Speech-based Depression Assessment: A Comprehensive Survey

Samara Soares Leal, Stavros Ntalampiras, and Roberto Sassi

***Abstract*—Depression (major depressive disorder) is one of the most common mental illnesses worldwide, causing feelings of sadness and loss of interest, and is a leading cause of suicidal ideation. Limited access to mental health services, stigma, patient privacy and delay in seeking help are the most significant barriers to assessment and effective treatment. In order to enhance the accuracy of depression prediction, automated strategies employing computational models have been widely explored in literature. To this end, automatic Speech Depression Recognition (SDR) methods stand out, as speech comprises a valuable marker of mental health. Interestingly, recording speech comprises a less intrusive and more portable approach than capturing video, thus more easily accepted, especially by the younger generations, who are at a considerable risk of social isolation due to addiction to social networks and excessive use of mobile devices. In this context, this paper presents an up-to-date survey on SDR. More specifically, we a) detail the major challenges and key issues on SDR, b) summarise the most recent approaches existing in the related literature, and c) highlight the open problems. At the same time, we illustrate a framework encompassing the latest tendencies for SDR, along with a suitable comparison of the achieved performances. Finally, we highlight future trends and present the overall findings, providing researchers with best practices and techniques to address the major challenges of SDR, as well as stimulating discussion and improvement in the field.**

***Index Terms*—Affective computing, paralinguistic speech processing, acoustic signal processing, audio pattern recognition, speech depression recognition, mental health.**

I. Introduction

DEPRESSION, also known as Major Depressive Disorder (MDD), has been evolving into a global health crisis with a considerable impact in modern society [1]. It is a common psychiatric disorder with serious negative consequences on an individual's cognitive, emotional and behavioural functioning. According to the World Health Organization (WHO), 322 million people worldwide suffer from MDD. That's 4.4% of the world's population [2]. In severe cases, such a disease can lead to suicidal ideation, as people suffering from depression are 20 times more likely to die by suicide [3]. Unfortunately, MDD has become increasingly common among young people, coinciding with trends like increased indoor lifestyles and higher use of video games and mobile devices [2].

The Diagnostic and Statistical Manual of Mental Disorders (DSM), the standard for psychiatric diagnosis, describes indicators of depression that are often overlooked in its assessments [5]. While objective physiological indicators are also used [1], depression is usually assessed on the basis of the patient's own verbal description of symptoms during appointments with psychiatrists who may leverage traditional questionnaires.

Although these tools are useful, there are still concerns about bias in professional judgment and issues with subjective patient responses due to stigma [7]. Also self-reported responses lack accuracy and screening based on clinical interviews is labor intensive [6]. In addition, the questionnaires do not include visual, acoustic or textual indicators of depression [7]. To overcome these limitations, recent research has focused on depression recognition from video [8], [9], text [10], [11], speech signals [13]–[15], electroencephalogram (EEG) [17], [18], multi-modal fusion approaches [7], [19], [20], and even from social media feeds and comments [2], [21], using computational algorithms.

Among these solutions, speech signals have several advantages as an application and are known to provide indicators of mental disorders [22], [23], [25]: a) speech signals can reflect a subject's tendency to be depressed with slowed speech rate, prolonged pauses and different pitch changes [1], b) it is one of the least intrusive and costly methods, c) includes the possibility of detecting depression via smartphones or laptop microphones [23]; as such, it can be conveniently implemented on a wide variety of platforms, unlike EEG and other approaches [22], and d) speech directly expresses emotions [12], [24] and reflects neural modulation through motor and acoustic changes, making it difficult to hide symptoms of depression [25].

Therefore, several recent researches have focused on artificial intelligence (AI), in particular Machine Learning (ML) models for automatic assessment of MDD from speech signals. These models have been continuously proposed and updated, with a transition from the early traditional hand-crafted feature based models to the application of end-to-end Deep Learning (DL) architectures [1], [25].

A. Focus of this Survey

Despite the advantages of SDR, there are many challenges that still require attention. The majority of the existing research deals with the problem of small and imbalanced depression assessment datasets to train the computational models [15], [19], [27]–[30] for depression prediction. This is often due to privacy issues [31], as most datasets are not publicly available [9], mainly given the data contains sensitive information that could affect relationships and employment [32]. Furthermore, such small and unbalanced datasets present an increased risk of model overfitting and a reduce chance of generalising [30].

The authors are with the Department of Computer Science, University of Milan, Milan, 20133, Italy. E-mail: name.surname@unimi.it

Manuscript received April 19, 2021; revised August 16, 2021.

In this context, this paper presents a survey of recent developments in speech signal processing for automatic depression detection. The main goal of this work is to provide a complete and up-to-date review of recent efforts to address the major challenges and issues that remain to be tackled in this area, such as: a) the dataset size and diversity [33], b) the interference from the places where the patient's speech is recorded [34], c) the way in which they are recorded [35], d) the type of speech recording [29], [36], e) usability issues between AI systems and healthcare professionals [37], f) whether the identity of the patient is taken into account or not [14], [38], g) the interpretability and reliability of AI-based decision-making systems [39], and h) the potential biases in the data that favor certain sub-populations [25].

The above-mentioned key issues will be discussed in detail in section III, thus filling the gap existing in other recent surveys which do not thoroughly consider such aspects. These issues have an impact on the results of depression prediction models because they affect the preprocessing of the data as well as the model accuracy [41]. Several of the previous surveys have also become outdated, such as [25], [40], mostly because of the achievements and rapid innovations observed in recent years.

Given this, the aim of the present study is to provide a comprehensive and in-depth analysis of the latest advances and existing challenges in the automatic detection of depression through speech signals. To this end, the following specific objectives (SO) are defined:

- **SO1**: to present the limitations of the traditional depression assessment procedures;
- **SO2**: to discuss the challenges and issues that have not been yet fully addressed by existing surveys;
- **SO3**: To present the latest scientific works using handcrafted Machine Learning (ML) models and end-to-end deep architectures to predict depression and their performances;
- **SO4**: to provide a comprehensive survey of the current progress made against major challenges in SDR. It is intended to offer researchers good practices and techniques, as well as to foster discussions and improvements in the field;
- **SO5**: to highlight the open issues and discuss future research directions, towards improving current solutions and contribute to the field.

B. Survey Organization

The rest of the survey is organized as follows: section II provides the definition of MDD and its standard questionnaires and assessment procedures. Subsequently, section III lists the key issues in the field as well as the major challenges in SDR. Section IV presents a framework encompassing the latest SDR methods, taking into account the datasets used, the acoustic feature extraction tools and the computational algorithms, as well as the reported performances. Finally, open issues and future directions are pointed out in section V while section VI offers a discussion and the conclusions of this study.

II. Major Depressive Disorder

According to the WHO, MDD is a mental disorder that involves a depressed mood or loss of pleasure or interest in activities over a long period of time [52]. WHO ranks depressive disorders as the most common mental disorder and the second most common disease [53].

Furthermore, the WHO has indicated that depressive illness can precipitate suicidal ideation, with individuals experiencing depressive disorders being 20 times more susceptible to suicidal death. [3], [52]. Thereby, MDD is associated with high morbidity and mortality and affects multiple domains of an individual's life, including work, school, and social life [52], [55]. Despite existing investments in therapeutic and psychiatric interventions, treatment for this disorder is usually delayed and inaccessible to the majority of the patients [56], not to mention the stigma and biases that still surround mental illnesses [25], [32].

The American Psychiatric Association (APA) classifies depression as a mood disorder in the DSM [54]. The key physiological symptoms of MDD are agitation and fatigue. Furthermore, psychomotor retardation and cognitive impairment are other common and early symptoms of depression that can manifest as changes in speech [15], [57]. However, the most widely used method to screen this illness is still through traditional standardized questionnaires [5], despite their limitations [7].

The Patient Health Questionnaire (PHQ) is the standardized method for screening depression [**?**], [58]. APA designed the PHQ as a brief self-report instrument that packages the DSM depression criteria and asks individuals about their experiences with symptoms such as fatigue, trouble concentrating, difficulty sleeping, poor appetite, and enjoyment of family [19], [59]. Unfortunately, diagnostic methods relying on patient cooperation, expressiveness, and clinician judgment [6], [25], may be subjective and inaccurate [6]. Moreover, these questionnaires also lack visual, acoustic and textual indicators of depression [7].

Misdiagnosis of depression or delay in diagnosis can interfere with treatment and may worsen the patient's condition [1]. In this context, the design of automated systems is attracting ever-increasing interest from the affective computing and AI communities to assist clinicians in efficiently assessing depression [66]. Although the specific scientific field has grown significantly in the recent years with considerable contributions, there are still several challenges and issues that the community needs to address, which are discussed in the next section.

