HEAR4Health: a blueprint for making computer audition a staple of modern healthcare

Andreas Triantafyllopoulos, Alexander Kathan, Alice Baird, Lukas Christ, Alexander Gebhard, Maurice Gerczuk, Vincent Karas, Tobias Hübner, Xin Jing, Shuo Liu, Adria Mallol-Ragolta, Manuel Milling, Sandra Ottl, Anastasia Semertzidou, Srividya Tirunellai Rajamani, Tianhao Yan, Zijiang Yang, Judith Dineley, Shahin Amiriparian, Katrin D. Bartl-Pokorny, Anton Batliner, Florian B. Pokorny, Björn W. Schuller

Angaben zur Veröffentlichung / Publication details:

Triantafyllopoulos, Andreas, Alexander Kathan, Alice Baird, Lukas Christ, Alexander Gebhard, Maurice Gerczuk, Vincent Karas, et al. 2023. "HEAR4Health: a blueprint for making computer audition a staple of modern healthcare." *Frontiers in Digital Health* 5: 1196079. https://doi.org/10.3389/fdgth.2023.1196079.

frontiers | Frontiers in Digital Health

TYPE Review
PUBLISHED 12 September 2023
DOI 10.3389/fdgth.2023.1196079

OPEN ACCESS

EDITED BY
Dian Tjondronegoro,
Griffith University, Australia

REVIEWED BY
Sidra Abbas,
COMSATS University Islamabad - Sahiwal campus, Pakistan
Heysem Kaya,
Utrecht University, Netherlands
Elena Lyakso,
Saint Petersburg State University, Russia

*CORRESPONDENCE
Andreas Triantafyllopoulos
✉ andreas.triantafyllopoulos@uni-a.de

RECEIVED 29 March 2023
ACCEPTED 01 September 2023
PUBLISHED 12 September 2023

CITATION
Triantafyllopoulos A, Kathan A, Baird A, Christ L, Gebhard A, Gerczuk M, Karas V, Hübner T, Jing X, Liu S, Mallol-Ragolta A, Milling M, Ottl S, Semertzidou A, Rajamani ST, Yan T, Yang Z, Dineley J, Amiriparian S, Bartl-Pokorny KD, Batliner A, Pokorny FB and Schuller BW (2023) HEAR4Health: a blueprint for making computer audition a staple of modern healthcare. Front. Digit. Health 5:1196079. doi: 10.3389/fdgth.2023.1196079

HEAR4Health: a blueprint for making computer audition a staple of modern healthcare

Andreas Triantafyllopoulos[1]*, Alexander Kathan[1], Alice Baird[1], Lukas Christ[1], Alexander Gebhard[1], Maurice Gerczuk[1], Vincent Karas[1], Tobias Hübner[1], Xin Jing[1], Shuo Liu[1], Adria Mallol-Ragolta[1,3], Manuel Milling[1], Sandra Ottl[1], Anastasia Semertzidou[1], Srividya Tirunellai Rajamani[1], Tianhao Yan[1], Zijiang Yang[1], Judith Dineley[1], Shahin Amiriparian[1], Katrin D. Bartl-Pokorny[1,2], Anton Batliner[1], Florian B. Pokorny[1,2,3] and Björn W. Schuller[1,3,4]

[1]EIHW – Chair of Embedded Intelligence for Healthcare and Wellbeing, University of Augsburg, Augsburg, Germany, [2]Division of Phoniatrics, Medical University of Graz, Graz, Austria, [3]Centre for Interdisciplinary Health Research, University of Augsburg, Augsburg, Germany, [4]GLAM – Group on Language, Audio, & Music, Imperial College London, London, United Kingdom

Recent years have seen a rapid increase in digital medicine research in an attempt to transform traditional healthcare systems to their modern, intelligent, and versatile equivalents that are adequately equipped to tackle contemporary challenges. This has led to a wave of applications that utilise AI technologies; first and foremost in the fields of medical imaging, but also in the use of wearables and other intelligent sensors. In comparison, computer audition can be seen to be lagging behind, at least in terms of commercial interest. Yet, audition has long been a staple assistant for medical practitioners, with the stethoscope being the quintessential sign of doctors around the world. Transforming this traditional technology with the use of AI entails a set of unique challenges. We categorise the advances needed in four key pillars: Hear, corresponding to the cornerstone technologies needed to analyse auditory signals in real-life conditions; Earlier, for the advances needed in computational and data efficiency; Attentively, for accounting to individual differences and handling the longitudinal nature of medical data; and, finally, Responsibly, for ensuring compliance to the ethical standards accorded to the field of medicine. Thus, we provide an overview and perspective of HEAR4Health: the sketch of a modern, ubiquitous sensing system that can bring computer audition on par with other AI technologies in the strive for improved healthcare systems.

KEYWORDS

computer audition, digital health, digital medicine, speech and language disorders, auscultation

1. Introduction

Following the rapid advancements in artificial intelligence (AI), and in particular those related to deep learning (DL) (1), digital health applications making use of those technologies are accordingly on the rise. Most of them are focused on diagnosis: from computer vision techniques applied to digital imaging (2) to wearable devices monitoring a variety of signals (3, 4), AI tools are being increasingly used to provide medical practitioners with a more comprehensive view of their patients—a trend which has been

accelerating in the aftermath of the COVID-19 pandemic (5). Computer audition complements this assortment of tools by providing access to the audio generated by a patient's body. Most often, this corresponds to speech produced by the patients—sometimes natural, mostly prompted (6–8). However, there exists a plethora of auditory signals emanating from the human body, all of which are potential carriers of information relating to disease.

These biosignals can be analysed either through specialised instruments or, more interestingly, through the use of off-the-shelf microphones embedded in everyday devices, such as smartphones, which are already being widely used by healthcare professionals in their day-to-day jobs (9). As such, they are poised to be an indispensable tool to assist doctors in making better decisions and acquiring a more holistic understanding of their patients. While AI monitoring systems could, theoretically, be deployed as standalone applications and make decisions without supervision, we envision the components of our system to assist doctors in their decision making processes, rather than substitute them.

Acquiring auditory biosignals is the first, crucial step in a computer audition pipeline. Oftentimes, this must be done in noisy environments where audio engineers have little to no control, e.g., in a busy hospital room or the patient's home. This results in noisy, uncurated signals which must be pre-processed in order to become usable, a process which is extremely laborious if done manually. Automating this process becomes the domain of the first of four outlined pillars, **(I) Hear**, which is responsible for denoising, segmenting, and altogether preparing the data for further processing by the downstream algorithms.

Those algorithms typically comprise learnable components, i.e., functions whose parameters are going to be learnt from the consumed data; in the current generation of computer audition systems, the backbone of those algorithms consists of DL models. These models, in turn, are typically very data "hungry," and require an enormous amount of computational resources and experimentation to train successfully. However, in the case of healthcare applications, such data might not exist, either due to privacy regulations which prohibit their open circulation, or, as in the case of rare or novel diseases, simply because this data does not exist. Yet doctors, and subsequently the tools they use, are commonly required to operate in such low-data regimes. Therefore, it is imperative to make these algorithms operational **(II) Earlier** than what is currently possible; this can be done, for example, by transferring knowledge from domains where data is widely available to data-sparse healthcare applications.

The first two pillars are of a more "engineering" nature; the third one requires more theoretical advances. Statistical learning theory, which forms the foundation of DL, is based on the core assumption that data are *independent and identically distributed* (10). In the healthcare domain, this translates to the hypothesis that the population of training patients is representative of the entire population—an assumption that often does not hold in practice. Instead, patients come from different backgrounds and are typically organised in sub-populations. Oftentimes, the level of analysis reaches all the way down to the individual; in this case, every patient is considered "unique." Furthermore, the larger upside of using AI in medicine lies in providing more fine-grained information in the form of longitudinal observations. Handling the need for individualised analysis with multiple observations over time requires algorithms to operate **(III) Attentively** to individual—often changing—needs.

The last pillar corresponds to the translation of the mandate enshrined in the Hippocratic oath to computer audition, and more generally AI: any developed technologies must be developed and be able to operate **(IV) Responsibly**. The responsibility lies with the developers and users of the technology and is targeted towards the patients who become its objects. This informs a set of guidelines and their accompanying technological innovations on how data needs to be sourced, how algorithms must meet certain fairness requirements, and, ultimately, on "doing good" for mankind.

Finally, we would be amiss not to mention the potential applications that can benefit from the introduction of computer audition in healthcare. This becomes the central component which permeates all aspects of the four pillars: they exist insofar as they serve the overarching goal of providing medical practitioners with novel tools that can help them understand, analyse, diagnose, and monitor their patients' **Health**.

An overview of the four pillars, as well as their interconnections are shown in **Figure 1**. In the following sections, we begin with an overview of the particular diseases in which we expect computer audition to make a decisive contribution, which lays the setting for the four pillars. We then proceed to analyse each of our four pillars in more detail and end with a discussion of how all four of them can be integrated in a practical architecture. Thus, we present **HEAR4Health**: an overview of recent advances and a blueprint for what needs to be done for audition to assume its rightful place in the toolkit of AI technologies that are rapidly revolutionising healthcare systems around the world.

