disease. A Scoping Systematic Review. Clin. Neurol. Neurosurg. 2020, 198, 106242. https://doi.org/10.1016/j.clineuro.2020.106242.

Evgeniou, T.; Pontil, M. Support vector machines: Theory and applications. In Advanced Course on Artificial Intelligence, Springer: Berlin/Heidelberg, Germany, 1999; pp. 249–257.

Explainable Emotion Recognition Using Xception- Based Feature Extraction and Supervised Machine Learning on the RAVDESS Dataset

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & others. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, 7(2), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

Eyben, F., Wöllmer, M., Schuller, B. (2010). Opensmile: the Munich versatile and fast open-source audio feature extractor. Proceedings of the 18th ACM International Conference on Multimedia, 1459–1462. https://doi.org/10.1145/1873951.1874246

F. Chollet, "Xception: Deep Learning With Depthwise Separable Convolutions," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258. Accessed: Mar. 07, 2024. [Online]. Available:

https://openaccess.thecvf.com/content_cvpr_2017/html/Chollet_Xception_Deep_Learning_CVPR_2017_paper.html

F. M. Talaat, Z. H. Ali, R. R. Mostafa, and N. El-Rashidy, "Real-time facial emotion recognition model based on kernel autoencoder and convolutional neural network for autism children," Soft Comput, vol. 28, no. 9, pp. 6695–6708, May 2024, doi: 10.1007/s00500-023-09477-y.

Fanos, V., Dessì, A., Deledda, L., Lai, A., Ranzi, P., Avellino, I., ... & Colangelo, A. (2023). Postpartum depression screening through artificial intelligence: preliminary data through the Talking About algorithm. Journal of Paediatric and Neonatal Individualised Medicine, 12(2), 1-11.

Farrelly, S., & Lester, H. (2014). Therapeutic relationships between mental health service users with psychotic disorders and their clinicians: A critical interpretive synthesis. Health & Social Care in the Community, 22(5), 449–460. https://doi.org/10.1111/hsc.12090 SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Faurholt-Jepsen, M., Rohani, D. A., Busk, J., Vinberg, M., Bardram, J. E., & Kessing, L. V. (2021). Voice analyses using smartphone-based data in patients with bipolar disorder, unaffected relatives and healthy control individuals, and during different affective states. In International Journal of Bipolar Disorders (Vol. 9, Issue 1). https://doi.org/10.1186/s40345-021-00243-3

Fawcett, T. An introduction to ROC analysis. Pattern Recognit Lett. 2006, 27, 861–874.

Fayaz, S., Shah, S.Z.A., Din, N.M., Gul, N. & Assad, A. (2024). Advancements in data augmentation and transfer learning: a comprehensive survey to address data scarcity challenges. Recent Advances in Computer Science and Communications 17, 14–35.

Fellman, D. et al. (2020) 'Training working memory updating in Parkinson's disease: A randomised controlled trial', Neuropsychological rehabilitation, 30(4), pp. 673–708. Available at: https://doi.org/10.1080/09602011.2018.1489860.

Filippa, M. & Kuhn, P. (2024). Early parental vocal contact in neonatal units: rationale and clinical guidelines for implementation. Frontiers in Neurology 15, 1441576

Filippa, M., Della Casa, E., D'amico, R., Picciolini, O., Lunardi, C., Sansavini, A. & Ferrari, F. (2021) Effects of early vocal contact in the neonatal intensive care unit: study protocol for a multi-centre, randomised clinical trial. International Journal of Environmental Research and Public Health 18, 3915.

Filler, T., Jameel, B., & Gagliardi, A. R. (2020). Barriers and facilitators of patient-centred care for immigrant and refugee women: A scoping review. BMC Public Health, 20(1), 859. https://doi.org/10.1186/s12889-020-09159-6

Fox, G., Fergusson, D. A., Daham, Z., Youssef, M., Foster, M., Poole, E., ... & Lalu, M. M. (2021). Patient engagement in preclinical laboratory research: A scoping review. EBioMedicine, 70, 103484. https://doi.org/10.1016/j.ebiom.2021.103484

Frank, M.C. (2020).Quantifying sources of variability in infancy research using the infant-directed-speech preference. Advances in Methods and Practices in Psychological Science 3, 24–52.

Fu, M., Li, D., Gadhiya, A., Lambright, B., Alowais, M., Bahnassy, M., Elletter, S.E.D., Toyin, H.O., Jiang, H., Zhang, K., Aldarmaki, H. (2025). Infant cry detection using causal temporal representation. In ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Gaffney, H. J., & Hamiduzzaman, M. (2022). Factors that influence older patients' participation in clinical communication within developed country hospitals and GP clinics: A systematic review of current literature. PLoS ONE, 17(6), e0269840. https://doi.org/10.1371/journal.pone.0269840

Gagliardi, A. R., Nyhof, B. B., Dunn, S., Grace, S. L., Green, C., Stewart, D. E., & Wright, F. C. (2019). How is patient-centred care conceptualised in women's health: A scoping review. BMC Women's Health, 19(1), 7. https://doi.org/10.1186/s12905-019-0852-9

Galatzer-Levy, I., Abbas, A., Ries, A., Homan, S., Sels, L., Koesmahargyo, V., Yadav, V., Colla, M., Scheerer, H., Vetter, S., Seifritz, E., Scholz, U., & Kleim, B. (2021). Preliminary support for visual and auditory digital markers of suicidality in acutely suicidal psychiatric inpatients: Proof-of-concept study. Journal of Medical Internet Research, 23(6). https://doi.org/10.2196/25199

Ganti, V.K.A.T. (2025). Beyond the stethoscope: how artificial intelligence is redefining diagnosis, treatment, and patient care in the 21st century. Deep Science Publishing.

Gartner, J. B., Abasse, K. S., Bergeron, F., Landa, P., Lemaire, C., & Côté, A. (2022). Definition and Conceptualisation of the patient-centered care pathway: A proposed integrative framework for consensus. BMC Health Services Research, 22(1), 790. https://doi.org/10.1186/s12913-022-07960-0 SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Gavelin, H.M. et al. (2020) 'Cognition-oriented treatments for older adults: A systematic overview of systematic reviews', Neuropsychology review, 30(2), pp. 167–193. Available at: https://doi.org/10.1007/s11065-020-09434-8.

Gavelin, H.M. et al. (2022) 'Computerised cognitive training in Parkinson's disease: A systematic review and meta-analysis', Ageing research reviews, 80(101671), p. 101671. Available at: https://doi.org/10.1016/j.arr.2022.101671.

Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M. & Ritter, M. (2017) Audio set: an ontology and human-labelled dataset for audio events. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 776–780.

Gentile, F.R., Shah, S.T.H., Sperti, M., Panagiotopoulos, K., Primi, R., Bendotti, S., Currao, A., Compagnoni, S., Baldi, E., Lopiano, C., Vicini Scajola, L., Marconi, G., Deriu, M.A., & Savastano, S. (2023.) An innovative medical decision support tool for neurological outcome prediction from post-resuscitation electrocardiograms (MILESTONE). European Heart Journal 44, ehad655.650.

George, S. M., & Ilyas, P. M. (2024). A review on speech emotion recognition: a survey, recent advances, challenges, and the influence of noise. Neurocomputing, 568, 127015.

Géron, A. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems; "O'Reilly Media, Inc.": Sebastopol, CA, USA, 2022.

Gichoya, J.W., Thomas, K., Celi, L.A., Safdar, N., Banerjee, I., Banja, J.D., Seyyed-Kalantari, L., Trivedi, H., & Purkayastha, S. (2023). AI pitfalls and what not to do: mitigating bias in AI. British Journal of Radiology 96, 20230023.

Giddens, C. L., Barron, K. W., Byrd-Craven, J., Clark, K. F., & Winter, A. S. (2013). Vocal indices of stress: A review. Journal of Voice, 27(3), 390.e21-390.e29. https://doi.org/10.1016/j.jvoice.2012.12.010

Giovagnoli, A.R. et al. (1996) 'Trail making test: normative values from 287 normal adult controls', Italian journal of neurological sciences, 17(4), pp. 305–309. Available at: https://doi.org/10.1007/bf01997792.

Girotti, F. et al. (1988) 'Dementia and cognitive impairment in Parkinson's disease', Journal of neurology, neurosurgery, and psychiatry, 51(12), pp. 1498–1502. Available at: https://doi.org/10.1136/jnnp.51.12.1498.

Giustiniani, A. et al. (2022) 'Effects of cognitive rehabilitation in Parkinson disease: a meta-analysis', Neurological sciences: official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology, 43(4), pp. 2323–2337. Available at: https://doi.org/10.1007/s10072-021-05772-4.

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M. B., Dodel, R., & others. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. Movement Disorders: Official Journal of the Movement Disorder

Society, 23(15), 2129–2170. https://doi.org/10.1002/mds.22340

Goetz, C.G.; Poewe, W.; Rascol, O.; Sampaio, C.; Stebbins, G.T.; Counsell, C.; Giladi, N.; Holloway, R.G.; Moore, C.G.; Wenning, G.K.; et al. Movement Disorder Society Task Force report on the Hoehn and Yahr staging scale: Status and recommendations. Mov. Disorders. 2004, 19, 1020–1028.