III. Key Issues and Challenges

This section organizes the major challenges (MC) and the associated key issues (KI) faced by researchers in SDR. To identify these significant challenges, a comprehensive review of the literature was conducted, with studies selected based on their relevance and credibility. During this review, several recurring themes emerged across multiple studies, forming the foundation of the challenges discussed. The key issues were derived directly from the literature, through the recent surveys

TABLE I
AN OVERVIEW OF EXISTING SURVEYS ON SDR

Author	Title	Main Contributions	MC/KI NOT addressed
Cummins (2015) [167]	A review of depression and suicide risk assessment using speech analysis	• Reviews the key non-speech biological, physiological and behavioural markers to assess depression; • Review the main features of SDR databases; • Discusses how common paralinguistic speech features are affected by MDD and the use of this information in classification and prediction systems.	MC4: KI9 MC2: KI4
Schuller (2018) [41]	Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends	• Covers progress on speech emotion recognition up to 2013; • Benchmarks of the AudioVisual Emotion Challenge (AVEC) competitions held in the field up to 2017 [69]; • Points out that there exists practically no engine on speech emotion recognition that is adaptive to cultural differences.	MC2: KI5, KI6 MC3: KI7, KI8 MC4: KI9
Robin et al., (2020) [71]	Evaluation of Speech-Based Digital Biomarkers: Review and Recommendations	• Outlines the literature of evaluation steps for speech-based digital biomarkers, based on the V3 framework [72]; • Discusses that speech features vary widely across studies, and no speech measure has yet been evaluated across all three categories of the V3 framework: verification, analytical validation, and clinical validation.	MC1: KI1, KI3 MC2: KI4 MC3: KI7, KI8 MC4: KI9
Low et al., (2020) [25]	Automated assessment of psychiatric disorders using speech: A systematic review	• Presents the studies on the use of speech for automated assessments in a range of psychiatric disorders, including depression, up to 2019; • Discusses how algorithms can be susceptible to learning biases that are inherent in the data used to train them.	MC4: KI9
Latif et al., (2021) [73]	Speech Technology for Healthcare: Opportunities, Challenges, and State of the Art	• Reviews the approaches in automatic speech recognition (ASR), speech synthesis or text to speech (TTS), and health detection and monitoring using speech signals; • Points out that enabling interoperability in healthcare systems can accelerate diagnostic procedures and provide medical practitioners with a complete patient history.	MC1: KI3 MC2: KI4, KI6
He (2022) [75]	Deep learning (DL) for depression recognition with audiovisual cues: A review	• Reviews the automatic depression detection methods based on deep neural networks up to 2021, discusses their challenges, and points to future research directions; • Discusses the ability of DL models to distinguish between MDD and other types of depression.	MC1: KI2 MC2: KI4, KI6 MC3: KI7, KI8 MC4: KI9
Maithri et al., (2022) [76]	Automated emotion recognition: Current trends and future perspectives	• Provides insights into methods employed using EEG, facial, and speech signals coupled with multi-modal emotion recognition (ER) techniques; • Discusses that there is still a need to achieve high performance in ML and DL based systems for SDR in an uncontrolled environment.	MC1: KI2, KI3 MC2: KI6 MC3: KI7, KI8 MC4: KI9
Wu et al., (2022) [1]	Automatic depression recognition by intelligent speech signal processing: A systematic survey.	• A review of the literature on speech depression recognition (SDR) up to 2021; • Explore research directions for SDR in the future.	MC2: KI6 MC3: KI7, KI8 MC4: KI9

in Table I and all referenced papers, ensuring a transparent and well-supported analysis.

- **MC1** - Size, balance and diversity of the datasets:
  - **KI1**: The size of depression datasets is limited given that in this domain, data collection requires the respect of rigorous ethical constraints that, inevitably, impact the proportions of the corpora that can be collected [19], [38], [67], [68].
  - **KI2**: The diverse range of people participating in the studies, while taking into account factors such as gender, age, language and cultural background. Special attention is placed on age since, nowadays, young people are at higher risk of social isolation due to addiction to social networks and excessive use of mobile devices [2], [4]. Gender is a particularly important factor as well, as WHO highlights that depression rates are higher in women than in men and can emerge during pregnancy [52], [55].
  - **KI3**: The impacts of data biases during the construction of ML models, as small and unbalanced datasets can

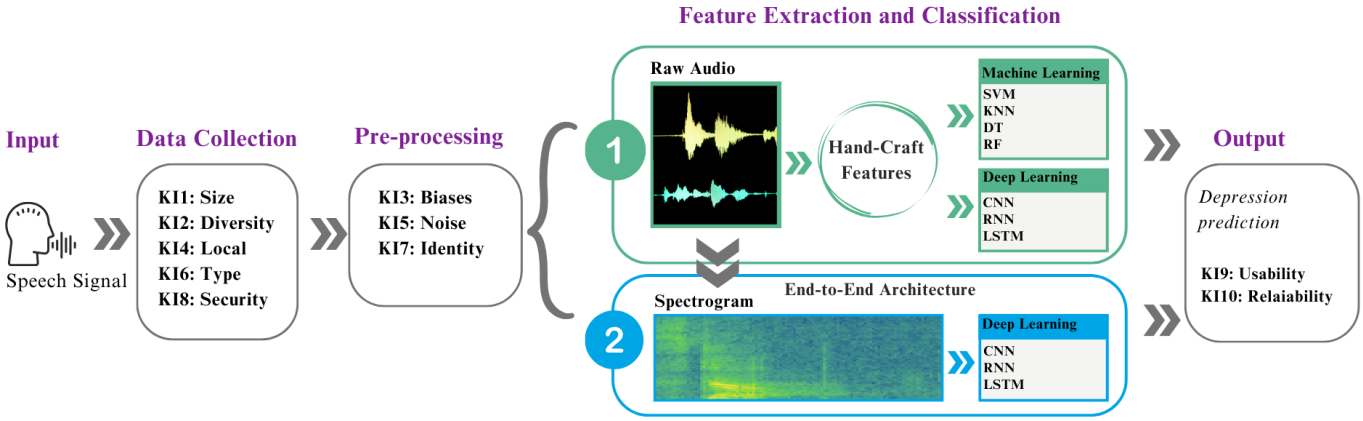

Fig. 1. A summarized framework for using computational methods to automatically predict depression from speech signals.

cause the model to overfit and favor specific subpopulations [30].

- **MC2** - Speech recording procedure:
  - **KI4**: The location where the patient's speech is recorded, such as in a laboratory, in the wild, or on social media, can affect the data quality [34].
  - **KI5**: The way speech signals are recorded can affect the model fit; for instance, the performance of the models can be significantly affected by noise in smartphone-based recording [35]. This could be a pre-processing issue as well as a recording issue. As noted by [168], recording biases in audio duration and intensity can lead to dataset-specific differences between patient and control audios.
  - **KI6**: The type of speech to record (interview, read or spontaneous) may affect SDR efficacy [36].
- **MC3** - Privacy and Ethics constraints:

  KI7: Taking patients' identity into account when evaluating speech representations for personalized or patient-independent models can lead to improved results [14], [38].

  KI8: Creating a more secure and welcoming environment within data collection interfaces can encourage patients to record their speech for depression assessment. By ensuring confidentiality and fostering trust, these interfaces can help minimize the stigma surrounding mental health.
- **MC4** - AI application for SDR:

  KI9: The usability and reliability issues existing between AI systems and healthcare professionals [37]. In accordance with the EU's AI Act, the interpretability and reliability of AI-based decision-making systems must be considered prior to their deployment in clinical settings [39]. Recent studies indicate that automatic speech analysis can assist in the assessment of cognitive disorders. However, further research is required to ascertain the reliability of the acoustic measures [160]. Moreover, any deployed digital health tools should be based on verifiable claims with published evidence of functionality.

We provide a comprehensive overview of these issues, highlighting the main contributions of existing works in addressing them in Table I.

IV. Automatic Depression Detection in Speech

Before reviewing the recent works of the literature in the field, in Fig. 1 we present a generic speech pattern recognition pipeline able to encompass all the processing steps, computational models, and KIs present in SDR.