Our work aims to go beyond existing surveys, which only concern themselves with the different technical aspects of computer audition, and link those aspects to the pragmatic requirements of the digital health setting. Therefore, instead of diving deep into technical details, we provide a broad, holistic coverage of the different components that are needed. Our work represents a roadmap and a blueprint for healthcare researchers and practitioners aiming to utilise computer audition to tackle a diverse set of challenges.

2. Healthcare applications

Naturally, any advances in computer audition targeted towards healthcare applications are inextricably tied to the specific medical conditions that lend themselves to modelling via audio; the necessary pre-requisite is that these conditions manifest themselves, at least to some extent, in auditory biomarkers emanating from the patients' bodies. Historically, a significantly higher emphasis has been placed on vocalisations compared to other body acoustics such as heart sounds (6). Accordingly, this choice has shaped most of the existing approaches and, thus, also becomes the central point of our review. **Table 1** shows the main

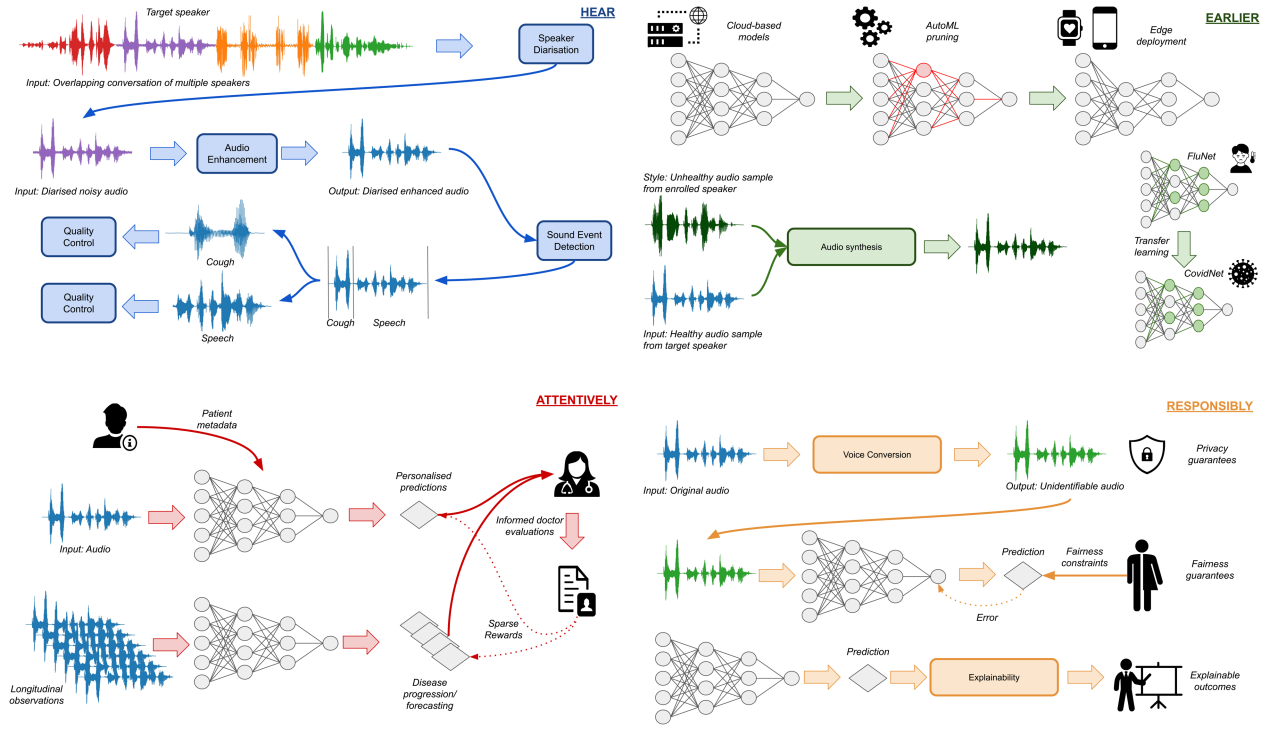

FIGURE 1
Overview of the four pillars for computer audition in healthcare. HEAR is tasked with capturing, identifying, and enhancing the target signal. EARLIER accounts for the need to deploy models at the edge to facilitate ubiquitous monitoring, as well as adapt to novel diseases using transfer learning and synthetic data generation. ATTENTIVELY incorporates personal information in the decision-making process, either using patient metadata or longitudinal observations, while assimilating doctor decisions in a human-in-the-loop framework. RESPONSIBLY ensures that developed models adhere to ethical requirements by safeguarding privacy, promoting fairness, and adding a layer of explainability.

ICD-11[1] categories on which previous research has focused, as well as the specific diseases that previous studies have focused on and the auditory signals and symptoms used for acoustic disease monitoring and characterisation. In the following sections, we proceed to analyse each of those categories, presenting prior computer audition works that have focused on specific diseases, and discussing the impact that our HEAR4Health framework can have on them.

2.1. Infectious or parasitic diseases

This broad category covers several communicable diseases, from bacterial, gastrointestinal infections, to sexually transmitted diseases and viral infections, the majority of which do not manifest in auditory biomarkers; the ones that do, however, number several auditory symptoms such as (persistent) coughing or having a sore throat. The ones predominantly appearing in computer audition literature are: (respiratory) *tuberculosis* (1B10) (11–14); *pertussis* (1C12) (15, 16); and *influenza*(1E) (17). Existing works have predominantly focused on detecting and analysing coughs; in particular, the onset of DL and the increase in available data have unveiled the potential of detecting coughs and subsequently categorising them as pathological or not.

2.2. Mental, behavioural, or neurodevelopmental disorders

Disorders belonging to this category are described by the WHO as "syndromes characterised by clinically significant disturbance in an individual's cognition, emotional regulation, or behaviour that reflects a dysfunction in the psychological, biological, or developmental processes that underlie mental and behavioural functioning." The wide variety of symptoms of these diseases, often manifesting as speech and language pathologies, along with their widespread prevalence, have made them prime targets for the computer audition community (7, 18–20). Typical disorders are: (developmental) *aphasia* (6A01.20) (21, 22); *schizophrenia* (6A20) (23–25); *autism* (6A02) (26–30); *mood disorders* (6A60; 6A70), of which *depression* is the most commonly researched (31–33); and *anxiety or fear-related disorders* (6B) (34, 35). For example, aphasia has been linked to mispronunciation errors and increased effort (21, 22); schizophrenia manifests in slower response times in conversations (24); and blunted affect (23–25); depression results in a flatter tone (lower mean F0 values and

[1]https://icd.who.int/

TABLE 1 Overview of diseases per ICD-11 category that are frequently investigated in the computer audition literature, as well as a list of auditory symptoms and signals used for their monitoring.

Disease group	Diseases	Symptoms	Relevant biosignals
Infectious or parasitic diseases	Tuberculosis Pertussis Influenza	Coughing Sore throat	Coughing Speech
Mental, behavioural, or neurological disorders	Aphasia Schizophrenia Autism Depression	Flatter tone Pauses Strained articulation	Speech
Sleep-wake disorders	Apnoeas	Snoring	Breathing Snoring
Diseases of the nervous system	Parkinson's Alzheimer's Multiple sclerosis ALS Cerebral palsy	Articulation problems Phonation problems	Speech Sustained vowels
Diseases of the circulatory system	Arrythmias	Irregular heartbeat	Heartbeat
Diseases of the respiratory system	Asthma Bronchitis COPD COVID-19 Pneumonia	Coughing Articulation problems Phonation problems	Breathing Coughing Speech Sustained vowels
Developmental anomalies	Angelman syndrome Fragile X syndrome Rett syndrome	Abnormal sounds Abnormal speech	Baby sounds Speech

range) with more pauses (32), and an increase in jitter and shimmer, indicative of more strained articulation.

2.3. Sleep-wake disorders

Research in sleep-wake disorders has been typically targeted to *breathing-disorders*—mainly apnoeas (36, 37), while some research has been focused on the detection of the resulting *sleepiness* (38). Apnoeas, on the one hand, mostly manifest as very loud snoring, which is caused by a prolonged obstruction of the airways and subsequent "explosive" inspirations. These signals can be automatically detected and analysed using auditory machine learning (ML) systems (39). Daytime sleepiness, on the other hand, has been mostly studied as a speech and language disorder; it manifests in lower speaking rates and irregular phonation (40).

2.4. Diseases of the nervous system

This family of diseases has adverse effects on memory, motor control, and cognitive performance. Its most widely studied sub-categories from a speech pathology perspective are Parkinson's (8A00.0) (41, 42); Alzheimer's (8A20) (43–45); multiple sclerosis (8A40) (46); amyotrophic lateral sclerosis (8B60) (47); and cerebral palsy (8D) (48). These manifest primarily in the speech signal, with dysarthria and dysphonia being the most common symptoms. For example, studies find that Parkinson's shows up as increased roughness, breathiness and dysphonia, and higher F0 values (41, 42), Alzheimer's results in more hesitation (43), multiple sclerosis leads to slower and more imprecise articulation, pitch and loudness instability, longer and more frequent pauses (46), and cerebral palsy shows up as dysarthria, hypernasality, and imprecise articulation of consonants (48).