Golubinski, V., Oppel, E. M., & Schreyögg, J. (2020). A systematic scoping review of psychosocial and psychological factors associated with patient activation. Patient Education and Counselling, 103(10), 2061–2068. https://doi.org/10.1016/j.pec.2020.05.005

Gómez-Vilda, P., Gómez-Rodellar, A., Palacios-Alonso, D., Rodellar-Biarge, V. & Álvarez-Marquina, A. (2022). The role of data analytics in the assessment of pathological speech—a critical appraisal. Applied Sciences 12, 11095.

Gorriz, J.M.; Segovia, F.; Raírez, J.; Ortiz, A.; Suckling, J. Is K-fold cross-validation the best model selection method for Machine Learning? arXiv 2024, arXiv:2401.16407

Gournay, P., Lahaie, O., & Lefebvre, R. (2018, Juin). A Canadian French emotional speech dataset. In Proceedings of the 9th ACM Multimedia Systems Conference (pp. 399-402).

Graffigna, G., & Barello, S. (2018). Spotlight on the patient health engagement model (PHE model): A psychosocial theory to understand people's meaningful engagement in their own health care. Patient Preference and Adherence, 12, 1261–1271. https://doi.org/10.2147/PPA.S145646

Graffigna, G., Barello, S., & Triberti, S. (2015). Patient engagement handbook. CreateSpace Independent Publishing Platform. Retrieved from https://www.researchgate.net/publication/287808329

Gratch, J., Artstein, R., Lucas, G. M., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., & others. (2014). The distress analysis interview corpus of human and computer interviews. LREC, 3123–3128. http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf

Greene, J., & Hibbard, J. H. (2012). Why does patient activation matter? An examination of the relationships between patient activation and health-related outcomes. Journal of General Internal Medicine, 27(5), 520–526. https://doi.org/10.1007/s11606-011-1931-2

Greenhalgh, T., Robert, G., Macfarlane, F., Bate, P., Kyriakidou, O., & Peacock, R. (2005). Storylines of research in diffusion of innovation: a meta-narrative approach to systematic review. Social Science & Medicine, 61(2), 417–430. https://doi.org/10.1016/j.socscimed.2004.12.001

Groene, O., Klazinga, N., Wagner, C., Arah, O. A., Thompson, A., Bruneau, C., & Suñol, R. (2010).

Investigating organisational quality improvement systems, patient empowerment, organisational culture, professional involvement and the quality of care in European hospitals: The "Deepening our Understanding of Quality Improvement in Europe (DUQuE)" project. BMC Health Services Research, 10, 281. https://doi.org/10.1186/1472-6963-10-281

Grosman, J. (2021). Fine-tuned XLSR-53 large model for speech recognition in Italian. Hugging Face. https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-italian.

Grossi, E.; Buscema, M. Introduction to artificial neural networks. Eur. J. Gastroenterol Hepatol. 2007, 19, 1046–1054.

Grossi, G., Perski, A., Osika, W., & Savic, I. (2015). Stress-related exhaustion disorder – Clinical manifestation of burnout, as a focal but bounded construct,? Scandinavian Journal of Psychology, 56(6), 626–636. https://doi.org/10.1111/sjop.12251

Grządzielewska, M. (2021). Using machine learning in burnout, as a focal but bounded construct, prediction: A survey. Child and Adolescent Social Work Journal, 38(2), 175–180. https://doi.org/10.1007/s10560-020-00695-0

Guarrasi, V., Aksu, F., Caruso, C.M., Di Feola, F., Rofena, A., Ruffini, F., & Soda, P. (2025). A systematic review of intermediate fusion in multimodal deep learning for biomedical applications. Image and Vision Computing 158, 105509.

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. Field Methods, 18(1), 59–82. https://doi.org/10.1177/1525822X05279903

Guglietti, B., Hobbs, D. and Collins-Praino, L.E. (2021) 'Optimising cognitive training for the treatment of cognitive dysfunction in Parkinson's disease: Current limitations and future directions', Frontiers in ageing neuroscience, 13, p. 709484. Available at: https://doi.org/10.3389/fnagi.2021.709484.

Hamilton, M. (1960). A rating scale for depression. Journal of Neurology, Neurosurgery, and Psychiatry, 23(1), 56 –62. https://doi.org/10.1136/jnnp.23.1.56

Hammers, D.B. et al. (2020) 'A survey of international clinical teleneuropsychology service provision prior to and in the context of COVID-19', The clinical neuropsychologist, 34(7-8), pp. 1267–1283. Available at: https://doi.org/10.1080/13854046.2020.1810323. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Hammoud, M., Getahun, M.N., Baldycheva, A. & Somov, A. (2024) Machine learning-based infant crying interpretation. Frontiers in Artificial Intelligence 7,1337356.

Hansen, L., Zhang, Y. P., Wolf, D., Sechidis, K., Ladegaard, N., & Fusaroli, R. (2022). A generalizable speech emotion recognition model reveals depression and remission. Acta Psychiatrica Scandinavica, 145(2), 186–199. https://doi.org/10.1111/acps.13388

Harati, S., Crowell, A., Mayberg, H., and Nemati, S. (2018). Depression Severity Classification from Speech Emotion. 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 5763-5766. doi: 10.1109/EMBC.2018.8513610.

Harel, B.; Cannizzaro, M.; Snyder, P.J. Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study. Brain Cogn. 2004, 56, 24–29.

Harrington, R. L., Hanna, M. L., Oehrlein, E. M., Camp, R., Wheeler, R., Cooblall, C., ... & Perfetto, E. M. (2020). Defining patient engagement in research: Results of a systematic review and analysis. Value in Health, 23(6), 677–688. https://doi.org/10.1016/j.jval.2020.01.019

Harrison, J. D., Auerbach, A. D., Anderson, W., Fagan, M., Carnie, M., Hanson, C., ... & Weiss, R. (2019). Patient stakeholder engagement in research: A narrative review to describe foundational principles and best practice activities. Health Expectations, 22(3), 307–316. https://doi.org/10.1111/hex.12873

Hassani, H., Royer-Carenzi, M., Mashhad, L.M., Yarmohammadi, M. & Yeganegi, M.R. (2024).Exploring the depths of the autocorrelation function: its departure from normality. Information 15, 449.

Hawi, S.; Alhozami, J.; AlQahtani, R.; AlSafran, D.; Alqarni, M.; El Sahmarany, L.Automatic Parkinson's disease detection based on the combination of long-term acoustic features and Mel frequency cepstral coefficients (MFCC). Biomed Signal Process Control. 2022, 78, 104013. https://doi.org/10.1016/j.bspc.2022.104013.

Heinemann, L. V., & Heinemann, T. (2017). burnout, as a focal but bounded construct, research: Emergence and scientific investigation of a contested diagnosis. SAGE Open, 7(1), 2158244017697154. https://doi.org/10.1177/2158244017697154

Herz, N.B. et al. (2013) 'Nintendo Wii rehabilitation ("Wii-hab") provides benefits in Parkinson's disease', Parkinsonism & related disorders, 19(11), pp. 1039–1042. Available at: https://doi.org/10.1016/j.parkreldis.2013.07.014.

Hibbard, J. H., Mahoney, E. R., Stock, R., & Tusler, M. (2007). Do increases in patient activation result in improved self-management behaviours? Health Services Research, 42(4), 1443–1463. https://doi.org/10.1111/j.1475-6773.2006.00669.x SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Higuchi, M., Nakamura, M., Shinohara, S., Omiya, Y., Takano, T., Mizuguchi, D., Sonota, N., Toda, H., Saito, T., So, M., Takayama, E., Terashi, H., Mitsuyoshi, S., & Tokuno, S. (2022). Detection of Major Depressive Disorder Based on a Combination of Voice Features: An Exploratory Approach. International Journal of Environmental Research and Public Health, 19(18). https://doi.org/10.3390/ijerph191811397

Hoehn, M. and Yahr, M. (2011) 'Parkinsonism: Onset, progression, and mortality', Neurology, 77(9), pp. 874–874. Available at: https://doi.org/10.1212/01.wnl.0000405146.06300.91.

Hossain, M.A.; Amenta, F. Machine Learning-Based Classification of Parkinson's Disease Patients Using Speech Biomarkers. J Park. Dis. 2024, 14, 95–109.

Hou, X., Zhang, P., Mo, L., Peng, C., & Zhang, D. (2024). Neonatal sensitivity to vocal emotions: a milestone at 37 weeks of gestational age. eLife 13, RP95393.

Hughes, A.J. et al. (1992) 'Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases', Journal of neurology, neurosurgery, and psychiatry, 55(3), pp. 181–184. Available at: https://doi.org/10.1136/jnnp.55.3.181.