This framework reflects the most relevant structures in the literature addressing the problem at hand [15], [66], [76], while considering the following steps:

1) **Input**: The first step in processing speech signals is related to instrumentation, with the selection of a suitable recording equipment for speech acquisition. Recording speech signals is crucial despite being overlooked or not even mentioned in a considerable amount of works. There is a significant variability in quality and reliability among different speech acquisition methods [77]. To evaluate speech-based digital biomarkers using the V3 framework [72], it is recommended to analyze the type of recording device to ensure reliable results from acceptable quality recordings [71].
2) **Data Collection and Storage**: Audio recordings are stored in a data structure to train and test the computational models. The most commonly-used datasets are outlined in subsection IV-A. However, the availability of a reliable dataset comprises a significant challenge in the field (KI1, KI2, KI4, KI6, and KI8) mainly due to their limited size and unbalancedness [78]. At the same time, it is important that datasets are accompanied by a standardized, potentially patient-independent, experimental protocol allowing a reliable comparison between contrasted approaches.
3) **Pre-processing**: In order to avoid model overfitting, which are typically caused by dataset characteristics, various pre-processing steps, including silence removal and class imbalance addressing, are performed on the

available speech signals (KI3, KI5 and KI7) [79]. Furthermore, due to the limited number of audio recordings and their varying lengths, previous studies have employed a variety of techniques, including audio random sampling [31], clustering-based resampling [79], cut audio into shorter segments of equal length [80] and/or zero-pad them [118], and segmentation and fusion methods [169] to avoid model bias. While these techniques have yielded promising results, further investigation is necessary to ensure that no depression-related information has been lost and that the models are not learning patterns related to the individual rather than the disease, which could cause the model to overfit [169], [170].

4) **Feature Extraction and Classification**: Acoustic features of depressed patients' speech have been shown to be statistically different from those of non-depressed ones [81]. Thus, they could potentially be used as an objective marker for predicting depression [50]. In fact, the performance of the model depends significantly on the quality of the features [82]. Therefore, this step plays an important role in depression analysis tasks [66]. The acoustic features are described in detail in the subsection IV-B.

   Figure 1 illustrates the two common ways for designing speech signal features [1], [66], [76]: i) hand-crafted feature extraction from raw audio signals feeding an ML or deep neural network, or ii) applying an end-to-end deep architecture that feeds the spectrogram into a deep network to learn high-level features automatically. In classification problems, the goal of extracting features is to reduce the dimensionality of the data vector while revealing the most discriminating characteristics of the signal [83]. For the classification step, the most popular ML methods used for SDR are Support Vector Machines (SVM), k-Nearest Neighbor (KNN), Decision Tree (DT), and Random Forest (RF) [76]. When using ML, we need to design features to extract structured knowledge from data sets [84], [85], but when using DL classifiers (Convolutional Neural Network - CNN, Recurrent Neural Network - RNN or Long Short-Term Memory - LSTM), this is not strictly necessary. The deep methods automatically learn high-level abstract features by building more hidden layer models to improve the accuracy of classification or score prediction [1].The most relevant SDR studies using these computational methods are presented in section IV-C.

5) **Output**: A substantial number of studies in SDR utilize binary classification as the output or prediction task. This involves classifying patients as either depressed or not based on the sum scores of the MDD assessment questionnaires [15], [49], [79]. There are also works on predicting the severity of depression [38], [81], [86]. However, there are still major challenges in ensuring the usability and reliability of these outputs by physicians (KI9). Furthermore, as [163] has observed, the use of a questionnaire sum-score-based approach to predicting depression may impede progress by obscuring insights, given that MDD is a highly heterogeneous diagnosis, with two patients diagnosed according to the DSM-5 criteria potentially exhibiting no shared symptoms. Secondly, there are numerous scales designed to assess MDD, which contributes to a lack of clarity in the definition of this disorder. Additionally, depression rating scales are not unidimensional, a psychometric fact that cannot properly be reflected in one sum-score [164]. Thus, Section V will present a potential future research direction to address these limitations of the sum-score approach. This will entail an investigation of changes in MDD symptom sum-scores over time in clinical trials.

In recent years, studies have also started investigating whether fusing acoustic features can improve the performance of SDR solutions [66]. Such a line of thought combines hand-crafted features with deep-learned ones to predict depression from speech signals (the input, data collection, and pre-processing steps are the same as those from Figure 1).

Combining different modalities can be an effective method for developing an automatic depression detection system [49]. However, the use of other input data, such as video or text, may affect the acceptability of the system in real-world situations [66], [79]. Hence, the development of a depression detection system based on a single modality has received much attention in recent years [79], especially audio [66]. Thus, the next subsection presents and discusses widely employed speech datasets used in the specific field.

A. The Datasets

Unfortunately, there is a limited amount of publicly available datasets designed for conducting research in SDR [9], [33], mostly due to privacy concerns. The Distress Analysis Interview Corpus (DAIC) is one of the most widely used open-access depression datasets [88], containing clinical interviews to identify people with depression [9], [33], [49]. There are a few extensions of DAIC: the DIAC-WOZ dataset annotated by PHQ-8, designed and proposed by the Audio Visual Emotion Challenge (AVEC) 2016 [79], [90] and the E-DAIC dataset collected from semi-clinical interviews [91].

A suitable speech corpus should meet the following requirements [83], [85]: 1) include a representative sample of the patients in the population under study [78]; 2) the speech recordings should be well-documented with metadata about the subjects; 3) the corpus should also include syntactic and prosodic annotation, as well as phonetic transcription [70].

A summary of the most relevant datasets used in the SDR field is presented in Table II. These datasets encompass recordings representing the diverse modalities, i.e. audio (a), visual (v), and text (t), and recording types, i.e. clinical assessment (ca), self-report (self), on social media (sm), reading (r), spontaneous (s) or in the wild (wild).

These datasets have been crucial in analyzing speech patterns of depression and gaining an improved understanding of the specific illness [9].

B. Acoustic Features for SDR

Interestingly, speech signals not only include linguistic content, but also exhibit time-frequency characteristics that

TABLE II
SUMMARY OF THE DATASETS USED IN SDR.

Name	Year	Subjects	Modality	Type
Mundt-35 [92]	2007	35	a	r
AVEC2014 [93]	2014	292	a + v	self + r
DAIC-Woz [88]	2014	110	a + v + t	self + r
E-DAIC [94]	2014	351	a + t	self + r
Rochester [95]	2015	27	v	self+ r
CHI-MEI [96]	2016	53	v	ca + r
BlackDog [178]	2018	130	a + v	self
ORYGEN [179]	2018	245	a + v	s
BD [97]	2018	46	a + v	ca + r
Pittsburgh [98]	2018	57	a + v	ca + r
PRIORI [34]	2018	11	a	s + wild
MODMA [99]	2020	55	a	ca + r
SH2-FS [35]	2020	887	a	s
EATD-Corpus [89]	2022	162	a + t	r
D-Vlog [9]	2022	816	a + v	sm + s
Sorrow Analysis Dataset [22]	2023	64	a	r
CMDC [7]	2023	78	a + v + t	s

are directly related to the severity of various types of mental disorders, including depression [79].

The structure of speech ranges from acoustic information at the lowest level to prosodic, phonetic and ultimately conversational at the highest level [100], [101]. The performance of the SDR models significantly depends on the quality of extracted features [79]. According to [25], the commonly-used features for SDR may be organized in the following classes:

- **Prosodic**: Prosodic features refer to changes over longer time segments, perceived in the rhythm, stress, and intonation of speech [41]. Among these features, the fundamental frequency (F0) has been extensively analyzed as a marker of depression, with numerous studies consistently showing a decrease in F0 among depressed patients [36]. F0 is an objective index of the rate at which the vocal folds open and close across the glottis during phonation [180]. It corresponds to the number of vibratory cycles the vocal folds complete per second and is the primary determinant of the auditory impression of vocal pitch [25]. Additionally, many studies have investigated the pause rate, as depressed individuals often exhibit a higher number of pauses and a slower speech rate [47];
- **Cepstral features**: This class includes features that describe the filter properties of speech by analyzing the vocal tract's resonances and the spectral characteristics of the signal [106]. Among these, the Mel-Frequency Cepstrum Coefficients (MFCCs) are widely studied for SDR, as they provide a parametric representation of the speech signal by modeling the vocal tract independently of vocal fold vibrations [103], [104]. Similarly, the Linear Predictive Cepstrum Coefficients (LPCCs) are another cepstral feature that has been identified as a potential indicator for depression recognition [15], [105], [146]. Both MFCCs and LPCCs operate at the level of the cepstrum, offering distinct but complementary perspectives on the speech signal [15], [16]. Additionally, formants, which represent the resonant frequencies of the vocal tract, and harmonics, derived from the spectrogram, also provide valuable insights into the filter properties of speech and have been analyzed as indicators of MDD [146].
- **Source features**: These features reflect airflow from the lungs through the glottis (glottal features) or vocal fold vibrations (voice quality features). They include jitter, shimmer, tremor, harmonic-to-noise ratio, frequency disturbance ratio, quasi-open quotient, normalized amplitude quotient, and peak slope. Among these, jitter and shimmer have been investigated as potential features for depression recognition, as shown in the work of [146].