2.5. Diseases of the circulatory system

Auscultation has been a mainstay of a medical examination since the invention of the stethoscope by Renë Laennec in 1816, by now a trademark of medical practitioners around the world (49). It is particularly useful when listening to the sounds of the heart or the lungs of a patient. Accordingly, its digital equivalent can be immensely useful in detecting pathologies of the circulatory system, such as arrhythmias or congenital heart diseases. Analysing those signals has become the topic of multiple PhysioNet challenges (50) was also featured in the 2018 version of the ComParE series (51), with computer audition systems being developed to detect and classify abnormal events ("murmurs") in phonocardiograms (52, 53).

2.6. Diseases of the respiratory system

These diseases can be broadly taxonomised as being related to the upper or lower respiratory tract. Prominent examples are bronchitis (CA20) (16); chronic obstructive pulmonary disease (COPD; CA22) (54, 55); asthma (CA23) (56); pneumonia (CA40) (57); and COVID-19 (RA01; designated under "codes for special purposes" due to the pandemic emergency) (58, 59). By nature of their symptomatology, these diseases are prototypical examples of ones that manifest in auditory biomarkers. Thus, different signals have been used to detect their presence, such as speech (60), breathing and coughing (61, 62), or sustained vowels (63). The most exemplary of those is COVID-19, whose devastating impact was felt around the world since its emergence in late 2019 and has led to a wave of renewed interest in computer audition for healthcare applications. In general, these diseases lead to more coughs, irregular breathing, phonation and articulation, and constrained airflow resulting in less loud and more strained vocalisations.

2.7. Developmental anomalies

Developmental disorders, such as the Angelman syndrome (LD90.0), Rett syndrome (LD90.4), and fragile X syndrome (LD55) manifest in divergent vocalisation and speech development patterns from an early age (29, 64–66). Infants with specific developmental disorders produce abnormal cooing sounds and less person-directed vocalisations, and their vocalisations are found to be of lower complexity as compared to typically developing infants. From a signal perspective, these anomalies manifest in speech, first in pre-linguistic sounds and later on in linguistic vocalisations of young children. As the

emphasis is on children and young adults, they present an additional challenge to data collection, on the one hand due to ethical and privacy reasons, and on the other due to a potentially reduced compliance of children with recording requirements.

3. Hear

A cornerstone of computer audition applications for healthcare is the ability to *Hear*: that is, the set of steps required to capture and pre-process audio waves and transform them into a clear, useful, and high-quality signal. This is all the more true in the healthcare domain, where recordings are often made in hospital rooms bustling with activity or conducted at home by the non-expert users themselves. Therefore, the first fundamental step in an application is to extract only the necessary components of a waveform.

In general, this falls under a category of problems commonly referred to as *source separation and diarisation* (67, 68): the separation part corresponds to the extraction of a signal coming from a particular source amongst a mixture of potentially overlapping sources, whereas diarisation corresponds to the identification of temporal start and end times of components assigned to specific subjects. In healthcare applications, these target components are the relevant sounds; this can include vocalisations (both verbal and non-verbal) but also other bodily sounds that can be captured by specialised auditory sensors attached to their body, or general ones that are monitoring the environment. These sounds need to be separated from all other sources; these may include a medical practitioner's own body sounds (e.g., their voice in doctor-patient conversations) or background environmental noise (e.g., babble noise in a hospital). Accordingly, successful preparation entails a) the ability to recognise which sounds belong to the target subject, b) the ability to detect their precise start and end times, and c) the ability to remove all other signals that co-occur during that time from the waveform.

Traditionally, these steps are tackled by specialised pipelines, which include learnable components that are optimised in supervised fashion (68). For example, the ability to recognise which sounds belong to the target subject is generally referred to as *speaker identification* (69). While this term is usually reserved for applications where speech is the sound of interest, it can also be generalised to other bodily sounds (70). Similarly, separation is typically done in a supervised way (68). During the training phase, clean audio signals are mixed with different noises, and a network is trained to predict the original, clean signal from the noisy mixture. As generalisability to new types of noise sources is a necessary pre-requisite, researchers often experiment with test-time adaptation methods, which adaptively configure a separation model to a particular source (71).

The crucial role of the *Hear* pillar becomes evident when considering data collection. There are three main data collection paradigms employed in healthcare applications: (a) the (semi-) structured doctor-patient interview, (b) ecological momentary assessments (EMAs) based on prompts (72), and, (c) passive, continual monitoring (73). All of them require very robust patient identification and diarisation capabilities. However, each comes with its own set of unique challenges that can be tackled by the *Hear* pillar. Structured interviews are often conducted in relatively quiet environments (e.g. a doctor's office or laboratory); the challenge mainly relies in the use of far-field microphones that make the processing more complicated (e.g. resulting in reverberation) (74). The need for a robust *Hear* pillar is punctuated by the fact that response rates and speaking times during interviews are often very informative features for these types of diseases; their accurate estimation is only possible following a reliable diarisation step.

EMAs further complicate processing as they may take place in different environments, not necessarily quiet ones, as the patient can choose to conduct them in any environment of their choosing. Thus, denoising becomes a crucial factor for removing the unwanted interference from background noise. Passive monitoring represents the most challenging form of data collection. The auditory signals are potentially embedded in several, high-varying sources; detection then becomes the first crucial step. Voice activity detection is more mature than the detection of other types of bodily acoustics; still, even that suffers from robustness issues and is often a crucial bottleneck for successful applications (75). This is followed by a source separation step which attempts to extract the useful signals from any other sources, a feat which becomes more challenging for non-speech signals such as coughing, as this requires general source separation. For such symptoms, a major contribution of the *Hear* pillar would be to improve the detection of coughs in naturalistic environments; this would pave the way for continuous monitoring using smart wearables or smartphones to monitor (prospective) patients over time and detect a change in their frequency of their coughing over time.

Moreover, the audio processing can be formulated as a single, unified task of target audio extraction. The gold standard for digital health applications is not defined by human listening studies as in traditional source separation, but rather from the performance of downstream processing modules, with the goal being to increase their performance and robustness to noise. Overall, the aim of pre-processing is to reduce the uncertainty in real-life recordings by adapting to different environmental situations. Hence, it helps to provide a more robust interface that enables digital health applications. Finally, some techniques based on speech enhancement and source separation, such as signal-to-noise ratio (SNR) estimation, can be used to make a decision on whether a specific audio signal is suitable for further audio-based medical diagnosis, depending on the quality of the original recording and the processed audio.

4. Earlier

The major promise of digital health applications is their ubiquitous presence, allowing for a much more fine-grained monitoring of patients than was possible in the past. This requires the systems to work on mobile devices in an energy-efficient way. Additionally, these systems must be versatile, and

easy to update in the case of new diseases, such as COVID-19. This requires them to generalise well while being trained on very scarce data. However, training state-of-the-art DL models is a non-trivial process, in many cases requiring weeks or even months, and is furthermore notoriously data intensive. Moreover, the technology required, such as high-end GPUs, is often expensive and has exceptionally high energy consumption (76).

There have consequently been increasing efforts to develop AutoML approaches that optimise a large network until it is executable on a low-resource device (77, 78). Many of these approaches focus on reducing the memory footprint and the computational complexity of a network while preserving its accuracy. These techniques have shown promise across a range of different learning tasks, however, their potential has not yet been realised for audio-based digital health applications.

On the issue of data efficiency, there has been a lot of research on utilising transfer learning techniques for increasing performance and decreasing the required amount of data. This is usually done by transferring knowledge from other tasks (79, 80), or even other modalities (81, 82). However, in the case of audio in particular, an extra challenge is presented by the mismatch between the pre-training and downstream domains (83). Recently, large models pre-trained in self-supervised fashion have reached exceptional performance on a variety of different downstream tasks, including the modelling of respiratory diseases (54), while showing more desirable robustness and fairness properties (84).

The implementation details of the *Earlier* pillar largely depend on the biomarkers related to the specific medical condition of interest. For example, in terms of mental disorders, which mostly manifest as pathologies of speech and language, it is mostly tied to generalisation across different languages. On the one hand, linguistic content itself is a crucial biomarker; on the other hand, it serves to constrain the function of acoustic features; thus, there is a need to learn multi-lingual representations that translate well to low-resource languages. For diseases manifesting in sounds other than speech signals, the *Earlier* pillar would then improve the data efficiency of their categorisation. For example, contrary to speech signals, for which large, pre-trained models are readily available (85), there is a lack of similar models trained on cough data; a lack partially attributable to the dearth of available data. This can be overcome, on the one hand, through the use of semi-supervised methods that crawl data from public sources (86), and, on the other hand, by pursuing (deep) representation learning methods tailored to cough sound characteristics.

When COVID-19 took the world by storm in early 2020, it represented a new, previously unseen threat for which no data was available. However, COVID-19 is "merely" a coronavirus targeting the upper and lower respiratory tracts, thus sharing common characteristics with other diseases in the same family (87). Transferring prior knowledge from those diseases, while rapidly adapting to the individual characteristics of COVID-19, can be another crucial factor when deploying auditory screening tools in the face of a pandemic.