Husain, A., Knake, L., Sullivan, B., Barry, J., Beam, K., Holmes, E., Hooven, T., McAdams, R., Moreira, A., Shalish, W.,& Vesoulis, Z. (2.025). AI models in clinical neonatology: a review of modelling approaches and a consensus proposal for standardised reporting of model performance. Paediatric Research https://doi.org/10.1038/s41390-025-04207-6

Irizarry, T., De Vito Dabbs, A., & Curran, C. R. (2015). Patient portals and patient engagement: A state of the science review. Journal of Medical Internet Research, 17(6), e148. https://doi.org/10.2196/jmir.4255

Iyer, A.; Kemp, A.; Rahmatallah, Y.; Pillai, L.; Glover, A.; Prior, F.; Larson-Prior, L.; Virmani, T. A machine learning method to process voice samples for identification of Parkinson's disease. Sci. Rep. 2023, 13. 20615.

Iyer, A.; Kemp, A.; Rahmatallah, Y.; Pillai, L.; Glover, A.; Prior, F.; Larson-Prior, L.; Virmani, T. A machine learning method to process voice samples for identification of Parkinson's disease. Sci. Rep. 2023, 13. 20615.

J. Li, Y. Tian, and T. Zhou, Healthcare Information Systems: Progress, Challenges and Future Directions. in Innovative Medical Devices. Singapore: Springer Nature, 2024. doi: 10.1007/978-981-97-9551-2.

Jahandideh, S., Kendall, E., Low-Choy, S., Donald, K., & Jayasinghe, R. (2018). The process of patient engagement in cardiac rehabilitation: A model-centric systematic review. Behaviour Change,

35(4), 185–202. https://doi.org/10.1017/bec.2018.20

James, M.; Hastie, P.; Taylor, B. First Printing: July 5, 2023. 2023.

Jeancolas, L.; Benali, H.; Benkelfat, B.-E.; Mangone, G.; Corvol, J.-C.; Vidailhet, M. Automatic detection of early stages of Parkinson's disease through acoustic voice analysis with mel-frequency cepstral coefficients. In Proceedings of the 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Fez, Morocco, 22–24 May 2017; pp. 1–6.

Jeong, Y. & Ha, S. (2025). Early developmental changes in infants' vocal responses in interactions with caregivers—Infant Behaviour and Development 78, 102022.

Jin, Z.; Shang, J.; Zhu, Q.; Ling, C.; Xie, W.; Qiang, B. RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis. In Web Information Systems Engineering–WISE 2020: Proceedings of the 21st International Conference, Part II 21, Amsterdam, The Netherlands, 20–24 October 2020; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; Volume 12343, pp. 503–515.

Joo, S., Choi, J., Kim, N., & Lee, M.C. (2021). Zero-crossing rate method as an efficient tool for combustion instability diagnosis. Experimental Thermal and Fluid Science 123, 110340.

Jørgensen, C. R., Thomsen, T. G., Ross, L., Dietz, S. M., Therkildsen, S., Groenvold, M., ... & Johnsen, A. T. (2018). What facilitates "patient empowerment" in cancer patients during follow-up: A qualitative systematic review of the literature. Qualitative Health Research, 28(2), 292–304. https://doi.org/10.1177/1049732317721477

Jørgensen, K., & Rendtorff, J. D. (2018). Patient participation in mental health care: Perspectives of healthcare professionals: An integrative review. Scandinavian Journal of Caring Sciences, 32(2), 490–501. https://doi.org/10.1111/scs.12531

Kalia, L.V. et al. (2015) 'Clinical correlations with Lewy body pathology in LRRK2-related Parkinson disease', JAMA neurology, 72(1), pp. 100–105. Available at: https://doi.org/10.1001/jamaneurol.2014.2704.

Kamiloğlu, R.G. & Sauter, D.A. (2021) Voice production and perception. In: Oxford Research Encyclopedia of Psychology. Oxford University Press, Oxford.

Kane, P. M., Murtagh, F. E. M., Ryan, K., Mahon, N. G., McAdam, B., McQuillan, R., ... & Daveson, B. A. (2015). The gap between policy and practice: A systematic review of patient-centred care interventions in chronic heart failure. Heart Failure Reviews, 20(6), 673–687. https://doi.org/10.1007/s10741-015-9508-5

Kao, C. & Zhang, Y. (2025). Age and sex differences in infants' neural sensitivity to emotional prosodies in spoken words: a multifeature oddball study. Journal of Speech, Language, and Hearing Research 68, 332–348.

Keles, E. & Bagci, U. (2023). The past, current, and future of neonatal intensive care units with artificial intelligence: a systematic review. NPJ Digital Medicine 6, 220.

Khammissa, R. A. G., Fourie, J., & Lemmer, J. (2022). burnout, as a focal but bounded construct, phenomenon: Neurophysiological factors, clinical features, and aspects of management. Journal of Psychology and Psychotherapy, 12(1), 1–6. https://doi.org/10.35248/2161-0487.22.12.456

Kliem, S., Lohmann, A., Mößle, T., & Brähler, E. (2017). German Beck Scale for Suicide Ideation (BSS): psychometric properties from a representative population survey. BMC Psychiatry, 17 (1), 389. https://doi.org/10.1186/s12888-017-1559-9

Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., & Khudanpur, S. (2017, March). A study on data augmentation of reverberant speech for robust speech recognition. In 2017, the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 5220-5224). IEEE.

Kotsiantis, S.B.; Zaharakis, I.D.; Pintelas, P.E. Machine learning: A review of classification and combining techniques. Artif Intell Rev. 2006, 26, 159–190.

Kouba, P., Šmotek, M., Tichý, T., & Kopřivová, J. (2023). Detection of air traffic controllers' fatigue using voice analysis - An EEG validation study. International Journal of Industrial Ergonomics, 95(November 2021). https://doi.org/10.1016/j.ergon.2023.103442

Kraus, B., Zinbarg, R., Braga, R.M., Nusslock, R., Mittal, V.A. & Gratton, C. (2023) Insights from personalised models of brain and behaviour for identifying biomarkers in psychiatry. Neuroscience and BiobehaviouralReviews 152, 105259.

Kristian, Y., Simogiarto, N., Sampurna, M.T.A., Hanindito, E. & Visuddho, V. (2023) Ensemble of multimodal deep learning autoencoder for infant cry and pain detection. F1000Research 11, 359.

Krones, F., Marikkar, U., Parsons, G., Szmul, A. & Mahdi, A. (2025) Review of multimodal machine learning approaches in healthcare. Information Fusion 114, 102690.

Kumar Nukala, V., Reddy Motheline, S., Wesley Kolasanakoti, J., Vankayalapati, S., Velupula, V. & Reddy Dodda, V. (2024) Advanced machine learning approaches for infant cry classification using audio feature extraction. In, 2024 International Conference on Integrated Intelligence and Communication Systems (ICIICS), pp.1–7.

L. A. Ferreira et al., "Disclosing neonatal pain in real-time: AIderived pain sign from continuous assessment of facial expressions," Computers in Biology and Medicine, vol. 189, p. 109908, May

2025, doi: 10.1016/j.compbiomed.2025.109908.

L. Jibb and J. Stinson, "Pain Assessment," in Managing Pain in Children and Young People, 2024, pp. 73–93. doi: 10.1002/9781119645641.ch6. "Pain Assessment in Neonatal Clinical Practice via Facial Expression Analysis and Deep Learning | SpringerLink." Accessed: Jan. 30, 2025. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-64636-2_19

L. Liao, S. Wu, C. Song, and J. Fu, "RS-Xception: A Lightweight Network for Facial Expression Recognition," Electronics, vol. 13, no.16, Art. no. 16, Jan. 2024, doi: 10.3390/electronics13163217.

La Rosa, V.L., Geraci, A., Iacono, A. & Commodari, E. (2024) Affective touch in preterm infant development: neurobiological mechanisms and implications for child–caregiver attachment and neonatal care. Children 11, 1407.

Langer, M., König, C. J., Siegel, R., Fredenhagen, T., Schunck, A. G., Hähne, V., & Baur, T. (2022). Vocal-stress diary: A longitudinal investigation of the association of everyday work stressors and human voice features. Psychological Science, 33(7), 1027–1039. https://doi.org/10.1177/09567976221088347

Lee, J.H., Lee, G.W., Bong, G., Yoo, H.J. & Kim, H.K. (2020) Deep-learning-based detection of infants with autism spectrum disorder using auto-encoder feature representation. Sensors 20, 6762.

Leung, I.H. et al. (2015) 'Cognitive training in Parkinson disease: a systematic review and meta-analysis', Neurology, 85, pp. 1843-1851. Available at: https://doi.org/10.1212/WNL.0000000000002145.

Liang, L., Cako, A., Urquhart, R., Straus, S. E., Wodchis, W. P., Baker, G. R., & Gagliardi, A. R. (2018). Patient engagement in hospital health service planning and improvement: A scoping review. BMJ Open, 8(1), e018263. https://doi.org/10.1136/bmjopen-2017-018263 SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Lin, Y., Liyanage, B. N., Sun, Y., Lu, T., Zhu, Z., Liao, Y., Wang, Q., Shi, C., & Yue, W. (2022). A deep learning-based model for detecting depression in senior population. Frontiers in Psychiatry, 13. https://doi.org/10.3389/fpsyt.2022.1016676

Little, M.; McSharry, P.; Hunter, E.; Spielman, J.; Ramig, L. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Trans. Biomed. Eng. 2008, 56, 1015–1022.