The openSMILE toolkit is a commonly utilized software for the extraction of the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [34], [108], [109]. This set contains functionals on the 38 low-level descriptors (LLDs) used as acoustic features in AVEC and in the Computational Paralinguistics Challenge (ComParE) [34]. Although it is a conventional hand-crafted feature extraction tool, some studies have highlighted that automatic feature extraction is more successful in depression detection when compared to high-dimensional openSMILE hand-crafted feature sets and their low-dimensional projections [79]. For purposes of reference, acoustic feature extraction can also be conducted using the Kaldi toolkit to extract the log mel-frequency bank (log-MFB) [34], [110], as well as the COVAREP toolkit. Additionally, open-source software such as the Octave toolkit, pyAudioAnalysis, openEAR, Praat, and Librosa may also be employed for this purpose [15], [25], [26], [79], [80].

The research findings suggest that analyzing the phonetic, prosodic, spectral, and cepstral features of speech could serve as valuable markers for identifying depression [25], [51]. Some features, such as F0, are influenced by factors including age, sex, and the nature of the task [15], [25], [111], reflecting the sexual dimorphism of the vocal folds resulting from hormonal changes during puberty. While F0 is a key acoustic feature, it should not be conflated with voice quality, which is determined by the configuration of the vocal tract, laryngeal anatomy, and learned components [181]. As highlighted by [181], two individuals can phonate at the same F0 but exhibit markedly different voice qualities (e.g., "rough" or "breathy"), underscoring the importance of distinguishing between these concepts in speech-based analyses for depression markers. Furthermore, as spectral and cepstral features describe the filter properties of speech, they can also be indirectly shaped by F0 variations, particularly in voiced phonemes [25].

In addition, the type of speech task might also influence the performance of features extraction for automatic depression assessment [15]. Spontaneous speech, *e.g.* social interactions or interviews, have yielded higher classification performances than reading tasks [50].

C. Pattern Recognition Algorithms

Following the challenging step of feature extraction, the next stage of the framework outlined in Fig. 1 is focused on the classification model. Both traditional ML and DL architectures

TABLE III
REPORTED PERFORMANCE OF ML-BASED SOLUTIONS FOR SDR

Work	ML Method	Dataset	Accuracy (%)
[50]	SVM	own and private	80.3
[113]	SVM	own and private	80
[115]	SVM	own and private	75.8
[117]	SVM	own and private	90
[116]	GMM	DAIC-WOZ	85.88
[105]	LR	own and public	81.82
[35]	LR	SH2-FS	72.9
[15]	RF	own and private	87.5
[120]	RF	own and private	90
[122]	RF	own and private	94.1

have been employed in the literature as described in the following subsections.

1) Machine Learning: In recent years, several ML models have been proposed for SDR [66], [112]. For decades, speech technology has been dominated by models based on the Hidden Markov Model (HMM) and the Gaussian Mixture Model (GMM) [73]. Traditional ML-based algorithms such as Support Vector Machine (SVM), Random Forest (RF), Support Vector Regression (SVR), K-means, Logistic Regression (LR), etc. are also used in SDR [1], [119].

SVM is a frequently utilised model for the processing of speech features, as evidenced by numerous references in the literature [50], including those pertaining to SDR [2], [50], [113], [114]. The highest performance of SVM in depression recognition, as identified in this literature review, was observed in the work of [117], with an accuracy of 90% for detecting depression. However, it cannot be assumed that the observed performance is solely due to the use of SVM, as the performance of any given model is a function of several factors, including sample size, pre-processing, feature selection and the model itself, all of which depend on the specific data being used [25].

Another study with promising performance is presented in [121]; the authors used RF to detect depression and reached an accuracy of 94.1%. A summary of other recent and relevant works' performance using ML in SDR can be seen in Table III. The performance values (accuracy, precision and root mean square error (RMSE)) were obtained from published papers.

Hand-crafted feature-based ML models have been widely used for SDR. However, their ability to predict depression's performance may be limited [66]. Moreover, their design requires human expert knowledge, it may be particularly time-consuming and relies on people's subjective assumptions [79]. In addition, useful information about depression patterns may be lost in hand-crafted features using toolkits [66].

In addition to the constraints imposed by the handcrafted nature of the features, the high-dimensionality of health data and the information embedded within it represents a substantial challenge in developing generalized models that can be effectively deployed in real-world settings. In order to construct robust models through health data for the resolution of complex problems, it is necessary to offset the increase in variability by increasing the sample size in a corresponding manner. However, this results in datasets with a "blind spot," a contiguous region of feature space devoid of any observations, a phenomenon known as the curse of dimensionality [174]. According to [174], this phenomenon can result from one of three factors: the absence of samples from that region, an unfortunate random sampling process that failed to include samples from that region, or a training dataset that is inherently biased and fails to represent samples from that region.

Given such limitations and the recent success of DL models in automatic feature extraction and classification for different applications, researches have began to develop SDR systems using DL methods [79].

2) Deep learning: The target of DL-based methods is to automatically learn high-level features by stacking hidden layer models towards improving the accuracy of classification and/or score prediction [1], [47], [48]. While these predictive DL models have proven to be highly effective, they require significantly more data than traditional ML-based classifiers,such as SVM [25]. Various DL models have been used in SDR, including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory networks (LSTM), Transformer (T), Generative Adversarial Networks (GAN), Deep Convolutional Generative Adversarial Network (DCGAN) and encoder–decoder and auto-encoder architectures (AE) [1], [25], [73].

According to the framework of Figure 1, DL can be used either as a classifier after hand-crafted feature extraction or in an end-to-end fashion to automatically learn high-level features [66] and carry out classification. Although several works discuss the subjectivity and labour cost of hand-crafted features, it remains the most popular method [84]. This is mainly due to the limited scale of current datasets and the fact that end-to-end deep architectures have been criticised for their poor interpretability and flexibility [1].

Nevertheless, recently there has been significant progress in end-to-end deep learned features from raw speech data. This approach has shown promising performance compared to using hand-crafted features, making it a current breakthrough in the field [47]. For instance, [79] used an end-to-end CNN-based autoencoder (CNN AE) technique with a SVM classifier to learn features from raw sequential audio data. The accuracy of the CNN AE was found to be 71%, outperforming the classification model using the hand-crafted openSMILE feature set.

Autoencoder (AE) is an unsupervised learning and special type of deep neural network that aims to reconstruct the input signal from a representation of reduced dimensionality [79]. The combination of this method with a convolutional neural network results in a stronger feature learning ability, which also addresses the issue of sample imbalance in the dataset through the resampling method based on clusters, according to [1].

As regards to automatically-learned features, the most popular solution in the literature is CNNs trained to extract multi-level speech representations [47], [76]. Furthermore, the combination of CNN and LSTM has demonstrated great performance in modelling affective behaviours and associated disorders from speech [73], [74]. For instance, in [124] a model with four local feature learning blocks (LFLBs) along

TABLE IV
REPORTED PERFORMANCE OF DL-BASED SOLUTIONS FOR SDR

Work	DL Method	Dataset	Performance
[28]	CNN	own and private	Accuracy of 78.14%
[47]	CNN AE	DAIC-WOZ	Precision of 71%
[124]	CNN-LSTM	Berlin EmoDB	Accuracy of 95.89%
[125]	CNN	DAIC-WOZ	Accuracy of 86.6%
[126]	CNN	DAIC-WOZ	Accuracy of 82.9%
[127]	CNN	AVEC2014	RMSE of 8.96
[43]	CNN T	DAIC-WOZ MODMA	Precision of 95% Precision of 98.7%
[45]	CNN T	MODMA DAIC-WOZ E-DAIC	Accuracy of 99% Precision of 92.7% Precision of 93%
[128]	DCGAN	AVEC2016	RMSE of 5.52
[129]	RNN	AVEC2014	RMSE of 9.28
[130]	AE	DAIC-WOZ	RMSE of 5.68
[131]	AE	AVEC2017	RMSE of 5.51
[13]	LSTM T	DAIC-WOZ	RMSE of 6.31
[132]	LSTM	DAIC-WOZ	Accuracy of 73.35%
[133]	T	D-vlog	Precision of 85.4%
[144]	Word2Vec GloVe	DAIC-WOZ SH2-FS	Accuracy of 82.9% Accuracy of 72.7%
[172]	Word2Vec	DAIC-WOZ CMDC	F1 score of 79% F1 score of 90.53%

with LSTM was proposed. This work used a combination of CNN and LSTM and achieved an accuracy of 95.33% for depression detection in speaker-dependent evaluation and 95.89% in speaker-independent cases. Table IV summarises the performance reported by the latest studies employing DL-based techniques.