Nevertheless, even after using transfer learning techniques, the problem of data sparsity still remains. In the audio domain, acquisition of data representative of the variety of signals seen at population level is time-consuming, costly and inefficient. A potential remedy could be found in *generating* new data. Many state-of-the-art, high-fidelity approaches for generating audio computationally are being developed, and these could be used to facilitate targeted data generation for handling underrepresented diseases. Co-opting these approaches for the digital health domain to generate (personalised) utterances of pathological speech and use them to augment the training data holds a lot of promise for mitigating the sparsity issue.

5. Attentively

Most contemporary digital health applications focus on the identification of subject states in a static setting, where it is assumed that subjects belong to a certain category or have an attribute in a certain range. However, many conditions have symptoms that manifest gradually (88), which makes their detection and monitoring over time a key proposition for future digital health applications. Furthermore, disease emergence and progression over time can vary between individuals (89–91). For example, the age at onset and the progression rate of age-related cognitive decline varies between individuals (89), while there is substantial heterogeneity in the manifestation and development of (chronic) cough across different patients (92). Focusing on these aspects of digital health by adapting to changes in distributions and developing personalised approaches can drastically improve performance.

Recent deep neural network (DNN)-based methods for personalised ML (30) and speaker adaptation (93) already pave the way for creating individualised models for different patients. However, these methods are still in their nascent stage in healthcare (94). Personalised ML is a paradigm which attempts to jointly learn from data coming from several individuals while accounting for differences between them. Advancing this paradigm for speech in digital health by utilising longitudinal data from several patients for learning to track changes in vocal and overall behaviour over time is a necessary precondition for the digital health systems of the future. This means that time-dependent, individualised distributions are taken into account for each patient, by that requiring the development of novel techniques better suited to the nature of this problem; in particular, developing versatile DL architectures consisting of global components that jointly learn from all subjects, and specialised ones which adapt to particular patients (95, 96). This novel framing will also enable faster adaptation to new patients by introducing and adapting new models for those patients alone.

On the other hand, speaker adaptation corresponds to disentangling speaker effects from biomarkers related to specific speaker states. This will be achieved by breaking down the input audio signal to a set of independent factors, enabling factors unrelated to the task at hand to be disregarded, such as speaker characteristics or speaker traits. The novel framework of causal representation learning (97), where deep neural networks are trained to disentangle independent factors, has yet to be utilised

in the healthcare domain. Accordingly, DL architectures must be developed that can utilise this implicit factorisation to differentiate between speaker-specific factors and disease biomarkers.

Finally, a major proposition of computer audition is to supplement clinical evaluations and doctor visits, which are resource-intensive procedures, with cost-efficient AI-driven measurements, remotely collected in advance of the appointment with finer temporal resolution. In addition to potential research applications, in clinical practice, this detailed, objective record will provide insights to clinicians and enable more timely diagnostic investigations; for example, by using change point detection to identify changes in the patient's state. The feedback from expert examinations, which is collected in more infrequent intervals, can then be incorporated using reinforcement learning. Reinforcement learning remains underutilised in the audio domain, largely because of the lack of an interactive environment where sparse rewards are available. However, digital health applications are ripe with sparse signals from medical practitioners in conjunction with asynchronous audio recordings that can be used to actively learn from sequences of observations in a constantly changing environment. Some of the audio recordings are collected in regular intervals, e.g. using smartphone apps or phone prompts, and are only supplemented by self-report measures when appropriate (96). However, such measures can only serve as a proxy to the target at hand; clinical evaluations, sometimes including specialised tests, are the gold standard in health state assessment. These more costly interventions are carried out infrequently compared to the remotely collected data, based on the decisions of medical practitioners. The asynchronous relationship between data recordings and targets present a fundamental problem for digital health applications. Tackling this challenge will become possible by developing a reinforcement learning framework for audio, where patient recordings will constitute the observations and clinical evaluations the "reward" from which the model will learn.

Adapting to individual characteristics is also of paramount importance. The *Attentively* pillar can become a cornerstone of future applications for monitoring mental health. Applications are already re-orienting towards longitudinal monitoring; this serves to provide more insight to a patient's mental state over time, and lends itself well to personalised modelling. This will additionally help elucidate differences within this family of diseases; as discussed above, symptoms are often similar across different mood disorders, making their differentiation difficult. This obstacle can be overcome by contextualising a model to individual characteristics, such as patient histories or demographics, resulting in a hybrid AI system, comprising both data-driven and knowledge-based components. To the best of our knowledge this has not been utilised in computational research for digital health—presenting a prime opportunity for the *Attentively* pillar.

Naturally, the requirement for personal data raises serious technical and ethical challenges. Firstly, this information might not be available to the same extent for each patient. Furthermore, there is an explicit trade-off between personalisation and privacy; the more individual-level that is needed by a system, the more privacy-infringing it becomes. It is therefore necessary that any personalisation methods are optional; their use should be turned on or off depending on whether the data is available and the patient agrees to its use. Most methods already support this level of controllability as they can be trained in a multi-condition scenario with different combinations of available/missing data. By selectively dropping out personal information during training, the system can learn to generalise in situations where this information is not available during deployment.

6. Responsibly

The development of responsible digital health technology is a key pillar of future healthcare applications. This ensures trustworthiness and encourages the adherence of users to monitoring protocols. Consequently, addressing crucial factors and technology-related consequences in automated disease detection concerning human subjects in a real-world context is of paramount importance (98).

This pillar intersects with all previous ones and informs their design, adhering to an 'ethical-by-design' principle which is fundamental for healthcare applications. Naturally, a first requirement that applies to all pillars is one of evaluation: all components of a healthcare application need to be comprehensively evaluated with respect to all sub-populations and sensitive attributes. This holds true for all components of a computer audition system: from extracting the target audio signal (Hear) to generating efficient representations (Earlier) and adapting to individual characteristics (Attentively), any developed methods should perform equally for different sub-populations. The evaluation could be complemented by explainability methods, which explicitly search for biases in model decisions (99).

Aside from comprehensively evaluating all methods with respect to fairness, explicit steps must be taken to improve on those (100). To this end, adversarial (101) and constraint-based methods (102) have been proposed to learn fair representations. In adversarial debiasing, the main predictive network learns to perform its task while an adversary pushes it toward representations which obfuscate the protected characteristics. Constraint-based methods instead solve the main prediction task subject to fairness constraints (such as equality of opportunity); these methods rely on convex relaxation or game-theoretic optimisation to efficiently optimise the constrained loss function.

The second requirement placed on the three other pillars is privacy. For example, the *Hear* pillar could be co-opted to remove private information (e.g., via using keyword spotting to remove sensitive linguistic information). The *Earlier* pillar would then take the extracted signal and remove any paralinguistic information unrelated to the task; this could be achieved by targeted voice conversion that preserves any required signal characteristics but changes the patient's voice to be unrecognisable (103).

A popular method to protect the privacy of an individual when analysing and releasing data is differential privacy (DP) (104). DP

tries to prevent "attackers" from being able to determine whether a certain individual is included in the dataset or not, i.e., the contribution of an individual in the dataset to the data or query output is obscured (105). This is achieved by adding controlled noise or randomness to the data or query results. In this connection, a parameter ϵ is applied to determine the strength of protection provided by a differential privacy mechanism, at which smaller values of ϵ may lead to better privacy but less data utility. Therefore, the non-trivial choice of ϵ depends on the specific privacy requirements, risk tolerance, and the data sensitivity in order to best deal with this tradeoff between privacy and utility. However, the personal information embedded in audio signals is not needed for a successful prediction and can be removed prior to storing the data thus safeguarding the privacy of individual patients irrespective of failsafe mechanisms that protect the collected datasets, which may prove insufficient against future, more competent attackers.

Satisfying this requirement, however, is particularly challenging for the *Attentively* pillar, as there is a natural privacy-personalisation trade-off: the more private information is removed, the less context remains to be utilised for the target patient. The main solution to this obstacle is the use of federated learning (106): to ensure that sensitive information cannot be derived from central models, differential privacy methods have been proposed, such as differentially private stochastic gradient descent (107) and a private aggregation of teacher ensembles (108). These methods would update the global model backbone discussed in Section V, which is shared among all "clients," while any personalised components would remain local—and thus under the protection of safety mechanisms implemented by the client institutions.

Finally, researchers need to focus on intersections between the investigated technology, the healthcare professional, and the patient. Understanding how and why a particular decision was made is critical for all stakeholders in the medical ecosystem. First and foremost, patients are entitled to an explanation for a particular diagnosis or proposed treatment plan (109, 110). These decisions will ultimately be made by doctors who utilise AI models as tools, and thus, they need to understand the outputs and workings of those models themselves. Finally, model developers can benefit from a better understanding of how their model works in order to improve it in future iterations. Explainable AI (XAI) provides a clear understanding of how an algorithm works and why it makes specific decisions. This information helps medical professionals trust and interpret the AI's outputs, and it also makes it easier for them to explain the AI's decisions to their patients. Expectedly, the community has recently engaged in substantial research efforts to mitigate this problem, leading to a wave of novel XAI techniques (111–114).