Little, M.; McSharry, P.; Hunter, E.; Spielman, J.; Ramig, L. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. Nat. Precedings. 2008, 1. https://doi.org/10.1038/npre.2008.2298.1.

Little, M.; McSharry, P.; Roberts, S.; Costello, D.; Moroz, I. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. BioMedical Eng. OnLine 2007, 6, 23.

Litvan, I., Goldman, J.G. and Troster, A.I. (2012) 'Diagnostic criteria for mild cognitive impairment in Parkinson's disease: Movement Disorder Society Task Force guidelines, Mov', Mov. Disord, 27, pp. 349–356.

Liu, Q., Song, L., Xu, D. & Long, Y. (2025) ICSD: an open-source dataset for infant cry and snoring detection.arXiv.org. Preprint.

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one, 13(5), e0196391.

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS One, 13(5), e0196391. https://doi.org/10.1371/journal.pone.0196391

Long, B., Xiang, V., Stojanov, S., Sparks, R.Z., Yin, Z., Keene, G.E., Tan, A.W.M., Feng, S.Y., Zhuang, C., Marchman, V.A., Yamins, D.L.K. & Frank, M.C.(2024) The BabyView dataset: high-resolution egocentric videos of infants' and young children's everyday experiences.arXiv.org. Preprint.

Long, H.L., Eichorn, N. & Oller, D.K. (2023) A probe study on vocal development in two infants at risk for cerebral palsy. Developmental Neurorehabilitation 26, 44–51.

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. In Laryngoscope Investigative Otolaryngology (Vol. 5, Issue 1, pp. 96–116). John Wiley and Sons Inc. https://doi.org/10.1002/lio2.354Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. Laryngoscope Investigative Otolaryngology, 5(1), 96–116. https://doi.org/10.1002/lio2.354

Luz, S., Haider, F., de la Fuente Garcia, S., Fromm, D., & Macwhinney, B. (2020). Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge. https://doi.org/10.21437/Interspeech.2020-2571

M. Ouhammou, N. Ababou, M. Baslam, and S. L. Aouragh, "Deep Facial Expression Recognition Using Xception Model," in Arabic Language Processing: From Theory to Practice, B. Hdioud and S. L.

Aouragh, Eds., Cham: Springer Nature Switzerland, 2025, pp. 209–220. doi: 10.1007/978-3-031-80438-0_16.

M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session, pp. 97–101, Feb. 2016, doi: 10.18653/v1/n16-3020.

Ma, A.; Lau, K.K.; Thyagarajan, D. Voice changes in Parkinson's disease: What are they telling us? J. Clin. Neurosci. 2020, 72, 1–7. https://doi.org/10.1016/j.jocn.2019.12.029.

Ma, F., Li, Y., Xie, Y., He, Y., Zhang, Y., Ren, H., Liu, Z., Yao, W., Ren, F., Yu, F.R. & Ni, S. (2024) A review of human emotion synthesis based on generative technology.IEEE Transactions on Affective Computing (Preprint).

Ma, Z., Chen, M., Zhang, H., Zheng, Z., Chen, W., Li, X., ... & Hain, T. (2024). Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark. arXiv preprint arXiv:2406.07162.

MacLachlan, M., & McVeigh, J. (Eds.) (2021). Macropsychology: A population science for sustainable development goals. Springer Nature. https://doi.org/10.1007/978-3-030-70198-6

MacLachlan, M., McVeigh, J., Hogan, M. J., & McAuliffe, E. (2019). Macropsychology: Challenging and changing social structures and systems to promote social inclusion. In G. Riva, B. K. Wiederhold, & M. Cipresso (Eds.), The SAGE handbook of applied social psychology (pp. 166–182). SAGE Publications. https://doi.org/10.4135/9781526426044.n12

Madanian, S., Chen, T., Adeleye, O., Templeton, J. M., Poellabauer, C., Parry, D., & Schneider, S. L. (2023). Speech emotion recognition using machine learning — A systematic review. In Intelligent Systems with Applications (Vol. 20). Elsevier B.V. https://doi.org/10.1016/j.iswa.2023.200266

Madanian, S., Parry, D., Adeleye, O., Poellabauer, C., Mirza, F., Mathew, S., & Schneider, S. (2022, January). Automatic speech emotion recognition using machine learning: digital transformation of mental health. In Proceedings of the Annual Pacific Asia Conference on Information Systems (PACIS).

Maggio, M.G. et al. (2018) 'What about the role of virtual reality in Parkinson's disease's cognitive rehabilitation? Preliminary findings from a randomised clinical trial, Journal of geriatric psychiatry and neurology, 31(6), pp. 312–318. Available at: https://doi.org/10.1177/0891988718807973.

Maggio, M.G. et al. (2024) 'Effectiveness of telerehabilitation plus virtual reality (Tele-RV) in cognitive e social functioning: A randomised clinical study on Parkinson's disease', Parkinsonism & related

disorders, 119, p. 105970. Available at: https://doi.org/10.1016/j.parkreldis.2023.105970.

Magrinelli, F.; Picelli, A.; Tocco, P.; Federico, A.; Roncari, L.; Smania, N.; Zanette, G.; Tamburin, S. Pathophysiology of Motor Dysfunction in Parkinson's Disease as the Rationale for Drug Treatment and Rehabilitation. Park. Dis. 2016, 2016, 9832839.

Mallamaci, R.; Musarò, D.; Greco, M.; Caponio, A.; Castellani, S.; Munir, A.; Guerra, L.; Damato, M.; Fracchiolla, G.; Coppola, C.; et al. Dopamine- and Grape-Seed-Extract-Loaded Solid Lipid Nanoparticles: Interaction Studies between Particles and Differentiated SH-SY5Y Neuronal Cell Model of Parkinson's Disease. Molecules 2024, 29, 1774.

Mallegni, N., Molinari, G., Ricci, C., Lazzeri, A., La Rosa, D., Crivello, A., & Milazzo, M. (2022). Sensing devices for detecting and processing acoustic signals in healthcare. Biosensors 12, 835.

Manafo, E., Petermann, L., Mason-Lai, P., & Vandall-Walker, V. (2018). Patient engagement in Canada: A scoping review of the "how" and "what" of patient engagement in health research. Health Research Policy and Systems, 16(1), 5. https://doi.org/10.1186/s12961-018-0282-4

Manafò, E., Petermann, L., Vandall-Walker, V., & Mason-Lai, P. (2018). Patient and public engagement in priority setting: A systematic rapid review of the literature. PLoS ONE, 13(3), e0193579. https://doi.org/10.1371/journal.pone.0193579

Mantri, S.; Morley, J.F. Prodromal and early Parkinson's disease diagnosis. Pr. Neurol. 2018, 35, 28–31.

Marschik, P.B., Widmann, C.A.A., Lang, S., Kulvicius, T., Boterberg, S., Nielsen-Saines, K., Bölte, S., Esposito, G., Nordahl-Hansen, A., Roeyers, H., Wörgötter, F., Einspieler, C., Poustka, L. & Zhang, D. (2022.a) Emerging verbal functions in early infancy: lessons from observational and computational approaches on typical development and neurodevelopmental disorders. Advances in Neurodevelopmental Disorders 6, 369–388.

Marschik, P.B., Widmann, C.A.A., Lang, S., Kulvicius, T., Boterberg, S., Nielsen-Saines, K., Bölte, S., Esposito, G., Nordahl-Hansen, A., Roeyers, H., Wörgötter, F., Einspieler, C., Poustka, L. & Zhang, D. (2022b) Emerging verbal functions in early infancy: lessons from observational and computational approaches on typical development and neurodevelopmental disorders. Advances in Neurodevelopmental Disorders 6, 369–388.

Marwan, N., Romano, M.C., Thiel, M. & Kurths, J. (20.07) Recurrence plots for the analysis of complex systems. Physics Reports 438, 237-329

Marzban, S., Najafi, M., Agolli, A., & Ashrafi, E. (2022). Impact of patient engagement on healthcare quality: A scoping review. Journal of Patient Experience, 9(1), 1–12.

https://doi.org/10.1177/23743735221125439

McCue, M., Fairman, A. and Pramuka, M. (2010) 'Enhancing quality of life through telerehabilitation', Physical Medicine and Rehabilitation Clinics of North America, 21(1), pp. 195–205. Available at: https://doi.org/10.1016/j.pmr.2009.07.005.

Meissen, F., Breuer, S., Knolle, M., Buyx, A., Müller, R., Kaissis, G., Wiestler, B., & Rückert, D. (2024). Predictable performance bias in unsupervised anomaly detection. eBioMedicine 101,105002.