One of the approaches using CNN that can be highlighted in Table IV is [43], where the authors used a parallel CNN and a transformer module with linear attention mechanisms, which reached a 95% precision rate in detecting depression on the DAIC-WOZ dataset. It also achieved a precision rate of 98.7% on the MODMA dataset. Another notable approach in terms of performance [45] used a temporal convolutional transformer with knowledge embedding addressing the joint task of detecting depression and recognising emotions. This method illustrates the 'two birds with one stone' scenario of handling multiple tasks with an unique model.

The authors in [144] proposed a framework for analyzing speech as a sequence of acoustic events, where they tokenize regions of acoustic space into 'words' representing speech events at fixed or irregular intervals. This tokenization allows acoustic word features to be exploited by natural language processing (NLP) methods such as Word2Vec, GloVe and Bag of Words (BoW). This framework achieved improvements in F1 (depressed) of up to 15% and 13% for the SH2-FS and DAIC-WOZ datasets, respectively.

Despite the recent success of DL, the lack of large datasets (KI1) represents a significant challenge for many mental health applications. As a result, instead of training from scratch, studies have also used pre-trained large language models and fine-tuned them on their respective target depression datasets [171]. For instance, the work of [172] presents a transfer learning approach to SDR. In their work, the authors obtained depression-related features by fine-tuning Wav2vec 2.0 and by integrating 1D-CNN and attention pooling structures, generating advanced features at the segment level. They then integrated LSTM and self-attention mechanisms in the realm of prediction results. This approach achieved an F1 score of 79% on the DAIC-WOZ dataset and 90.53% on the CMDC dataset.

Regarding the GAN model, the first study to to apply it for SDR was [128]. The authors proposed a DCGAN model to augment features and improve the estimation of depression severity from speech. They conducted the experiments on speech signals from the AVEC2016 depression dataset and obtained an RMSE of 5.52.

D. Progress Against Major Challenges

In addition to the significant progress made in computational models' performance to predict depression from speech, the literature has also made strides in addressing the major challenges outlined in Section III. Table V presents an overview of this progress, as one of the specific objectives of this paper (SO5) is to provide researchers with a comprehensive survey of good practices and techniques in SDR, as well as to encourage discussions and present the advancements made in the field.

V. OPEN ISSUES AND FUTURE DIRECTIONS

As can be seen in Table V, recent works have made a significant progress in SDR. Even so, there are still several open issues that need to be addressed, especially those related to the usability and reliability of AI-based systems in this field. The lack of these aspects can significantly delay the clinical applications of depression detection systems in practice [39], [134], [135]. Therefore, in order to begin utilizing such systems in real-world settings, it is necessary to shift the focus towards improved privacy, transparency and interpretability.

Researchers have increasingly considered the perspective of explainable AI (XAI) in their studies on affective computing and SDR. According to [151], three major challenges in XAI are: i) how to relate explanations to multimodal and time-dependent data; ii) how to integrate context and inductive biases into explanations using mechanisms such as attention, GAN, or graph-based methods; and iii) how to capture both intra-modal and cross-modal interactions in post-hoc explanations. Exploring the V3 framework proposed by [72] and the EU's AI Act [71] can set future directions to evaluate the applicability and reliability of speech-based digital biomarkers. Future researches also need to address issues such as fidelity and causality in SDR [151].

Both privacy and ethical issues have to be thoroughly considered while using speech processing systems in healthcare. In Table V there are several methods proposed to address them (KI7 and KI8). Beyond these contributions, future directions for protecting speaker and gender identity include the use of privacy-preserving deep learning algorithms [152], [153]. Furthermore, federated learning also offers a promising solution for safeguarding user privacy in SDR by allowing data

TABLE V
PROGRESS OF MOST RELEVANT WORKS AGAINST MAJOR CHALLENGES

Major Challenges	Key Issues	Methods to Address the Key Issues
MC1 - Size, balance and diversity of the datasets:	**KI1: Dataset Size**	- Data augmentation using GAN to generate synthetic spectrograms for the minority class [143]. - The SpecAugment data augmentation method may be employed for end-to-end ASR tasks with Listen, Attend, and Spell (LAS) networks, including warping log mel spectrogram features, masking blocks of frequency channels, and masking blocks of time steps [137]. - Cropping audio into fixed and same-size samples, increasing the number of samples extracted from an audio signal by reusing parts of a signal in a new instance [79], [80].
	KI2: Dataset Diversity	- Investigating the effect of specific features in detecting depression in different genders. In [138], glottal pulse duration was found to be the most effective parameter for discriminating between depressed and non-depressed speech for males, whereas the fundamental frequency F0 was the most determinant for females. - Developing a gender-based model, *i.e.*, considering gender as a variable for SDR [139]. - Proposing a modeling paradigm for SDR that takes age into account as in [140].
	KI3: Data Biases	- In signal processing, the Non-Negative Matrix Factorization (NMF) is a commonly used for decomposing spectrograms into two matrices: one representing frequency patterns and another representing how those patterns combine over time [159]. This allows NMF to separate useful signals from noise or unwanted variations such as accents, enabling the model to focus on the features that are more relevant to MDD recognition [41].
MC2 - Speech recording procedure	**KI4: Patient's Speech Recording Location**	- Use "in the wild" emotion data and consider it as a meta-feature for mood state monitoring [34]. - Analyzing speech as a sequence of acoustic events, where regions of acoustic space are tokenized into "words" representing speech events at fixed or irregular intervals. This tokenization enables NLP methods to utilize acoustic word features and achieve good results on both laboratory-grade audio recordings and smartphone recordings [144].
	KI5: Data Noise	- The speaker diarization can be employed to extract speech segments from the patients, rather than the interviewers. A set of acoustic features and deep representations are then extracted from these segments. The representations obtained are used to create a Bag-of-Features, which is a fixed-length histogram representation of each audio recording. This quantization step has been demonstrated to be an effective means of filtering small amounts of noise present in the original feature space [145]. - The Time Delay Neural Network-based Denoising Autoencoder (TDNN-DAE) is a viable solution for estimating clean MFCCs from noisy MFCCs. [142]. - Voice Activity Detection (VAD) is a technique that can be used to segment portions of a speech audio signal, discarding segments of silence or background noise that could affect the extraction of relevant features for depression analysis [35], [141].
	KI6: Type of Speech to Record	- In [50], the authors modeled males and females separately in the classifier, providing different weights for different speech types (interview, picture description, and reading) and emotions according to their respective contributions in detecting depression. - In [146], the authors used a multiple classifier method using different speech types and speech emotions. Their findings indicate that the recognition rate for depression is higher when using interview speech compared to picture description speech and reading speech.
MC3 - Privacy and Ethics constraints	**KI7: Identity**	- Use a speaker-independent cross-validation technique [38], which involves applying the group shuffle split technique to ensure that the same groups of data don't appear in both the training and test sets. Additionally, samples from both low and high MDD classes per speaker are included in the training process. As stated in [38], this approach ensures that the model learns patterns related to MDD symptoms, rather than speaker identity. - A speaker disentanglement method that utilizes a non-uniform mechanism to maximize the adversarial speaker identification (SID) loss. This is achieved by varying the adversarial weight between different layers of a model during training [31].
	KI8: Security	- Explaining the study procedures to all participants. Ensuring that they provide consent for the use of data for the study [28]. However, recording sensitive information, such as that shared in a clinical interview, can be risky. Therefore, according to [25], two approaches can be suggested: (i) Filter recordings in real time or use bone conduction microphones [147] or (ii) Capture audio on the participant's device, extract encrypted acoustic features from which the raw audio cannot be reconstructed, and then send the features to a secure server to download later [148].
MC4 - AI application for SDR	**KI9: Usability and Reliability**	- Speech synthesis is a text-to-speech (TTS) technology that can improve the usability of wearable health trackers [149]. WaveNet is an auto-regressive generative TTS model that uses linguistic features to generate relatively realistic human-like speech. However, it has the disadvantage of a long inference time due to its auto-regressive architecture. To overcome this problem, the FFT-Net, WaveRNN, and WaveGlow models have been proposed [73]. - The EU's AI Act sets out rules for the development of AI in the European Union [39]. - The test-retest reliability of acoustic and linguistic features shows how consistent test results are over time. It helps to understand the source of acoustic variability and how to make it more reliable, thus ensuring measurement accuracy during speech assessment practice [160].

to remain on local devices while enabling collaborative model training. This method enhances data security since only model updates are shared, not raw data [73], [154].

Although electronic health record (EHR) based systems have shown a great improvement lately, their usability with any input modality is an area that requires continuous development. In particular, the addition of a speech recognition (SR) component to an EHR system can result in a significant reduction in perceived usability by clinicians, as pointed out by [150]. This is largely due to significantly lower overall usability scores when using SR, as it imposes greater costs in terms of learnability through training and support for EHR-based documentation compared to, for example, using a keyboard or mouse.