XAI methods can be broadly categorised into two main categories: model-based (global) and instance-based (local) (115). Model-based methods, such as surrogate models or layer visualisation techniques, attempt to understand the inner mechanisms of a particular model. These methods give an understanding of how a model makes decisions over multiple instances. In contrast, instance-based methods focus on why a particular decision was made for each particular instance. These methods attempt to attribute the decision to the characteristics of that specific instance. Finally, a particularly pertinent explainability method for healthcare applications is counterfactional explainability (116), which provide a natural interface for doctors to evaluate alternative outcomes through the language of counterfactuals ("what would the decision have been if feature X had a different value?"). Ultimately, the goal of a comprehensive system should be to combine an assortment of different XAI methods and provide a well-rounded understanding of how auditory models work and why they make specific decisions.

In addition to these by now established XAI methods, recent advances in generative AI have paved the way for a more natural presentation of explanations to the end-user. For instance, the recent success in mapping different modalities to text by aligning the learnt representation spaces of large multimodal foundation models has enabled the provision of *textual explanations*, which can be seen as a form of captioning (117), with the difference that the must conform to XAI requirements (i.e., fidelity and correctness). For auditory models in particular, a more natural way to present explanations would be the *sonification* of explanatory information (118). In a nutshell, sonification entails the generation of audio that conveys information in an easily digestible way. At a basic level, this might simply correspond to identifying those constituents of an auditory signal which were most relevant for a particular decision and playing them back to the user (a method often used in biofeedback for training (119)), though with the advent of generative audio models, more advanced explanations will become possible.

Finally, like *Earlier*, low-resource languages become the subject of the *Responsibly* pillar as well; the majority of studies has been performed on English data, due to their wider availability. However, due to the widespread nature of mood disorders, it is imperative to extend the applicability of computer audition algorithms to a wider gamut of languages (and cultures). There is an equal lack of work on fairness aspects relating to cough detection and categorisation; in particular, we expect age to play a crucial work both in the frequency, and the acoustic properties of cough signals; this would fall under the auspices of the *Responsibly* pillar, which should be tasked first with understanding, and subsequently mitigating, differences in performance across different populations. Similar to explainability, these aspects of fairness could be embedded in a counterfactual framework (120) which would allow medical practitioners to examine alternative scenarios for algorithmic predictions ("what would the decision be if the patient was female instead of male?").

7. HEAR4Health: a blueprint for future auditory digital health

Early diagnosis, ideally even before symptoms become obvious to individuals in their daily lives, allows very early interventions, maximising the likelihood of successful treatments and a positive outcome, and optimising public health expenditures. While early

diagnosis will not enable a curative treatment of all diseases in all cases, it provides the greatest chance of preventing irreversible pathological changes in the organ, skeletal, or nervous system, as well as reducing chronic pain and psychological stress. In some cases, early intervention can prevent the emergence of related long-term consequences. From a public health perspective, early detection is also an effective way to minimise the spread of contagious diseases—as became evident during the COVID-19 pandemic.

Audio signals are well suited to such a non-invasive early diagnosis strategy, as they can be easily acquired anywhere and anytime using ubiquitous smart devices. A key differentiating factor of audition, as opposed to other modalities, is the nature of the signals that are used for monitoring patients. This can be audio recordings of the voice (e.g. sustained vowels, social interactions and interviews), body sounds (e.g. heartbeat, breathing, coughing, and snoring sounds), and audio recordings of an individual's acoustic environment (e.g. extracting information about frequency of communicative acts and emotional states during interactions, or noise exposure) with the aim of developing tools and methods to support the earlier diagnosis of acute and chronic diseases. The nature of those signals presents new challenges, and new opportunities, for future healthcare systems. In the present section, we attempt to sketch out a blueprint for bringing existing and upcoming advances of computer audition technologies out of the lab and into the real world of contemporary medicine.

Unifying the four pillars results in a working digital health system which we name *HEAR*. Our system can be used to supplement the decision-making of practitioners across a wide facet of diseases. In general, we anticipate two distinct functioning modes for it. On the one hand, it can be used as a general-purpose *screening* tool to monitor healthy individuals and provide early warning signs of a potential disease. This hearing with "all ears open" mode takes a holistic approach, and emphasises a wide coverage of symptoms and diseases, thus functioning as an early alarm system that triggers a follow-up investigation. Following that, it can be utilised to *monitor* the state of patients after they have been diagnosed with a disease, or for measuring the effect of an intervention. This second, more constrained setting necessitates a 'human-in-the-loop' paradigm, where the doctor isolates a narrower set of biomarkers for the system to monitor—now with more focus and prior information about the patient's state—which is then reported back for each new follow-up. Through this loop, HEAR provides vital information of a high-resolution temporal scale, thus facilitating more personalised interventions and helping strengthen the doctor-patient link.

A key enabling factor for both operating modes will be the co-opting of ubiquitous auditory sensors as medical screening devices: the most obvious candidates would be smartphones, but also other IoT devices with audio recording capabilities, such as smartwatches. These would rely on active or passing auditory monitoring to identify potential symptoms. In the case of passive monitoring, the onus would be on the *Hear* pillar to detect them: in the case of speech, utilising voice activity detection, identification, diarisation, and separation to extract the target voice, while additionally performing audio event detection to detect coughs, sneezes, snores, or other bodily sounds. Analysing the frequency of symptoms (in the case of coughs or similar sounds) could serve as the first indication of a disease. Further exploring the nature of those symptoms would require the use of other pillars, most notably the *Attentively* one to determine whether the identified sounds represent a deviation from the "norm" of a given subject; this would serve as another indication of disease. In general, both systems would collaborate with the *Responsibly* pillar to take into account subject demographics. This would help contextualised detected patterns with respect to the particular risks faced by the individual.

This early screening system would mostly serve to provide warnings which trigger a subsequent medical evaluation. During a visit to a medical professional, the subject would present an account of their symptoms, which would be complemented by intelligent analytics from the auditory monitoring system. This highlights the need for the system to be explainable (a component of the *Responsibly* pillar), as merely reporting concerning findings without additional details or explanations about the nature of the detected pathology would be of little help to the practitioner.

Following a medical examination, the nature of which could also entail the use of computer audition technologies as well, the healthcare professional would prescribe an intervention (e.g. in the form of medication or surgery) or highlight potential causes of concern. The success of this intervention or the potential risks require subsequent monitoring, entrusted (partially) to a similar auditory monitoring system. This time, however, instead of a "broad sweep" for different comorbidities, the monitoring would be targeted to a particular disease, or at least a constrained set of alternative diagnoses prescribed by the practitioner. A diagnosis serves as a "primer" for all components of the *HEAR* system to look for a particular disease: the *Hear* component would be more sensitive to the biomarkers of choice, the *Earlier* pillar would draw on existing knowledge for those biomarkers to enhance their detection, *Attentively* would track changes according to the initial states, *Responsibly* would provide the missing link to patients and practitioners by interpreting those changes and transforming them to features understandable by both professionals and laymen.

In general, this second, more targeted phase would mean the beginning of a human-machine loop, where a medical practitioner prescribes interventions whose success the auditory system helps to quantify, or identifies missing information that the system needs to gather. Each time, a new prescription signals a new configuration of the *HEAR* system: *Hear* looks for the missing information, assisted by *Earlier* and *Attentively*, and their findings are reported back with the help of *Responsibly*.

Each pillar may encompass different capabilities across the two different operating modes. While all of them need to be active during the initial screening phase, with the system hearing with "all ears open," there is an inherent trade-off in such a holistic approach: the more targets the system is looking for, the bigger the possibility of false alarms or confusions, and the greater the

complexity of the overall system, with the accompanying increases in energy consumption and computational resources. This necessitates the presence of the second, more guided phase, where the system is looking for a more constrained set of biomarkers. In either case, there are stringent requirements for reliability and explainability that can only be satisfied with the use of prior knowledge, attention to the individual, and an adherence to ethical principles. Ultimately, it is user trust that is the deciding factor behind the adoption of a transformative technology. The use of computer audition in healthcare applications is currently in its nascent stages, with a vast potential for improvement. Our blueprint, *HEAR4Health*, incorporates the necessary design principles and pragmatic considerations that need to be accounted for by the next wave of research advances to turn audition into a cornerstone of future, digitised healthcare systems.

Author contributions

BWS conceptualised the structure and content of the four pillars. Literature review was conducted for each pillar independently as follows: Hear: AT, XJ, and SL Earlier: AlB, SO, TY, TH, ZY, and SA. Attentively: AT, AK, AG, AS, STR, and SL Responsibly: AlB, LC, MG, VK, AnB, and FBP; and for the healthcare applications by: AlB, AG, AM-R, AS, SO, MM, STR, JD, and KDB-P. Following that, AT, AK, FBP, and BWS synthesised the resulting literature reviews. AT wrote the first manuscript draft. All authors contributed to the article and approved the submitted version.

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.