Menear, M., Dugas, M., Careau, E., Chouinard, M. C., Dogba, M. J., Gagnon, M. P., ... & Légaré, F. (2020). Strategies for engaging patients and families in collaborative care programs for depression and anxiety disorders: A systematic review. Journal of Affective Disorders, 263, 528–539. https://doi.org/10.1016/j.jad.2019.11.008

Menichetti, J., Graffigna, G., & Steinsbekk, A. (2018). What are the contents of patient engagement interventions for older adults? A systematic review of randomised controlled trials. Patient Education and Counselling, 101(6), 995–1005. https://doi.org/10.1016/j.pec.2017.12.009

Miao, X., Li, Y., Wen, M., Liu, Y., Julian, I. N., & Guo, H. (2022). Fusing features of speech for depression classification based on higher-order spectral analysis. Speech Communication, 143(October 2021), 46–56. https://doi.org/10.1016/j.specom.2022.07.006

Min, S., Shin, D., Rhee, S. J., Park, C. H. K., Yang, J. H., Song, Y., Kim, M. J., Kim, K., Cho, W. I., Kwon, O. C., Ahn, Y. M., & Lee, H. (2023). Acoustic Analysis of Speech for Screening for Suicide Risk: Machine Learning Classifiers for Between- and Within-Person Evaluation of Suicidality. Journal of Medical Internet Research, 25. https://doi.org/10.2196/45456

Minafra, B. et al. (2014) 'Eight-year failure of subthalamic stimulation rescued by globus pallidus implant', Brain stimulation, 7(2), pp. 179–181. Available at: https://doi.org/10.1016/j.brs.2013.12.011.

Monaco, M. et al. (2015) 'Erratum to: Forward and backward span for verbal and visuospatial data: standardisation and normative data from an Italian adult population', Neurol Sci, 36, pp. 345–347.

Mosca, I.E. et al. (2020) 'Analysis of feasibility, adherence, and appreciation of a newly developed Tele-rehabilitation program for people with MCI and VCI', Frontiers in Neurology, 11, p. 583368. Available at: https://doi.org/10.3389/fneur.2020.583368.

Murphy, M.M., Colquitt, G.T., Ryals, P.S., Shin, K., Kjeldsen, W.C., McIntyre, A., Whitten, S.V.W., Modlesky, C.M., & Maitre, N.L. (2025). Synergies, discrepancies, and action priorities: a statewide engagement study to strengthen clinical research in cerebral palsy. Health Expectations 28, e70257.

Muzammel, M., Salam, H., & Othmani, A. (2021). End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis. Computer Methods and Programs in Biomedicine, 211, 106433. https://doi.org/10.1016/j.cmpb.2021.106433

Naamanka, E. et al. (2024) 'Effectiveness of teleneuropsychological rehabilitation: Systematic review of randomised controlled trials', Journal of the International Neuropsychological Society: JINS, 30(3), pp. 295–312. Available at: https://doi.org/10.1017/S1355617723000565.

Nadon, L., De Beer, L. T., & Morin, A. J. S. (2022). Should burnout, as a focal but bounded construct, be conceptualised as a mental disorder? Behavioural Sciences, 12(3), 82. https://doi.org/10.3390/bs12030082

Naismith, S.L. et al. (2013) 'Improving memory in Parkinson's disease: a healthy brain ageing cognitive training program', Movement disorders: official journal of the Movement Disorder Society, 28(8), pp. 1097–1103. Available at: https://doi.org/10.1002/mds.25457.

Naranjo, L.; Pérez, C.J.; Martín, J.; Campos-Roca, Y. A two-stage variable selection and classification approach for Parkinson's disease detection by using voice recording replications. Comput. Methods Programs Biomed 2017, 142, 147–156. https://doi.org/10.1016/j.cmpb.2017.02.019.

Narayanan, D.Z., Takahashi, D.Y., Kelly, L.M., Hlavaty, S.I., Huang, J., & Ghazanfar, A.A. (2022). Prenatal development of neonatal vocalisations. eLife 11, e78485.

Narayanan, S.P., Manikandan, M.S., & Cenkeramaddi, L.R. (2024). Fast autocorrelation feature-based infant cry detector for resource-efficient, affordable edge cry sound analysis systems. In 2024 IEEE 19th Conference on Industrial Electronics and Applications (ICIEA), pp.1–6.

Natraj, S., Kojovic, N., Maillart, T. & Schaer, M. (2024) Video-audio neural network ensemble for comprehensive screening of autism spectrum disorder in young children. PLOS ONE 19, e0308388.

Newman, B., Chauhan, A., Holly, E., Jiadai, L., Merrilyn, W., & Stephen, W. (2021). Do patient engagement interventions work for all patients? A systematic review and realist synthesis of interventions to enhance patient safety. Health Expectations, 24(6), 1974–1991. https://doi.org/10.1111/hex.13343

Ng, M. M., Firth, J., Minen, M., & Torous, J. (2019). User engagement in mental health apps: A review of measurement, reporting, and validity. Psychiatric Services, 70(7), 538–544. https://doi.org/10.1176/appi.ps.201800519 SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Novelli, G. et al. (1986) 'Tre test clinici di ricerca e produzione lessicale'. Taratura su soggetti normali.

[Three clinical tests to research and rate the lexical performance of normal subjects, Arch Psicol Neurol Psichiatr, 47, pp. 477–506.

Nussbaum, C., Frühholz, S. & Schweinberger, S.R. (2025) Understanding voice naturalness. Trends in Cognitive Sciences 29, 467–480.

O'Sullivan, S.B.; Schmitz, T.J. Physical Rehabilitation, 5th ed.; Davis Company: Philadelphia, PA, USA, 2007.

Ocloo, J., Garfield, S., Franklin, B. D., & Dawson, S. (2021). Exploring the theory, barriers and enablers for patient and public involvement across health, social care and patient safety: A systematic review of reviews. Health Research Policy and Systems, 19(1), 8. https://doi.org/10.1186/s12961-020-00644-3

Oganian, Y.; Bhaya-Grossman, I.; Johnson, K.; Chang, E.F. Vowel and formant representation in the human auditory speech cortex. Neuron. 2023, 111, 2105–2118.e4.

Oktay, L. A., Abuelgasim, E., Abdelwahed, A., Houbby, N., Lampridou, S., Normahani, P., ... & Jaffer, U. (2021). Factors affecting engagement in web-based health care patient information: Narrative review of the literature. Journal of Medical Internet Research, 23(9), e19896. https://doi.org/10.2196/19896

Onciul, R., Tataru, C.-I., Dumitru, A.V., Crivoi, C., Serban, M., Covache-Busuioc, R.-A., Radoi, M.P., & Toader, C. (2025). Artificial intelligence and neuroscience: transformative synergies in brain research and clinical applications. Journal of Clinical Medicine 14, 550. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Oostendorp, L. J., Durand, M. A., Lloyd, A., & Elwyn, G. (2015). Measuring organisational readiness for patient engagement (MORE): An international online Delphi consensus study. BMC Health Services Research, 15(1), 61. https://doi.org/10.1186/s12913-015-1038-8 SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Ophey, A. et al. (2020) 'Effects of working memory training in patients with Parkinson's disease without cognitive impairment: A randomised controlled trial', Parkinsonism & related disorders, 72, pp. 13–22. Available at: https://doi.org/10.1016/j.parkreldis.2020.02.002.

Orgeta, V. et al. (2020) 'Cognitive training interventions for dementia and mild cognitive impairment in Parkinson's disease', Cochrane database of systematic reviews, 2(2), p. CD011961. Available at: https://doi.org/10.1002/14651858.CD011961.pub2.

Ozcan, T. & Gungor, H. (202.5) Baby cry classification using structure-tuned artificial neural networks with data augmentation and MFCC features. Applied Sciences 15, 2648.

Ozkanca, Y., Göksu Öztürk, M., Ekmekci, M. N., Atkins, D. C., Demiroglu, C., & Hosseini Ghomi, R. (2019). Depression Screening from Voice Samples of Patients Affected by Parkinson's Disease. Digital Biomarkers, 3(2), 72–82. https://doi.org/10.1159/000500354

Paggetti, A. et al. (2024) 'The efficacy of cognitive stimulation, cognitive training, and cognitive rehabilitation for people living with dementia: a systematic review and meta-analysis', GeroScience [Preprint]. Available at: https://doi.org/10.1007/s11357-024-01400-z.

Palmirotta, C. et al. (2024) 'Unveiling the diagnostic potential of linguistic markers in identifying individuals with Parkinson's disease through artificial intelligence: A systematic review', Brain Sciences, 14(2). Available at: https://doi.org/10.3390/brainsci14020137.

París, A.P. et al. (2011) 'Blind randomised controlled study of the efficacy of cognitive training in Parkinson's disease', Movement disorders: official journal of the Movement Disorder Society, 26(7), pp. 1251–1258. Available at: https://doi.org/10.1002/mds.23688.

Parker, G., & Tavella, G. (2022). The diagnosis of burnout, as a focal but bounded construct,: Some challenges. The Journal of Nervous and Mental Disease, 210(7), 475–478. https://doi.org/10.1097/NMD.0000000000000146

Pel-Littel, R. E., Snaterse, M., Teppich, N. M., Buurman, B. M., van Etten-Jamaludin, F. S., van Weert, J. C. M., ... & Scholte op Reimer, W. J. M. (2021). Barriers and facilitators for shared decision making in older patients with multiple chronic conditions: A systematic review. BMC Geriatrics, 21(1), 112. https://doi.org/10.1186/s12877-021-02050-y

Peretz, O.; Koren, M.; Koren, O. Naive Bayes classifier—An ensemble procedure for recall and precision enrichment. Eng. Appl. Artif. Intell. 2024, 136, 108972.