The use of multiple depression datasets for generalization represents a challenging domain of inquiry, largely due to the considerable discrepancies in recording environments, recording procedures and depression assessment. Furthermore, there are a number of ethical, clinical and legal considerations regarding the acquisition and sharing of such datasets [182]. Consequently, the number of cross-cultural studies remains limited, with only a few studies currently aiming to develop an adaptive depression engine capable of responding to different populations [182]–[184].

For example, the work of [182] investigates the generalizability of a cross-cultural depression detection model that extracts nonverbal temporal patterns of depression to cross-cultural datasets. They concluded that when the classifiers are trained on varied observations, they have a stronger ability to generalize to new observations than when they are trained on observations with less variability.

However, these studies still raise concerns that the differences in the recording environments could affect the classification results [182], and some studies have shown that there are downgrades in acoustic emotion recognition when crossing populations [41], [42]. Therefore, a key step towards breakthroughs in depression analysis is the creation of a large cross-cultural database with open standards, accurate and consistent labeling [1]. Thus, as highlighted by [185], future research should focus on conducting more studies using multiple datasets and sharing their extracted features, as well as the code used for training and evaluating the models. They also advise to consolidate the data, code, and environment into a single package that can be easily redistributed, for example using containers like Dockers. This strategy will enhance the generalizability of models and assist in resolving discrepancies regarding crucial and predictive features. Furthermore, the integration of multi-modal features in machine learning models holds potential for improving mental health assessment and treatment.

To date, there has been limited research on using speech signals to assess different types and severity levels of depression. In addition, the assessment of emotion of speaker groups has hardly been addressed [36]. Most of the works are focused only in the binary classification of MDD [38].

However, the binary MDD prediction through a questionnaire sum-score has limitations due to the highly heterogeneous nature of this disorder [163]. Therefore, as pointed out by [163], an approach that could better benefit patients and open up new possibilities for early depression prevention and intervention is to investigate the symptoms of MDD. This would involve, firstly, the study of the MDD symptoms individually to understand how they differ from each other in terms of their impact on functioning, their response to specific life events, their relations with biological markers, and their risk factors [165]. Second, to examine associations between symptoms, as causal influences between them are ignored when examining sum scores. For instance, in the work of [166], the authors present a Bayesian framework to capture the relationships between depression symptoms, and features derived from speech, facial expression and cognitive game data. This multi-modal approach yielded good results that were not subject to demographic biases. Finally, investment in personalised medicine may also be an approach to overcome the limitations of the highly heterogeneous depression phenotype [163].

In terms of computational models, future trends in the field of depression detection have shifted from traditional ML models to DL architectures, with many recent works exploring the use of transformers and parallel convolutional neural networks [43]–[45], multimodal fusion models [7], [46], [47], GANs [128], additive cross-modal attention networks [49] and transfer learning [171], [172]. The way adversarial attacks can influence the performance of such models still requires attention [173].

Although latest works have shown high performance, the experimental protocol should thoroughly assess the statistical significance of the obtained results. To this end, statistical tests can be used, such as a paired t-test or a permutation test to increase confidence in the generalizability of the results, or even a bootstrapping procedure [36]. Furthermore, while these studies demonstrate the potential for MDD assessment, their application in cases with comorbidities and overlapping symptoms presents a significant challenge [175]. This is particularly evident during depressive episodes, in which bipolar disorder (BD) and MDD may be difficult to distinguish [176]. In patients with MDD, models trained solely on binary MDD classifications might misclassify these other conditions as MDD, thereby reducing their reliability [175], [177]. As misdiagnosis may lead to delays in effective treatment and to exposure to ineffective treatment, these models need further development and validation to be safely deployed in clinical practice [175].

A lot of effort have been made into avoiding data biases in SDR (KI3); however, there are still open issues. To prevent model biases, analytical validation can be achieved through cross-validation, independent replication of results, and testing the model's generalizability to new independent datasets [71]. It is important to ensure that factors such as age, gender, education level, and accent are evenly distributed in both the training and testing datasets [139], [140]. At the same time, future research needs to follow the open science principles including standardized experimental protocols.

[162] has put together a set of recommendations for avoiding bias in explanatory ML models using audio recordings. These include: (i) formulating hypotheses about the

importance of predictor variables; (ii) implementing quality control measures for controlled and remote recording settings; (iii) collecting representative samples of the different groups; (iv) providing instructions to patients for the recording procedure; (v) removing non-natural outliers; (vi) being cautious with preprocessing steps that may affect the properties associated with the disorder; (vii) training multiple ML models of different complexity; (viii) performing explainability analyses, including variable dependency, feature robustness, potential biases, expert ratings; (ix) using bias mitigation strategies; (x) continuous assessment; and other techniques.

As previously stated, the WHO has indicated that MDD rates are higher in women than in men. Additionally, we have previously emphasized the importance of investigating the effect of specific features in detecting depression in different genders and age groups to avoid biases and increase the model's generalizability capability. However, research considering female factors such as menstruation and hormones across various phases of the menstrual cycle remains limited.

The study of [186] employs a ML model to predict depressive symptoms in menopausal women, as the prevalence of depression during menopause represents a significant public health concern, with reported incidences ranging from 5.9% to 23.8% [187]. Although the work presented promising results, underscoring the significance of incorporating various factors, such as hormonal levels, lifestyle, and demographic variables, when addressing depression during menopause, reliance on self-reported data in the study introduces potential biases. Furthermore, the applicability of the findings might be restricted due to the specific demographic focus.

Depression self-assessment mobile applications have become popular in the last decade. The studies of [155]–[158] present several such depression detection tools. Although they provide a less costly and quick mental health assessment service when compared to professional services [155], only a few applications currently offer comprehensive programs that can be proven to benefit users [157]. In addition, most of them share users' data with third parties, which raises serious concerns about user privacy. These apps should complement an ongoing patient-provider therapeutic relationship and not replace professional monitoring [156], while following evidence-based clinical guidelines, be patient-centered, and enhance user privacy [157].

As shown here, SDR is a fertile field of research with many challenges as well as opportunities. While progress has been made in the literature, SDR systems still have a long way to go before they can be safely and effectively used in real-world settings. Certain factors, such as age, gender, and speech type, can impact the depression assessment. Thus, analyzing these factors in SDR could lead to improved assessment accuracy and a reduction in the significant socioeconomic impact associated with this disorder [15], [51].

VI. Conclusion

This paper presented a comprehensive survey of the latest improvements, existing open issues, and challenges in SDR. The concept of depression and its traditional clinical approaches were presented (SO1) as well as the most recent surveys in the field and their limitations, particularly in addressing the major challenges and key issues faced by most researches (SO2). Based on our research findings, the analysis of speech acoustic features could potentially serve as an objective marker for the identification of MDD [50]. This could lead to improved accuracy of assessment and a reduction in the significant socioeconomic impact associated with this condition [51].

Importantly, we presented the state-of-the-art approaches while organising the existing major challenges and key issues (S04). A comparison between latest scientific works' performance using hand-crafted and/or end-to-end deep architectures to predict depression were presented (SO3), showing that recent works have shifted towards DL approaches, particularly utilizing technologies such as transformers, GANs, attention mechanisms, and multi-modal model fusion. In addition, this article offers researchers good practices and techniques fostering discussions and improvements in the field. Last but not least, we highlighted several concerns, open issues and future directions for developing SDR solutions (SO5).

Acknowledgments

Part of this research was supported by the project SOLITAIRE - "Digital interventions for Social isOLation In youThs And theIR familiEs", funded by the European Union - Next Generation EU - NRRP M6C2 - Investment 2.1 "Enhancement and strengthening of biomedical research in the NHS". Project number: PNRR-MAD-2022-12376834, CUP: G43C22004010006.

We would like to thank the SOLITAIRE group for the contribution (Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona, Verona, Italy & Azienda Ospedaliera Universitaria Integrata Verona, Verona, Italy: Mirella Ruggeri, Marcella Bellani, Maria Gloria Rossetti, Cinzia Perlini, Francesca Girelli, Niccolò Zovetti, Maria Diletta Buio; Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy: Paolo Brambilla, Cinzia Bressi, Antonella delle Fave, Virginia Pupi; CNR Institute of Neuroscience, Vedano al Lambro, Italy: Fabrizia Guarnieri, Edoardo Moretto; Department of Computer Science, University of Milan, Milan, Italy: Roberto Sassi, Maria Renata Guarneri, Stavros Ntalampiras, Samara Soares Leal; Unit for Severe Disabilities in Developmental Age and Young Adults, Associazione La Nostra Famiglia - IRCCS E. Medea, Scientific Hospital for Neurorehabilitation, Brindisi, Italy: Isabella Fanizza, Lara Scialpi, Giorgia Carlucci, Mariangela Leucci; Scientific Institute IRCCS Eugenio Medea, Scientific Direction, Bosisio Parini, Lecco, Italy: Antonio Trabacca).