Publisher's note

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. (2015) 521:436–44. doi: 10.1038/nature14539

2. Esteva A, Chou K, Yeung S, Naik N, Madani A, Mottaghi A, et al. Deep learning-enabled medical computer vision. *NPJ Digit Med*. (2021) 4:1–9. doi: 10.1038/s41746-020-00376-2

3. Amft O. How wearable computing is shaping digital health. *IEEE Pervasive Comput*. (2018) 17:92–8. doi: 10.1109/MPRV.2018.011591067

4. Tu J, Torrente-Rodríguez RM, Wang M, Gao W. The era of digital health: a review of portable, wearable affinity biosensors. *Adv Funct Mater*. (2020) 30:1906713. doi: 10.1002/adfm.201906713

5. Tarhini A, Harfouche A, De Marco M. Artificial intelligence-based digital transformation for sustainable societies: the prevailing effect of COVID-19 crises. *Pac Asia J Assoc Inf Syst*. (2022) 14:1. doi: 10.17705/1pais.14201

6. Cummins N, Baird A, Schuller BW. Speech analysis for health: current state-of-the-art, the increasing impact of deep learning. *Methods*. (2018) 151:41–54. doi: 10.1016/j.ymeth.2018.07.007

7. Latif S, Qadir J, Qayyum A, Usama M, Younis S. Speech technology for healthcare: opportunities, challenges, and state of the art. *IEEE Rev Biomed Eng*. (2020) 14:342–56. doi: 10.1109/RBME.2020.3006860

8. Milling M, Pokorny FB, Bartl-Pokorny KD, Schuller BW. Is speech the new blood? Recent progress in ai-based disease detection from audio in a nutshell. *Front Digit Health*. (2022) 4:886615. doi: 10.3389/fdgth.2022.886615

9. Hitti E, Hadid D, Melki J, Kaddoura R, Alameddine M. Mobile device use among emergency department healthcare professionals: prevalence, utilization, attitudes. *Sci Rep*. (2021) 11:1–8. doi: 10.1038/s41598-021-81278-5

10. Shalev-Shwartz S, Ben-David S. *Understanding machine learning: from theory to algorithms*. Cambridge: Cambridge University Press (2014).

11. Larson S, Comina G, Gilman RH, Tracey BH, Bravard M, López JW. Validation of an automated cough detection algorithm for tracking recovery of pulmonary tuberculosis patients. *PLoS ONE*. (2012) 7:1–10. doi: 10.1371/journal.pone.0046229

12. Botha G, Theron G, Warren R, Klopper M, Dheda K, Van Helden P, et al. Detection of tuberculosis by automatic cough sound analysis. *Physiol Meas*. (2018) 39:045005. doi: 10.1088/1361-6579/aab6d0

13. Ijaz A, Nabeel M, Masood U, Mahmood T, Hashmi MS, Posokhova I, et al. Towards using cough for respiratory disease diagnosis by leveraging artificial intelligence: a survey. *Inform Med Unlocked*. (2022) 29:100832. doi: 10.1016/j.imu.2021.100832

14. Zimmer AJ, Ugarte-Gil C, Pathri R, Dewan P, Jaganath D, Cattamanchi A, et al. Making cough count in tuberculosis care. *Commun Med*. (2022) 2:1–8. doi: 10.1038/s43856-022-00149-w

15. Pramono RXA, Imtiaz SA, Rodriguez-Villegas E. A cough-based algorithm for automatic diagnosis of pertussis. *PLoS ONE*. (2016) 11:e0162128. doi: 10.1371/journal.pone.0162128

16. Imran A, Posokhova I, Qureshi HN, Masood U, Riaz MS, Ali K, et al. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. *Inform Med Unlocked*. (2020) 20:100378. doi: 10.1016/j.imu.2020.100378

17. Ward RJ, Jjunju FPM, Kabenge I, Wanyenze R, Griffith EJ, Banadda N, et al. FluNet: an AI-enabled influenza-like warning system. *IEEE Sens J*. (2021) 21:24740–8. doi: 10.1109/JSEN.2021.3113467

18. Voleti R, Liss JM, Berisha V. A review of automated speech and language features for assessment of cognitive and thought disorders. *J Sel Top Signal Process*. (2019) 14:282–98. doi: 10.1109/JSTSP.2019.2952087

19. Miner AS, Haque A, Fries JA, Fleming SL, Wilfley DE, Terence Wilson G, et al. Assessing the accuracy of automatic speech recognition for psychotherapy. *NPJ Digit Med*. (2020) 3:1–8. doi: 10.1038/s41746-020-0285-8

20. Zhang T, Schoene AM, Ji S, Ananiadou S. Natural language processing applied to mental illness detection: a narrative review. *NPJ Digit Med*. (2022) 5:1–13. doi: 10.1038/s41746-022-00589-7

21. Le D, Provost EM. *Modeling pronunciation, rhythm, and intonation for automatic assessment of speech quality in aphasia rehabilitation*. Singapore: ISCA (2014). p. 1–5.

22. Le D, Licata K, Provost EM. Automatic paraphasia detection from aphasic speech: a preliminary study. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Stockholm, Sweden: ISCA (2017). p. 294–8.

23. DeLisi LE. Speech disorder in schizophrenia: review of the literature and exploration of its relation to the uniquely human capacity for language. *Schizophr Bull*. (2001) 27:481–96. doi: 10.1093/oxfordjournals.schbul.a006889

24. Tahir Y, Chakraborty D, Dauwels J, Thalmann N, Thalmann D, Lee J. Non-verbal speech analysis of interviews with schizophrenic patients. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai, China: IEEE (2016). p. 5810–4.

25. He F, He L, Zhang J, Li Y, Xiong X. Automatic detection of affective flattening in schizophrenia: acoustic correlates to sound waves and auditory perception. *IEEE/ACM Trans Audio Speech Lang Process*. (2021) 29:3321–34. doi: 10.1109/TASLP.2021.3120591

26. Gernsbacher MA, Morson EM, Grace EJ. Language and speech in autism. *Annu Rev Linguist*. (2016) 2:413. doi: 10.1146/annurev-linguistics-030514-124824

27. Rynkiewicz A, Schuller B, Marchi E, Piana S, Camurri A, Lassalle A, et al. An investigation of the "female camouflage effect" in autism using a computerized

ADOS-2 and a test of sex/gender differences. *Mol Autism*. (2016) 7:1–8. doi: 10.1186/s13229-016-0073-0

28. Pokorny F, Schuller B, Marschik P, Brueckner R, Nyström P, Cummins N, et al. Earlier identification of children with autism spectrum disorder: an automatic vocalisation-based approach. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 2017. Stockholm, Sweden: ISCA (2017). p. 309–13.

29. Roche L, Zhang D, Bartl-Pokorny KD, Pokorny FB, Schuller BW, Esposito G, et al. Early vocal development in autism spectrum disorder, Rett syndrome, and fragile X syndrome: insights from studies using retrospective video analysis. *Adv Neurodev Disord*. (2018) 2:49–61. doi: 10.1007/s41252-017-0051-3

30. Rudovic O, Lee J, Dai M, Schuller B, Picard RW. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Sci Robot*. (2018) 3: eaao6760. doi: 10.1126/scirobotics.aao6760

31. France DJ, Shiavi RG, Silverman S, Silverman M, Wilkes M. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Trans Biomed Eng*. (2000) 47:829–37. doi: 10.1109/10.846676

32. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. *Speech Commun*. (2015) 71:10–49. doi: 10.1016/j.specom.2015.03.004

33. Ringeval F, Schuller B, Valstar M, Cummins N, Cowie R, Tavabi L, et al. AVEC 2019 workshop, challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. *Proceedings of the Audio/Visual Emotion Challenge and Workshop*. Nice, France: Association for Computing Machinery (2018). p. 3–12.

34. Laukka P, Linnman C, Åhs F, Pissiota A, Frans Ö, Faria V, et al. In a nervous voice: acoustic analysis and perception of anxiety in social phobics' speech. *J Nonverbal Behav*. (2008) 32:195–214. doi: 10.1007/s10919-008-0055-9

35. Baird A, Triantafyllopoulos A, Zänkert S, Ottl S, Christ L, Stappen L, et al. An evaluation of speech-based recognition of emotional and physiological markers of stress. *Front Comput Sci*. (2021) 3:1–19. doi: 10.3389/fcomp.2021.750284

36. Janott C, Schmitt M, Zhang Y, Qian K, Pandit V, Zhang Z, et al. Snoring classified: the Munich-Passau snore sound corpus. *Comput Biol Med*. (2018) 94:106–18. doi: 10.1016/j.compbiomed.2018.01.007

37. Korompili G, Amfilochiou A, Kokkalas L, Mitilineos SA, Tatlas NA, Kouvaras M, et al. PSG-audio, a scored polysomnography dataset with simultaneous audio recordings for sleep apnea studies. *Sci Data*. (2021) 8:1–13. doi: 10.1038/s41597-021-00977-w

38. Schuller BW, Batliner A, Bergler C, Pokorny FB, Krajewski J, Cychosz M, et al. The INTERSPEECH 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds & Orca activity. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Graz, Austria: ISCA (2019). p. 2378–82.

39. Duckitt W, Tuomi S, Niesler T. Automatic detection, segmentation and assessment of snoring from ambient acoustic data. *Physiol Meas*. (2006) 27:1047. doi: 10.1088/0967-3334/27/10/010

40. Hönig F, Batliner A, Nöth E, Schnieder S, Krajewski J. *Automatic modelling of depressed speech: relevant features and relevance of gender*. Singapore: ISCA (2014). p. 1–5.