Pérez-Toro, P. A., Arias-Vergara, T., Klumpp, P., Vásquez-Correa, J. C., Schuster, M., Nöth, E., & Orozco-Arroyave, J. R. (2022). Depression assessment in people with Parkinson's disease: The combination of acoustic features and natural language processing. Speech Communication, 145(September), 10–20. https://doi.org/10.1016/j.specom.2022.09.001

Pérez-Toro, P. A., Vásquez-Correa, J. C., Bocklet, T., Noth, E., & Orozco-Arroyave, J. R. (2023). User State Modelling Based on the Arousal-Valence Plane: Applications in Customer Satisfaction and Health-Care. IEEE Transactions on Affective Computing, 14(2), 1533–1546. https://doi.org/10.1109/TAFFC.2021.3112543 SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Petrelli, A. et al. (2014) 'Effects of cognitive training in Parkinson's disease: a randomised controlled trial', Parkinsonism & related disorders, 20(11), pp. 1196–1202. Available at:

https://doi.org/10.1016/j.parkreldis.2014.08.023.

Pfister, T., & Robinson, P. (2010). Speech Emotion Classification and Public Speaking Skill Assessment. In A. A. Salah, T. Gevers, N. Sebe, & A. Vinciarelli (Eds.), Human Behaviour Understanding (pp. 151–162). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-14715-9_15

Piczak, K.J. (2015) ESC: dataset for environmental sound classification. In Proceedings of the 23rd ACM International Conference on Multimedia, MM'15. Association for Computing Machinery, New York, NY, USA, pp.1015–1018.

Pigueiras-del-Real, J., Gontard, L.C., Benavente-Fernández, I., Lubián-López, S.P., Gallero-Rebollo, E., & Ruiz-Zafra, A. (2024a). NRP: a multi-source, heterogeneous, automatic data collection system for infants in neonatal intensive care units. IEEE Journal of Biomedical and Health Informatics 28, 678–689.

Pigueiras-del-Real, J., Ruiz-Zafra, A., Benavente-Fernández, I., Lubián-López, S.P., Shah, S.A.H., Shah, S.T.H. & Gontard, L.C. (2024b) NeoVault: empowering neonatal research through a neonate data hub. BMC Paediatrics 24, 787.

Pinto, J.O. et al. (2024) 'Ecological validity of neuropsychological interventions: A systematic review', Applied neuropsychology. Adult, pp. 1–20. Available at: https://doi.org/10.1080/23279095.2024.2328694.

Piron, L. et al. (2009) 'Exercises for paretic upper limb after stroke: a combined virtual-reality and telemedicine approach', Journal of rehabilitation medicine: official journal of the UEMS European Board of Physical and Rehabilitation Medicine, 41(12), pp. 1016–1102. Available at: https://doi.org/10.2340/16501977-0459.

Poewe, W.; Seppi, K.; Tanner, C.M.; Halliday, G.M.; Brundin, P.; Volkmann, J.; Schrag, A.E.; Lang, A.E. Parkinson's disease. Nat. Rev. Dis. Primers 2017, 3, 1–21.

Powers, D.M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv 2020, http://arxiv.org/abs/2010.16061.

Prediction of Parkinson's Disease Using Long-Term, Short-Term Acoustic Features Based on Machine Learning

Preliminary Qualitative study on AI and burnout, as a focal but bounded construct,, diagnostic potential of vocal biomarkers

Prey, J. E., Woollen, J., Wilcox, L., Sackeim, A. D., Hripcsak, G., Bakken, S., ... & Vawdrey, D. K. (2014). Patient engagement in the inpatient setting: A systematic review. Journal of the American

Medical Informatics Association, 21(4), 742–750. https://doi.org/10.1136/amiajnl-2013-002141

Prior, F.; Virmani, T.; Iyer, A.; Larson-Prior, L.; Kemp, A.; Rahmatallah, Y.; Pillai, L.; Glover, A. Voice Samples for Patients with Parkinson's Disease and Healthy Controls. Available online: https://figshare.com/articles/dataset/Voice_Samples_for_Patients_with_Parkinson_s_Disease_and_Healthy_Controls/23849127 (accessed on 7 July 2025).

R. Guo, H. Guo, L. Wang, M. Chen, D. Yang, and B. Li, "Development and application of emotion recognition technology — a systematic literature review," BMC Psychol, vol. 12, no. 1, p. 95, Feb. 2024, doi: 10.1186/s40359-024-01581-4.

R. Gutierrez, J. Garcia-Ortiz, and W. Villegas-Ch, "Multimodal AI techniques for pain detection: integrating facial gesture and paralanguage analysis," Front. Comput. Sci., vol. 6, Jul. 2024, doi: 10.3389/fcomp. 2024.1424935.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localisation," presented at the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618– 626.

Ramig, L.O.; Fox, C.; Sapir, S. Speech treatment for Parkinson's disease. Expert Rev. Neurother. 2008, 8, 297–309.

Randolph, C. et al. (1998) 'The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): preliminary clinical validity', Journal of clinical and experimental neuropsychology, 20(3), pp. 310–319. Available at: https://doi.org/10.1076/jcen.20.3.310.823.

Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., & Othmani, A. (2022). MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech. Biomedical Signal Processing and Control, 71(PA), 103107. https://doi.org/10.1016/j.bspc.2021.103107

Reyes-Galaviz, O.F., Cano-Ortiz, S.D. & Reyes-García, C.A. (2008) Evolutionary-neural system to classify infant cry units for pathologies identification in recently born babies. In 2008, the Seventh Mexican International Conference on Artificial Intelligence, pp. 330–335.

Ribolsi, M., Fiori Nastro, F., Pelle, M., Medici, C., Sacchetto, S., Lisi, G., Riccioni, A., Siracusano, M., Mazzone, L., & Di Lorenzo, G. (2022). Recognising psychosis in autism spectrum disorder. Frontiers in Psychiatry 13, 768586.

Romo, N., Robb, M.P., Lee, J. & Wermke, K. (2024) Noise phenomena in distress cries of term and very preterm infants at term-equivalent age. Logopaedics Phoniatrics Vocology 50, 48-54.

Rudenko, Y. (2023). Neurophysiological and neuropsychological mechanisms of vocalisation

contrasting with music perception. SSRN Blog.

Ruiz, R., Legros, C., & Guell, A. (1990). Voice analysis to predict the psychological or physical state of a speaker. Aviation, Space, and Environmental Medicine, 61(3), 266–271.

Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161 –1178. https://doi.org/10.1037/h0077714

Rusz, J.; Krupička, R.; Vítečková, S.; Tykalová, T.; Novotný, M.; Novák, J.; Dušek, P.; Růžička, E. Speech and gait abnormalities in motor subtypes of de-novo Parkinson's disease. CNS Neurosci Ther. 2023, 29, 2101–2110.

S. Dutta, V. Shukla, Y. Pant, and V. Tripathi, "Human Psychological Counselling Framework using Computer Vision," in 2024 11th International Conference on Computing for Sustainable Global Development (INDIACom), Feb. 2024, pp. 745–750. doi: 10.23919/INDIACom61295.2024.10498255.

S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems, vol. 2017-December, pp. 4766–4775, May 2017.

S. T. H. Shah et al., "Artificial intelligence coupled with the Internet of Things targeting neurodevelopmental challenges in preterm neonates," Neural Press, vol. 4, no. 1, pp. 32–56, Mar. 2025.

S. T. H. Shah et al., "Code of Explainable Emotion Recognition Using Xception-Based Feature Extraction and Supervised Machine Learning on the RAVDESS Dataset," Apr. 2025, Accessed: Apr. 14, 2025. [Online]. Available: https://zenodo.org/records/15211064

S. T. H. Shah et al., "Data-driven classification and explainable-AI in the field of lung imaging," Front. Big Data, vol. 7, Sep. 2024, doi: 10.3389/fdata.2024.1393758.

S. T. H. Shah, S. A. H. Shah, S. A. Qureshi, A. Di Terlizzi, and M. A. Deriu, "Automated facial characterisation and image retrieval by convolutional neural networks," Front. Artif. Intell., vol. 6, Dec. 2023, doi: 10.3389/frai. 2023.1230383.

Salekin, M.S. (2022). Generative spatio-temporal and multimodal analysis of neonatal pain. Ph.D.thesis University of South Florida, United States – Florida.

Sanchez-Luengos, I. et al. (2021) 'Effectiveness of cognitive rehabilitation in Parkinson's disease: A systematic review and meta-analysis', Journal of personalised medicine, 11(5), p. 429. Available at: https://doi.org/10.3390/jpm11050429.

Sanna, M. (2025). Proprioceptive resonance and multimodal semiotics: readiness to act, embodied cognition, and the dynamics of meaning. NeuroSci 6, 42.

Santangelo, G. et al. (2015) 'Normative data for the Montreal Cognitive Assessment in an Italian population sample', Neurological sciences: official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology, 36(4), pp. 585–591. Available at: https://doi.org/10.1007/s10072-014-1995-y.