References

[1] P. Wu *et al.,* "*Automatic depression recognition by intelligent speech signal processing: A systematic survey*", CAAI Trans on Intel Tech, vol. 8, no. 3, pp. 701–711, Sep. 2023, doi: 10.1049/cit2.12113.

[2] Z. N. Vasha *et al.,* "*Depression detection in social media comments data using machine learning algorithms*", Bulletin EEI, vol. 12, no. 2, pp. 987–996, Apr. 2023, doi: 10.11591/eei.v12i2.4182.

[3] M. Briley and Lépine, "*The increasing burden of depression*", NDT, p. 3, May 2011, doi: 10.2147/NDT.S19617.

[4] U.S. '*Department of Health and Human Services., Healthy people 2010 : understanding and improving health*", Washington, DC: U.S. Dept. Health Hum. Serv., 2000.

[5] A. P, "*Association, Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*", WA, D.C., USA: American Psychiatric Pub, 2013.

[6] S. Graham *et al.,* "*Artificial intelligence for mental health and mental illnesses: An overview*", Curr. Psychiatry Rep., vol. 21, no. 11, pp. 1–18, 2019, doi: 10.1007/s11920-019-1094-0.

[7] B. Zou *et al.,* "*Semi-Structural Interview-Based Chinese Multimodal Depression Corpus Towards Automatic Preliminary Screening of Depressive Disorders*", IEEE Trans. Affective Comput., vol. 14, no. 4, pp. 2823–2838, Oct. 2023, doi: 10.1109/TAFFC.2022.3181210.

[8] W. Xie *et al.,* "*Interpreting depression from question-wise long-term video recording of SDS evaluation*", IEEE J. Biomed. Health Inform., vol. 26, no. 2, pp. 865–875, Feb. 2022, doi: 10.1109/jbhi.2021.3092628.

[9] J. Yoon *et al.,* "*D-vlog: Multimodal Vlog Dataset for Depression Detection*", AAAI, vol. 36, no. 11, pp. 12226–12234, Jun. 2022, doi: 10.1609/aaai.v36i11.21483.

[10] H. Dinkel*et al.,* "*Text-based depression detection: What triggers an alert*", 2019, arXiv:1904.05154.

[11] D. William and D. Suhartono, "*Text-based depression detection on social media posts: A systematic literature review*", Procedia Computer Science, v. 179, p. 582-589, 2021, doi: 10.1016/j.procs.2021.01.043.

[12] S. Ntalampiras, "*Speech emotion recognition via learning analogies*", Pattern Recognition Letters, vol. 144, p. 21–26, 2021, doi: 10.1016/j.patrec.2021.01.018.

[13] P. Zhang *et al.,* "*DEPA: Self-Supervised Audio Embedding for Depression Detection*", Pingyue, New York, NY, USA: Association for Computing Machinery, p. 135–143, 2021, doi: 10.1145/3474085.3479236.

[14] M. Gerczuk *et al.,* "*Personalised deep learning for monitoring depressed mood from speech*", in Proceedings of the E- Health and Bioengineering Conference (EHB). Ias, i, Romania: IEEE, 2022, pp. 1–5.

[15] C. W. Espinola *et al.,* "*Detection of major depressive disorder using vocal acoustic analysis and machine learning — an exploratory study*", Research on Biomedical Engineering, v. 37, p. 53-64, 2021, doi: 10.1007/s42600-020-00100-9.

[16] S. Ntalampiras, "*Model Ensemble for Predicting Heart and Respiration Rate From Speech*", in IEEE Internet Computing, vol. 27, no. 3, pp. 15-20, May-June 2023, doi: 10.1109/MIC.2023.3257862.

[17] A. Khosla *et al.,* "*Automated diagnosis of depression from EEG signals using traditional and deep learning approaches: A comparative analysis*", Biocybern. Biomed. Eng., vol. 42, no. 1, pp. 108-142, Jan. 2022, doi: 10.1016/j.bbe.2021.12.005.

[18] B. Zhang *et al.,* "*Brain functional networks based on resting-state EEG data for major depressive disorder analysis and classification*", IEEE Trans. Neural Syst. Rehabil. Eng., vol. 29, pp. 215-229, 2021, doi: 10.1109/tnsre.2020.3043426.

[19] G. Lam, H. Dongyan, and W. Lin, "*Context-aware deep learning for multi-modal depression detection*", in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 3946–3950, doi: 10.1109/ICASSP.2019.8683027.

[20] H. Cai *et al.,* "*MODMA dataset: A multi-model open dataset for mental disorder analysis background and summary*", 2020, arXiv:2002.09283.

[21] N. V. Babu, E. Kanaga and M. Grace, "*Sentiment analysis in social media data for depression detection using artificial intelligence: a review*", SN Computer Science, v. 3, p. 1-20, 2022, doi: 10.1007/s42979-021-00958-1.

[22] M. F. Alghifari, S. G. Teddy and K. Mira, "*Development of Sorrow Analysis Dataset for Speech Depression Prediction.*", in 2023 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), 01–06. Kuala Lumpur, Malaysia: IEEE, 2023, doi: 10.1109/I2MTC53148.2023.10176040.

[23] N. Aloshban, A. Esposito, and A. Vinciarelli, "*What You Say or How You Say It? Depression Detection Through Joint Modeling of Linguistic and Acoustic Aspects of Speech*", Cogn Comput, vol. 14, no. 5, pp. 1585–1598, Sep. 2022, doi: 10.1007/s12559-020-09808-3.

[24] I. Mantegazza and S. Ntalampiras, "*Italian Speech Emotion Recognition*", 2023 24th International Conference on Digital Signal Processing (DSP), Rhodes (Rodos), Greece, 2023, pp. 1-5, doi: 10.1109/DSP58604.2023.10167766.

[25] D. M. Low *et al.,* "*Automated Assessment of Psychiatric Disorders Using Speech: A Systematic Review*", Laryngoscope Investigative Otolaryngol., vol. 5, no. 1, pp. 96–116, Jan. 2020, doi: 10.1002/lio2.354.

[26] P. A. Babu, V. Siva Nagaraju and R. R. Vallabhuni, '*Speech Emotion Recognition System With Librosa*", 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2021, pp. 421-424, doi: 10.1109/CSNT51715.2021.9509714.

[27] Or, F., J. Torous, and J.-P. Onnela, "*High potential but limited evidence: Using voice data from smartphones to monitor and diagnose mood disorders*", Psychiatric Rehabilitation Journal, vol. 40, no. 3, pp. 320–324, 2017, doi: 10.1037/prj0000279.

[28] A. Y. Kim *et al.,* "*Automatic Depression Detection Using Smartphone-Based Text-Dependent Speech Signals: Deep Convolutional Neural Network Approach*", Journal of Medical Internet Research 25, e34474, 2023, doi: 10.2196/34474.

[29] G. Kiss and K. Vicsi, "*Comparison of Read and Spontaneous Speech in Case of Automatic Detection of Depression*", in 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Debrecen: IEEE, Sep. 2017, pp. 000213–000218. doi: 10.1109/CogInfoCom.2017.8268245.

[30] M. Tasnim, M. Ehghaghi, B. Diep, and J. Novikova, "*DEPAC: A Corpus for Depression and Anxiety Detection from Speech*", In Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, 1–16. Seattle, USA: Association for Computational Linguistics, 2022, doi: 10.18653/v1/2022.clpsych-1.1.

[31] J. Wang, V. Ravi, and A. Alwan, *Non-Uniform Speaker Disentanglement For Depression Detection From Raw Speech Signals*, arXiv, 2023.

[32] S. D. Lustgarten *et al., Digital privacy in mental healthcare: current issues and recommendations for technology use*, Current Opinion Psychol., vol. 36, pp. 25–31, Dec. 2020, doi: 10.1016/j.copsyc.2020.03.012

[33] J. Gratch *et al.,* "*The distress analysis interview corpus of human and computer interviews*", In: LREC, p. 3123-3128, 2014.

[34] S. Khorram *et al.,* "*The PRIORI Emotion Dataset: Linking Mood to Emotion Detected In-the-Wild*", in Interspeech 2018, ISCA, Sep. 2018, pp. 1903–1907. doi: 10.21437/Interspeech.2018-2355.