41. J Holmes R, M Oates J, J Phyland D, J Hughes A. Voice characteristics in the progression of Parkinson's disease. *Int J Lang Commun Disord*. (2000) 35:407–18. doi: 10.1080/136828200410654

42. Midi I, Dogan M, Koseoglu M, Can G, Sehitoglu M, Gunal D. Voice abnormalities and their relation with motor dysfunction in Parkinson's disease. *Acta Neurol Scand*. (2008) 117:26–34. doi: 10.1111/j.1600-0404.2007.00965.x

43. Hoffmann I, Nemeth D, Dye CD, Pákáski M, Irinyi T, Kálmán J. Temporal parameters of spontaneous speech in Alzheimer's disease. *Int J Speech Lang Pathol*. (2010) 12:29–34. doi: 10.3109/17549500903137256

44. de la Fuente Garcia S, Ritchie CW, Luz S. Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review. *J Alzheimers Dis*. (2020) 78:1547–74. doi: 10.3233/JAD-200888

45. Luz S, Haider F, Fromm D, MacWhinney B. Alzheimer's dementia recognition through spontaneous speech: the ADReSS challenge. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Virtual Conference: ISCA (2020). p. 2172–6.

46. Noffs G, Perera T, Kolbe SC, Shanahan CJ, Boonstra FM, Evans A, et al. What speech can tell us: a systematic review of dysarthria characteristics in multiple sclerosis. *Autoimmun Rev*. (2018) 17:1202–9. doi: 10.1016/j.autrev.2018.06.010

47. Vieira FG, Venugopalan S, Premasiri AS, McNally M, Jansen A, McCloskey K, et al. A machine-learning based objective measure for ALS disease severity. *NPJ Digit Med*. (2022) 5:1–9. doi: 10.1038/s41746-022-00588-8

48. Nordberg A, Miniscalco C, Lohmander A. Consonant production and overall speech characteristics in school-aged children with cerebral palsy and speech impairment. *Int J Speech Lang Pathol*. (2014) 16:386–95. doi: 10.3109/17549507.2014.917440

49. Chizner MA. Cardiac auscultation: rediscovering the lost art. *Curr Probl Cardiol*. (2008) 33:326–408. doi: 10.1016/j.cpcardiol.2008.03.003

50. Clifford GD, Liu C, Moody B, Springer D, Silva I, Li Q, et al. Classification of normal/abnormal heart sound recordings: the physionet/computing in cardiology challenge 2016. *Proceedings of the Computing in Cardiology Conference (CinC)*. Vancouver, Canada: IEEE (2016). p. 609–12.

51. Schuller B, Steidl S, Batliner A, Marschik PB, Baumeister H, Dong F, et al. The INTERSPEECH 2018 computational paralinguistics challenge: atypical and self-assessed affect, crying and heart beats. In: *Proceedings of the annual conference of the international speech communication association, INTERSPEECH*. Stockholm: ISCA (2017). Vol. 2017. p. 3442–6.

52. Singh J, Anand RS. Computer aided analysis of phonocardiogram. *J Med Eng Technol*. (2007) 31:319–23. doi: 10.1080/03091900500282772

53. Oliveira J, Renna F, Costa PD, Nogueira M, Oliveira C, Ferreira C, et al. The CirCor DigiScope dataset: from murmur detection to murmur classification. *IEEE J Biomed Health Inform*. (2021) 26:2524–35. doi: 10.1109/JBHI.2021.3137048

54. Triantafyllopoulos A, Fendler M, Batliner A, Gerczuk M, Amiriparian S, Berghaus T, et al. Distinguishing between pre-, post-treatment in the speech of patients with chronic obstructive pulmonary disease. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Incheon, South Korea: ISCA (2022). p. 3623–7.

55. Claxton S, Porter P, Brisbane J, Bear N, Wood J, Peltonen V, et al. Identifying acute exacerbations of chronic obstructive pulmonary disease using patient-reported symptoms and cough feature analysis. *NPJ Digit Med*. (2021) 4:1–7. doi: 10.1038/s41746-021-00472-x

56. Kutor J, Balapangu S, Adofo JK, Dellor AA, Nyakpo C, Brown GA. Speech signal analysis as an alternative to spirometry in asthma diagnosis: investigating the linear and polynomial correlation coefficient. *Int J Speech Technol*. (2019) 22:611–20. doi: 10.1007/s10772-019-09608-7

57. Kosasih K, Abeyratne UR, Swarnkar V, Triasih R. Wavelet augmented cough analysis for rapid childhood pneumonia diagnosis. *IEEE Trans Biomed Eng*. (2014) 62:1185–94. doi: 10.1109/TBME.2014.2381214

58. Deshpande G, Batliner A, Schuller BW. AI-based human audio processing for COVID-19: a comprehensive overview. *Pattern Recognit*. (2022) 122:108289. doi: 10.1016/j.patcog.2021.108289

59. Han J, Xia T, Spathis D, Bondareva E, Brown C, Chauhan J, et al. Sounds of COVID-19: exploring realistic performance of audio-based digital testing. *NPJ Digit Med*. (2022) 5:1–9. doi: 10.1038/s41746-021-00553-x

60. Triantafyllopoulos A, Semertzidou A, Song M, Pokorny FB, Schuller BW. COVYT: introducing the Coronavirus YouTube and TikTok speech dataset featuring the same speakers with and without infection [Preprint] (2022). p. 1–12. Available at: https://arxiv.org/2206.11045

61. Sharma N, Krishnan P, Kumar R, Ramoji S, Chetupalli SR, Nirmala R, et al. Coswara: a database of breathing, cough, and voice sounds for COVID-19 diagnosis. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Virtual Conference: ISCA (2020).

62. Brown C, Chauhan J, Grammenos A, Han J, Hasthanasombat A, Spathis D, et al. Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data. *Proceedings of the International Conference on Knowledge Discovery & Data Mining (SIGKDD)*; San Diego, USA (2020). p. 3474–84.

63. Bartl-Pokorny KD, Pokorny FB, Batliner A, Amiriparian S, Semertzidou A, Eyben F, et al. The voice of COVID-19: acoustic correlates of infection in sustained vowels. *J Acoust Soc Am*. (2021) 149:4377–83. doi: 10.1121/10.0005194

64. Grieco JC, Bahr RH, Schoenberg MR, Conover L, Mackie LN, Weeber EJ. Quantitative measurement of communication ability in children with Angelman syndrome. *J Appl Res Intellect Disabil*. (2018) 31:e49–e58. doi: 10.1111/jar.12305

65. Bartl-Pokorny KD, Pokorny FB, Garrido D, Schuller BW, Zhang D, Marschik PB. Vocalisation repertoire at the end of the first year of life: an exploratory comparison of Rett syndrome and typical development. *J Dev Phys Disabil*. (2022) 34:1053–69. doi: 10.1007/s10882-022-09837-w

66. Pokorny FB, Schmitt M, Egger M, Bartl-Pokorny KD, Zhang D, Schuller BW, et al. Automatic vocalisation-based detection of fragile X syndrome and Rett syndrome. *Sci Rep*. (2022) 12:1–13. doi: 10.1038/s41598-022-17203-1

67. Anguera X, Bozonnet S, Evans N, Fredouille C, Friedland G, Vinyals O. Speaker diarization: a review of recent research. *IEEE/ACM Trans Audio Speech Lang Process*. (2012) 20:356–70. doi: 10.1109/TASL.2011.2125954

68. Wang D, Chen J. Supervised speech separation based on deep learning: an overview. *IEEE/ACM Trans Audio Speech Lang Process*. (2018) 26:1702–26. doi: 10.1109/TASLP.2018.2842159

69. Snyder D, Garcia-Romero D, Sell G, Povey D, Khudanpur S. X-vectors: robust DNN embeddings for speaker recognition. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Australia: IEEE (2018). p. 5329–33.

70. Jokić S, Cleres D, Rassouli F, Steurer-Stey C, Puhan MA, Brutsche M, et al. TripletCough: cougher identification and verification from contact-free smartphone-based audio recordings using metric learning. *IEEE J Biomed Health Inform* 26 (2022) 2746–57. doi: 10.1109/JBHI.2022.3152944

71. Liu S, Keren G, Parada-Cabaleiro E, Schuller B. N-HANS: a neural network-based toolkit for in-the-wild audio enhancement. *Multimed Tools Appl*. (2021) 80:28365–89. doi: 10.1007/s11042-021-11080-y

72. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. *Annu Rev Clin Psychol.* (2008) 4:1–32. doi: 10.1146/annurev.clinpsy.3.022806.091415

73. Cornet VP, Holden RJ. Systematic review of smartphone-based passive sensing for health and wellbeing. *J Biomed Inform.* (2018) 77:120–32. doi: 10.1016/j.jbi.2017.12.008

74. Jin Q, Schultz T, Waibel A. Far-field speaker recognition. *IEEE Trans Audio Speech Lang Process.* (2007) 15:2023–32. doi: 10.1109/TASL.2007.902876

75. Milling M, Baird A, Bartl-Pokorny K, Liu S, Alcorn A, Shen J, et al. Evaluating the impact of voice activity detection on speech emotion recognition for autistic children. *Front Comput Sci.* (2022) 4:837269. doi: 10.3389/fcomp.2022.837269

76. Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics (2019). p. 3645–50.