Sarrami-Foroushani, P., Travaglia, J., Debono, D., & Braithwaite, J. (2014). Key concepts in consumer and community engagement: A scoping meta-review. BMC Health Services Research, 14(1), 250. https://doi.org/10.1186/1472-6963-14-250 SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Savioni, L., Triberti, S., Durosini, I., Sebri, V., & Pravettoni, G. (2022). Cancer patients' participation and commitment to psychological interventions: A scoping review. Psychology and Health, 37(8), 1022–1055. https://doi.org/10.1080/08870446.2021.1916494

Schapira, A.H.V., Chaudhuri, K.R. and Jenner, P. (2017) 'Non-motor features of Parkinson's disease', Nature Reviews. Neuroscience, 18(8), p. 509. Available at: https://doi.org/10.1038/nrn.2017.91.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. Speech communication, 40(1-2), 227-256.

Schirmer, A. (2018). Is the voice an auditory face? An ALE meta-analysis comparing vocal and facial emotion processing. Social Cognitive and Affective Neuroscience, 13(1), 1-13.

Schirmer, A., Ng, T., Escoffier, N., & Penney, T. B. (2016). Emotional voices distort time: behavioural and neural correlates. Timing & Time Perception, 4(1), 79-98.

Schuller, B., Steidl, S., & Batliner, A. (2009). The Interspeech 2009 emotion challenge. Proc. Interspeech 2009, 312-315, doi: 10.21437/Interspeech.. 2009-103

Schultebraucks, K., Yadav, V., Shalev, A. Y., Bonanno, G. A., & Galatzer-Levy, I. R. (2022). Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilising visual and auditory markers of arousal and mood. In Psychological Medicine (Vol. 52, Issue 5, pp. 957–967). https://doi.org/10.1017/S0033291720002718

Selvan, K., Leekha, A., Abdelmeguid, H., & Malvankar-Mehta, M. S. (2022). Barriers faced by adult refugees to community health and patient engagement: A systematic review. Global Public Health, 17(12), 3412–3425. https://doi.org/10.1080/17441692.2022.2121846

Servello, D. et al. (2023) 'Complications of deep brain stimulation in Parkinson's disease: a single-centre experience of 517 consecutive cases', Acta neurochirurgica, 165(11), pp. 3385–3396. Available at: https://doi.org/10.1007/s00701-023-05799-w. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Shah, S.A.H., Di Terlizzi, A. & Deriu, M.A. (2022). Intelligent system development to monitor neonatal behaviour: a review. In Conference: International Workshop in Neurodevelopmental Impairments in Preterm Children - Computational Advancements (DETERMINED 2022). Ljubljana, Slovenia.

Shah, S.A.H., Shah, S.T.H., Khaled, R., Buccoliero, A., Shah, S.B.H., Di Terlizzi, A., Di Benedetto, G., & Deriu, M.A. (2024). Explainable AI-based skin cancer detection using CNN, particle swarm optimisation and machine learning. Journal of Imaging 10, 332.

Shah, S.T.H., 2025. Multimodal AI tools for predicting neurological and neurodevelopmental trajectories. PhD thesis. Politecnico di Torino, Italy – Torino.

Shah, S.T.H., Shah, S.A.H., Khan, I.I., Imran, A., Shah, S.B.H., Mehmood, A., Qureshi, S.A., Raza, M., Di Terlizzi, A., Cavaglià, M. and Deriu, M.A., 2024. Data-driven classification and explainable AI in the field of lung imaging. Frontiers in Big Data, 7, 1393758.

Shah, S.T.H., Shah, S.A.H., Panagiotopoulos, K., Pigueiras-del-Real, J., Qayyum, K., Shah, S.B.H., Qureshi, S.A., Di Terlizzi, A., Di Benedetto, G., & Deriu, M.A. (2025). Artificial intelligence, coupled with the Internet of Things, is targeting neurodevelopmental challenges in preterm neonates. Journal of Multiscale Neuroscience 4, 32–56.

Shah, S.T.H., Shah, S.A.H., Qureshi, S.A., Di Terlizzi, A., & Deriu, M.A. (2023). Automated facial characterisation and image retrieval by convolutional neural networks. Frontiers in Artificial Intelligence 6, 1230383.

Sheikh, S.A., Sahidullah, M. & Kodrasi, I. (2025). Deep learning for pathological speech: A survey. arXiv preprint.

Shimmin, C., Wittmeier, K. D. M., Lavoie, J. G., Wicklund, E. D., & Sibley, K. M. (2017). Moving towards a more inclusive patient and public involvement in health research paradigm: The incorporation of a trauma-informed intersectional analysis. BMC Health Services Research, 17(1), 539. https://doi.org/10.1186/s12913-017-2463-1 SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Shinohara, S., Nakamura, M., Omiya, Y., Higuchi, M., Hagiwara, N., Mitsuyoshi, S., Toda, H., Saito, T., Tanichi, M., Yoshino, A., & Tokuno, S. (2021). Depressive mood assessment method based on emotion level derived from voice: comparison of voice features of individuals with major depressive disorders and healthy controls. In International Journal of Environmental Research and Public Health (Vol. 18, Issue 10). https://doi.org/10.3390/ijerph18105435

Shippee, N. D., Domecq Garces, J. P., Prutsky Lopez, G. J., Wang, Z., Elraiyah, T. A., Nabhan, M., ... & Murad, M. H. (2015). Patient and service user engagement in research: A systematic review and

synthesised framework. Health Expectations, 18(5), 1151–1166. https://doi.org/10.1111/hex.12090 SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Skorvanek, M.; Martinez-Martin, P.; Kovacs, N.; Rodriguez-Violante, M.; Corvol, J.C.; Taba, P.; Seppi, K.; Levin, O.; Schrag, A.; Foltynie, T.; et al. Differences in MDS-UPDRS Scores Based on Hoehn and Yahr Stage and Disease Duration. Mov Disord Clin Pract. 2017, 4, 536–544.

Snow, M. E. (2022). Patient engagement in healthcare planning and evaluation: A call for social justice. International Journal of Health Planning and Management, 37(S1), 20–31. https://doi.org/10.1002/hpm.3509

Sogaard, M. B., Andresen, K., & Kristiansen, M. (2021). Systematic review of patient-engagement interventions: Potentials for enhancing person-centred care for older patients with multimorbidity. BMJ Open, 11(12), e048558. https://doi.org/10.1136/bmjopen-2020-048558

Spitzer, R. L., Kroenke, K., Williams, J. B. W. (1999). Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Jama, 282(18), 1737–1744. https://doi.org/10.1001/jama.282.18.1737

Suárez-González, A. et al. (2024) 'Rehabilitation Services for Young-Onset Dementia: Examples from High-and Low-Middle-Income Countries', International Journal of Environmental Research and Public Health, 21(6). SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Sullivan, J., Mei, M., Perfors, A., Wojcik, E. and Frank, M.C., 2021. SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. Open Mind 5, 20–29.

Sun, C. and Armstrong, M.J. (2021) 'Treatment of Parkinson's disease with cognitive impairment: Current approaches and future directions', Behavioural Sciences, 11(4), p. 54. Available at: https://doi.org/10.3390/bs11040054.

Suppa, A.; Costantini, G.; Asci, F.; Di Leo, P.; Al-Wardat, M.S.; Di Lazzaro, G.; Scalise, S.; Pisani, A.; Saggio, G. Voice in Parkinson's Disease: A Machine Learning Study. Front. Neurol. 2022, 13, 831428.

Suppa, A.; Costantini, G.; Asci, F.; Di Leo, P.; Al-Wardat, M.S.; Di Lazzaro, G.; Scalise, S.; Pisani, A.; Saggio, G. Voice in Parkinson's Disease: A Machine Learning Study. Front. Neurol. 2022, 13, 831428.

T. Uchiyama, N. Sogi, S. Iizuka, K. Niinuma, and K. Fukui, "Adaptive occlusion sensitivity analysis for visually explaining video recognition networks," Aug. 17, 2023, arXiv: arXiv:2207.12859. doi:

10.48550/arXiv.2207.12859.

Taguchi, T., Tachikawa, H., Nemoto, K., Suzuki, M., Nagano, T., Tachibana, R., Nishimura, M., & Arai, T. (2018). Major depressive disorder discrimination using vocal acoustic features. Journal of Affective Disorders, 225(January 2017), 214–220. https://doi.org/10.1016/j.jad.2017.08.038

Takano, T., Mizuguchi, D., Omiya, Y., Higuchi, M., Nakamura, M., Shinohara, S., Mitsuyoshi, S., Saito, T., Yoshino, A., Toda, H., & Tokuno, S. (2023). Estimating Depressive Symptom Class from Voice. International Journal of Environmental Research and Public Health, 20(5). https://doi.org/10.3390/ijerph20053965

Tao, J., Tan, T. (2005). Affective Computing: A Review. In: Tao, J., Tan, T., Picard, R.W. Affective Computing and Intelligent Interaction. ACII 2005. Lecture Notes in Computer Science, vol 3784. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11573548_125

Tobiano, G., Chaboyer, W., Teasdale, T., Raleigh, R., & Manias, E. (2019). Patient engagement in admission and discharge medication communication: A systematic mixed studies review. International Journal of Nursing Studies, 95, 87–102. https://doi.org/10.1016/j.ijnurstu.2019.04.009

Tonn, P., Seule, L., Degani, Y., Herzinger, S., Klein, A., & Schulze, N. (2022). Digital Content-Free Speech Analysis Tool to Measure Affective Distress in Mental Health: Evaluation Study. JMIR Formative Research, 6(8), 1–16. https://doi.org/10.2196/37061

Tracey, B.; Volfson, D.; Glass, J.; Haulcy, R.; Kostrzebski, M.; Adams, J.; Kangarloo, T.; Brodtmann, A.; Dorsey, E.R.; Vogel, A. Towards interpretable speech biomarkers: Exploring MFCCs. Sci. Rep. 2023, 13, 22787. https://doi.org/10.1038/s41598-023-49352-2.

Tsanas, A.; Little, M.A.; McSharry, P.E.; Spielman, J.; Ramig, L.O. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Trans Biomed Eng. 2012, 59, 1264–1271.

U. Bilotti, C. Bisogni, M. De Marsico, and S. Tramonte, "Multimodal Emotion Recognition via Convolutional Neural Networks: Comparison of different strategies on two multimodal datasets," Engineering Applications of Artificial Intelligence, vol. 130, p. 107708, Apr. 2024, doi: 10.1016/j.engappai.2023.107708.

Vakili, M.; Ghamsari, M.; Rezaei, M. Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification. arXiv 2020, http://arxiv.org/abs/2001.09636

Valstar, M., Schuller, B. W., Krajewski, J., Cowie, R., & Pantic, M. (2014). AVEC 2014: the 4th international audio/visual emotion challenge and workshop. Proceedings of the 22nd ACM

International Conference on Multimedia, 1243–1244. https://doi.org/10.1145/2647868.2647869

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., & Pantic, M. (2013). Avec 2013: the continuous audio/visual emotion and depression recognition challenge. Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, 3–10. https://doi.org/10.1145/2512530.2512533

Vellata, C. et al. (2021) 'Effectiveness of telerehabilitation on motor impairments, non-motor symptoms and compliance in patients with Parkinson's disease: A systematic review', Frontiers in neurology, 12, p. 627999. Available at: https://doi.org/10.3389/fneur.2021.627999.

Verbaan, D. et al. (2007) 'Cognitive impairment in Parkinson's disease', Journal of neurology, neurosurgery, and psychiatry, 78(11), pp. 1182–1187. Available at: https://doi.org/10.1136/jnnp.2006.112367.

Veres, G. (2025) gveres/donateacry-corpus. [online] GitHub, Inc.

Violos, J., Diamanti, K.-C., Kompatsiaris, I. & Papadopoulos, S. (2025). Frugal machine learning for energy-efficient and resource-aware artificial intelligence. arXiv preprint.

Wagner, L., Banchik, M., Tsang, T., Okada, N.J., Altshuler, R., McDonald, N., Bookheimer, S.Y., Jeste, S.S., Green, S., & Dapretto, M. (2025). Atypical early neural responses to native and non-native language in infants at high likelihood for developing autism. Molecular Autism 16, 6.

Wallace, E.R. et al. (2022) 'Meta-analysis of cognition in Parkinson's Disease mild cognitive impairment and dementia progression', Neuropsychology review, 32(1), pp. 149–160. Available at: https://doi.org/10.1007/s11065-021-09502-7.

Wang, T.V. & Song, P.C. (2022). Neurological voice disorders: A review. International Journal of Head and Neck Surgery 13, 32–40.

Wang, Y., Boumadane, A., & Heba, A. (2021). A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. arXiv preprint arXiv:2111.02735.

Wasserzug, Y., Degani, Y., Bar-Shaked, M., Binyamin, M., Klein, A., Hershko, S., & Levkovitch, Y. (2023). Development and preliminary support for a machine learning-based vocal predictive model for major depressive disorder. Journal of Affective Disorders, 325(April 2022), 627–632. https://doi.org/10.1016/j.jad.2022.12.117 SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Watson, G.S. and Leverenz, J.B. (2010) 'Profile of cognitive impairment in Parkinson's disease', Brain pathology (Zurich, Switzerland), 20(3), pp. 640–645. Available at: https://doi.org/10.1111/j.1750-

3639.2010.00373.x.

Weil, R.S.; Morris, H.R. REM sleep behaviour disorder: An early window for prevention in neurodegeneration? Brain 2019, 142, 498–501.

Wen, Y., Innuganti, A., Ramos, A.B., Guo, H., & Yan, Q. (2025). SoK: How robust is audio watermarking in generative AI models? arXiv preprint.

Wilson-Menzfeld, G., Erfani, G., Young-Murphy, L., Charlton, W., De Luca, H., Brittain, K., & Steven, A. (2025). Identifying and understanding digital exclusion: a mixedmethods study. Behaviour & Information Technology, 44(8), 1649-1666.

Wind, A., van der Linden, C., Hartman, E., Siesling, S., & van Harten, W. (2022). Patient involvement in clinical pathway development, implementation and evaluation – A scoping review of international literature. Patient Education and Counselling, 105(6), 1441–1448. https://doi.org/10.1016/j.pec.2021.10.007

Wong, G., Greenhalgh, T., Westhorp, G., Buckingham, J., & Pawson, R. (2013). RAMESES publication standards: meta-narrative reviews. Journal of Advanced Nursing, 69(5), 987–1004. https://doi.org/10.1111/jan.12092

Woodall, A., Morgan, C., Sloan, C., & Howard, L. (2010). Barriers to participation in mental health research: Are there specific gender, ethnicity and age-related barriers? BMC Psychiatry, 10(1), 103. https://doi.org/10.1186/1471-244X-10-103

Wright, H.; Postema, M.; Aharonson, V. Towards a voice-based severity scale for Parkinson's disease monitoring. Curr. Dir. Biomed. Eng. 2024, 10, 2024–2168. https://hal.science/hal-04737545v2

Wroge, T.J.; Özkanca, Y.; Demiroglu, C.; Si, D.; Atkins, D.C.; Ghomi, R.H. Parkinson's disease diagnosis using machine learning and voice. In Proceedings of the 2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), Philadelphia, PA, USA, 1 December 2018; IEEE: New York, NY, USA, 2018; pp. 1–7.

Wykes, T., Csipke, E., Williams, P., Koeser, L., Nash, S., Rose, D., ... & McCrone, P. (2017). Improving patient experiences of mental health inpatient care: A randomised controlled trial. Psychological Medicine, 47(4), 681–691. https://doi.org/10.1017/S003329171700188X SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Xu, L., Yildiz, A.M., Tuncer, I., Ozyurt, F., Dogan, S., & Tuncer, T. (2025). Detection of community emotions through sound: An investigation using the FF-Orbital chaos-based feature extraction model. Ain Shams Engineering Journal 16, 103248.

Yang, K.; Wu, Z.; Long, J.; Li, W.; Wang, X.; Hu, N.; Zhao, X.; Sun, T. White matter changes in

Parkinson's disease. NPJ Park. Dis. 2023, 9, 150.

Zaman, M.S., Ghahari, S. and McColl, M.A. (2021) 'Barriers to accessing healthcare services for people with Parkinson's disease: A scoping review', Journal of Parkinson's disease, 11(4), pp. 1537–1553. Available at: https://doi.org/10.3233/JPD-212735. SER is therefore presented as a promising assessment direction rather than a clinically validated tool.

Zayed, Y., Hasasneh, A., & Tadj, C. (2023). Infant cry signal diagnostic system using deep learning and fused features. Diagnostics 13, 2107.

Zewoudie, A.W.; Luque, J.; Hernando, J. The use of long-term features for GMM- and i-vector-based speaker diarization systems. EURASIP J Audio Speech Music. Process. 2018, 2018.

Zhang, E.Q. (2025). The influence of prenatal auditory input on newborn vocalisations. SSRN Blog.

Zhang, L., Duvvuri, R., Chandra, K. K. L., Nguyen, T., & Ghomi, R. H. (2020). Automated voice biomarkers for depression symptoms using an online cross-sectional data collection initiative. Depression and Anxiety, 37(7), 657–669. https://doi.org/10.1002/da.23020

Zhao, X., Sun, H., Lin, B., Zhao, H., Niu, Y., Zhong, X., Wang, Y., Zhao, Y., Meng, F., Ding, J., Zhang, X., Dong, L., & Liang, S. (2022). Markov transition fields and deep learning-based event-classification and vibration-frequency measurement for φ-OTDR. IEEE Sensors Journal 22, 3348–3357.

Ziebland, S., Hyde, E., & Powell, J. (2021). Power, paradox and pessimism: on the unintended consequences of digital health technologies in primary care. Social Science & Medicine, 289, 114419.

Plagiarism Declaration

I, Andrea Buccoliero, declare that the work presented in this doctoral thesis is the result of my own independent research, carried out during my enrolment in the PhD Programme in Human Sciences (38th Cycle) at the University of Verona, and the period in the hosting institution (GPI S.P.A.).

This thesis has not been submitted for the award of any other academic degree or qualification, either at this institution or at any other institution where contributions of others are involved, whether in terms of collaborative research, data collection, or joint publications. Full acknowledgement is given in the appropriate sections of the thesis.

I confirm that all sources used have been properly cited and referenced, and that the thesis complies with the ethical and academic integrity standards required by the University of Verona.