77. Cheng Y, Wang D, Zhou P, Zhang T. A survey of model compression and acceleration for deep neural networks [Preprint] (2017). p. 1–10. Available at: https://arxiv.org/1710.09282

78. Amiriparian S, Hübner T, Karas V, Gerczuk M, Ottl S, Schuller BW. DeepSpectrumLite: a power-efficient transfer learning framework for embedded speech and audio processing from decentralised data. *Frontiers in Artificial Intelligence, Section Language and Computation*. Frontiers Press (2022). 16 p.

79. Guedes V, Teixeira F, Oliveira A, Fernandes J, Silva L, Junior A, et al. Transfer learning with audioset to voice pathologies identification in continuous speech. *Procedia Comput Sci.* (2019) 164:662–9. doi: 10.1016/j.procs.2019.12.233

80. Sertolli B, Ren Z, Schuller BW, Cummins N. Representation transfer learning from deep end-to-end speech recognition networks for the classification of health states from speech. *Comput Speech Lang.* (2021) 68:101204. doi: 10.1016/j.csl.2021.101204

81. Amiriparian S, Gerczuk M, Ottl S, Cummins N, Freitag M, Pugachevskiy S, et al. Snore sound classification using image-based deep spectrum features. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Stockholm, Sweden: ISCA (2017). p. 3512–6.

82. Amiriparian S. *Deep representation learning techniques for audio signal processing* [Dissertation]. München: Technische Universität München (2019).

83. Triantafyllopoulos A, Schuller BW. The role of task and acoustic similarity in audio transfer learning: insights from the speech emotion recognition case. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, Canada: IEEE (2021). p. 7268–72.

84. Wagner J, Triantafyllopoulos A, Wierstorf H, Schmitt M, Eyben F, Schuller BW. Dawn of the transformer era in speech emotion recognition closing the valence gap [Preprint] (2022). p. 1–25. Available at: https://arxiv.org/2203.07378

85. Baevski A, Zhou Y, Mohamed A, Auli M. wav2vec 2.0: a framework for self-supervised learning of speech representations. *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*. Vancouver, Canada: Curran Associates Inc. (2020). p. 12449–60.

86. Amiriparian S, Pugachevskiy S, Cummins N, Hantke S, Pohjalainen J, Keren G, et al. Cast a database: rapid targeted large-scale big data acquisition via small-world modelling of social media platforms. *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*. San Antonio, USA: IEEE (2017). p. 340–5.

87. Zou L, Ruan F, Huang M, Liang L, Huang H, Hong Z, et al. SARS-CoV-2 viral load in upper respiratory specimens of infected patients. *N Engl J Med.* (2020) 382:1177–9. doi: 10.1056/NEJMc2001737

88. Amieva H, Le Goff M, Millet X, Orgogozo JM, Pérès K, Barberger-Gateau P, et al. Prodromal Alzheimer's disease: successive emergence of the clinical symptoms. *Ann Neurol.* (2008) 64:492–8. doi: 10.1002/ana.21509

89. Wilson RS, Beckett LA, Barnes LL, Schneider JA, Bach J, Evans DA, et al. Individual differences in rates of change in cognitive abilities of older persons. *Psychol Aging.* (2002) 17:179. doi: 10.1037/0882-7974.17.2.179

90. Pinto MF, Oliveira H, Batista S, Cruz L, Pinto M, Correia I, et al. Prediction of disease progression and outcomes in multiple sclerosis with machine learning. *Sci Rep.* (2020) 10:1–13. doi: 10.1038/s41598-020-78212-6

91. Hizel C, Tremblay J, Bartlett G, Hamet P. Chapter 1 - Introduction: every individual is different and precision medicine offers options for disease control and treatment. *Progress and Challenges in Precision Medicine*. Cambridge, MA: Academic Press (2017). p. 1–34.

92. Mazzone SB, Chung KF, McGarvey L. The heterogeneity of chronic cough: a case for endotypes of cough hypersensitivity. *Lancet Respir Med.* (2018) 6:636–46. doi: 10.1016/S2213-2600(18)30150-4

93. Triantafyllopoulos A, Liu S, Schuller BW. Deep speaker conditioning for speech emotion recognition. *Proceedings of the International Conference on Multimedia and Expo (ICME)*. Virtual Conference: IEEE (2021). p. 1–6.

94. Chén OY, Roberts B. Personalized health care and public health in the digital age. *Front Digit Health*. (2021) 3:595704. doi: 10.3389/fdgth.2021.595704

95. Gerczuk M, Triantafyllopoulos A, Amiriparian S, Kathan A, Bauer J, Schuller B, Personalised deep learning for monitoring depressed mood from speech. *Proceedings of the E-Health and Bioengineering Conference (EHB)*. Iași, Romania: IEEE (2022). p. 1–5.

96. Kathan A, Harrer M, Küster L, Triantafyllopoulos A, He X, Milling M, et al. Personalised depression forecasting using mobile sensor data and ecological momentary assessment. *Front Digit Health.* (2022) 4:964582. doi: 10.3389/fdgth.2022.964582

97. Schölkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, et al. Toward causal representation learning. *Proc IEEE.* (2021) 109:612–34. doi: 10.1109/JPROC.2021.3058954

98. Yunis M, Markarian C, El-Kassar A. A conceptual model for sustainable adoption of ehealth: role of digital transformation culture and healthcare provider's readiness. *Proceedings of the IMCIC* (2020). Orlando: International Institute of Informatics and Systemics (IIIS).

99. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. (2020) 58:82–115. doi: 10.1016/j.inffus.2019.12.012

100. Du M, Yang F, Zou N, Hu X. Fairness in deep learning: a computational perspective. *IEEE Intell Syst.* (2020) 36:25–34. doi: 10.1109/MIS.2020.3000681

101. Wang T, Zhao J, Yatskar M, Chang KW, Ordonez V. Balanced datasets are not enough: estimating and mitigating gender bias in deep image representations. *Proceedings of the International Conference on Computer Vision (ICCV)*. Seoul, South Korea: IEEE/CVF (2019). p. 5310–9.

102. Zafar MB, Valera I, Rogriguez MG, Gummadi KP. Fairness constraints: mechanisms for fair classification. *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*. Lauderdale, USA: JMLR (2017). p. 962–70.

103. Jordon J, Yoon J, Van Der Schaar M. ((((PATE-GAN generating synthetic data with differential privacy guarantees. *Proceedings of the International Conference on Learning Representations*; New Orleans, USA (2019). p. 1–21.

104. Dwork C. Differential privacy. *International colloquium on automata, languages, and programming*. Heidelberg, Germany: Springer (2006). p. 1–12.

105. Dankar FK, El Emam K. The application of differential privacy to health data. *Proceedings of the 2012 Joint EDBT/ICDT Workshops*. Berlin: ACM (2012). p. 158–66.

106. Fallah A, Mokhtari A, Ozdaglar A. Personalized federated learning with theoretical guarantees: a model-agnostic meta-learning approach. *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*. Vancouver, Canada: Curran Associates Inc. (2020). p. 3557–68.

107. Song S, Chaudhuri K, Sarwate AD. Stochastic gradient descent with differentially private updates. *Proceedings of the Global Conference on Signal and Information Processing*. Austin, USA: IEEE (2013). p. 245–8.

108. Papernot N, Song S, Mironov I, Raghunathan A, Talwar K, Erlingsson U. Scalable private learning with pate. *Proceedings of the International Conference on Learning Representations*; Vancouver, Canada (2018). p. 1–34.

109. Emanuel EJ, Emanuel LL. Four models of the physician-patient relationship. *JAMA*. (1992) 267:2221–6. doi: 10.1001/jama.1992.03480160079038

110. Percival T. *Medical ethics*. Cambridge: Cambridge University Press (2014).

111. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. (2017) 30.

112. Samek W, Binder A, Montavon G, Lapuschkin S, Müller KR. Evaluating the visualization of what a deep neural network has learned. *IEEE Trans Neural Netw Learn Syst.* (2016) 28:2660–73. doi: 10.1109/TNNLS.2016.2599820

113. Montavon G, Binder A, Lapuschkin S, Samek W, Múller KR. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*. New York: Springer (2019). p. 193–209.

114. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; San Francisco, CA, USA (2016). p. 1135–44.

115. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*. (2018) 6:52138–60. doi: 10.1109/ACCESS.2018.2870052

116. Mothilal RK, Sharma A, Tan C. Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*; Barcelona, Spain (2020). p. 607–17.

117. Drossos K, Adavanne S, Virtanen T. Automated audio captioning with recurrent neural networks. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY: IEEE (2017). p. 374–8.

118. Schuller BW, Virtanen T, Riveiro M, Rizos G, Han J, Mesaros A, et al. Towards sonification in multimodal and user-friendly explainable artificial intelligence. *Proceedings of the International Conference on Multimodal Interaction (ICMI)*; Montreal, Canada (2021). p. 788–92.

119. Jimenez Morgan S, Molina Mora JA. Effect of heart rate variability biofeedback on sport performance, a systematic review. *Appl Psychophysiol Biofeedback*. (2017) 42:235–45. doi: 10.1007/s10484-017-9364-2

120. Kusner MJ, Loftus J, Russell C, Silva R. Counterfactual fairness. *Adv Neural Inf Process Syst*. (2017) 30.