[35] Z. Huang *et al.,*, "*Depression Detection from Short Utterances via Diverse Smartphones in Natural Environmental Conditions*", in Interspeech 2018, ISCA, Sep. 2018, pp. 3393–3397. doi: 10.21437/Interspeech.2018-1743.

[36] L. Verde et al., "*A Lightweight Machine Learning Approach to Detect Depression from Speech Analysis*", in 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), Washington, DC, USA: IEEE, Nov. 2021, pp. 330–335. doi: 10.1109/ICTAI52525.2021.00054.

[37] M. Nazar, *et al., "A Systematic Review of Human–Computer Interaction and Explainable Artificial Intelligence in Healthcare With Artificial Intelligence Techniques*", in IEEE Access, vol. 9, pp. 153316-153348, 2021, doi: 10.1109/ACCESS.2021.3127881.

[38] E. L. Campbell *et al.,* "*Classifying depression symptom severity: Assessment of speech representations in personalized and generalized machine learning models*", in INTERSPEECH 2023, ISCA, Aug. 2023, pp. 1738–1742. doi: 10.21437/Interspeech.2023-1721.

[39] European Parliament. Directorate-General for Research., Ed., "*Parliamentary control of community finances*", 3rd ed. Luxembourg: Office Official Publications Eur. Communities, 1988.

[40] G. Kiss *et al.,* "*Language Independent Detection Possibilities of Depression by Speech*", in Recent Advances in Nonlinear Speech Processing. Cham: Springer Int. Publishing, 2016, pp. 103–114.

[41] B. W. Schuller, "*Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends*", Commun. ACM, vol. 61, no. 5, pp. 90–99, Apr. 2018, doi: 10.1145/3129340.

[42] S. Ntalampiras, "*Toward language-agnostic speech emotion recognition*", Journal of the Audio Engineering Society, 68(1/2), 7-13, 2020.

[43] F. Yin, J. Du, X. Xu, and L. Zhao, *Depression Detection in Speech Using Transformer and Parallel Convolutional Neural Networks*, Electronics, vol. 12, nº 2, p. 328, jan. 2023, doi: 10.3390/electronics12020328.

[44] H. Sun, Y.-W. Chen, and L. Lin, "*TensorFormer: A Tensor-Based Multimodal Transformer for Multimodal Sentiment Analysis and Depression Detection*", IEEE Trans. Affective Comput., vol. 14, nº 4, p. 2776–2786, out. 2023, doi: 10.1109/TAFFC.2022.3233070.

[45] W. Zheng, L. Yan, and F. Y. Wang, "*Two Birds With One Stone: Knowledge-Embedded Temporal Convolutional Transformer for Depression Detection and Emotion Recognition*", IEEE Trans. Affective Comput., vol. 14, nº 4, p. 2595–2613, out. 2023, doi: 10.1109/TAFFC.2023.3282704.

[46] M. Niu *et al.,* "*Multimodal Spatiotemporal Representation for Automatic Depression Level Detection*", IEEE Trans. Affective Comput., vol. 14, nº 1, p. 294–307, jan. 2023, doi: 10.1109/TAFFC.2020.3031345.

[47] L. He *et al.,* "*Deep learning for depression recognition with audiovisual cues: A review*", Inf. Fusion, vol. 80, pp. 56–86, 2022.

[48] S. Ntalampiras, "*Deep Learning of Attitude in Children's Emotional Speech*" 2020 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications, Tunis, Tunisia, 2020, pp. 1-5, doi: 10.1109/CIVEMSA48639.2020.9132743.

[49] N. K. Iyortsuun *et al.,* "*Additive Cross-Modal Attention Network (ACMA) for Depression Detection Based on Audio and Textual Features*", in IEEE Access, vol. 12, pp. 20479-20489, 2024, doi: 10.1109/ACCESS.2024.3362233

[50] H. Jiang *et al.,* "*Investigation of different speech types and emotions for detecting depression using different classifiers*", Speech Communication, vol. 90, pp. 39–46, 2017.

[51] N. Cummins *et al.,* "*A review of depression and suicide risk assessment using speech analysis*", Speech Communication, vol. 71, pp. 10–49, 2015.

[52] World Health Organization, "*Depressive disorder (depression)*", https://www.who.int/news-room/fact-sheets/detail/depression, 2023, Accessed in Feb 2024.

[53] W. Depression, "*Other common mental disorders: global health estimates*", Geneva: World Health Organization, pp. 1–24, 2017.

[54] A. American Psychiatric Association and A. P. Association, "*Diagnostic and statistical manual of mental disorders: Dsm-5,*", 2013.

[55] World Health Organization, "*Depression and Other Common Mental Disorders: Global Health Estimates*", 2017.

[56] A. H. Weinberger *et al.,* "*Trends in depression prevalence in the USA from 2005 to 2015: widening disparities in vulnerable groups*", Psychological medicine, v. 48, n. 8, p. 1308-1315, 2018.

[57] N. W. Hashim *et al.,* "*Evaluation of voice acoustics as predictors of clinical depression scores*", J. Voice, vol. 31, no. 2, pp. 256.e1–256.e6, Mar. 2017. doi: 10.1016/j.jvoice.2016.06.006.

[58] B. Arroll et al., "*Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population*", Ann. Family Medicine, vol. 8, no. 4, pp. 348–353, Jul. 2010, doi: 10.1370/afm.1139.

[59] T. Al Hanai, M. Ghassemi, and J. Glass, "*Detecting Depression with Audio/Text Sequence Modeling of Interviews*", in Interspeech 2018, ISCA, Sep. 2018, pp. 1716–1720. doi: 10.21437/Interspeech.2018-2522.

[60] M. Hamilton, *A RATING SCALE FOR DEPRESSION*, Journal of Neurology, Neurosurgery, and Psychiatry, vol. 23, no. 1, pp. 56–62, Feb. 1960.

[61] M. Ghisi *et al.,* "*Beck depression inventory-ii*", Manuale italiano. Firenze: Organizzazioni Speciali, 2006.

[62] A. J. *Rush et al.,* "*The 16-Item quick inventory of depressive symptomatology, clinician rating, and self-report: A psychometric evaluation in patients with chronic major depression*", Biol. Psychiatry, vol. 54, no. 5, pp. 573–583, Sep. 2003, doi: 10.1016/s0006-3223(02)01866-8.

[63] R. Young, *et al.,* "*A rating scale for mania: reliability, validity and sensitivity*", Brit. J. Psychiatry, vol. 133, no. 5, pp. 429–435, Nov. 1978, doi: 10.1192/bjp.133.5.429

[64] S. A. Montgomery and M. Åsberg. "*A new depression scale designed to be sensitive to change*", Brit. J. Psychiatry, vol. 134, no. 4, pp. 382–389, Apr. 1979, doi: 10.1192/bjp.134.4.382.

[65] W. W. K. Zung, "*A self-rating depression scale*", Arch. Gen. Psychiatry, vol. 12, no. 1, pp. 63–70, 1965.

[66] L. He and C. Cao, "*Automated depression analysis using convolutional neural networks from speech*", J. Biomed. Inform., vol. 83, pp. 103–111, Jul. 2018, doi: 10.1016/j.jbi.2018.05.007.

[67] R. Alsarrani, A. Esposito, and A. Vinciarelli, "*Thin Slices of Depression: Improving Depression Detection Performance Through Data Segmentation*", in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore: IEEE, May 2022, pp. 6257–6261. doi: 10.1109/ICASSP43922.2022.9746426.

[68] J. Huang *et al.,* "*Multimodal continuous emotion recognition with data augmentation using recurrent neural networks*", in MM '18: ACM Multimedia Conf., Seoul Republic of Korea. New York, NY, USA: ACM, 2018, doi: 10.1145/3266302.3266304.

[69] F. Ringeval *et al.,* "*Avec 2017: Real-life depression, and affect recognition workshop and challenge. In: Proceedings of the 7th annual workshop on audio/visual emotion challenge*", 2017. p. 3-9.

[70] Y. Li *et al.,* "*Speech databases for mental disorders: A systematic review*", Gen Psych, vol. 32, no. 3, p. e100022, Jul. 2019, doi: 10.1136/gpsych-2018-100022.

[71] J. Robin *et al.,* "*Evaluation of Speech-Based Digital Biomarkers: Review and Recommendations*", Digit. Biomarkers, vol. 4, no. 3, pp. 99–108, Oct. 2020, doi: 10.1159/000510820.

[72] J. C. Goldsack *et al.,* "*Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs)*", npj Digit. Medicine, vol. 3, no. 1, Apr. 2020, doi: 10.1038/s41746-020-0260-4.

[73] S. Latif *et al.,* "*Speech Technology for Healthcare: Opportunities, Challenges, and State of the Art*", IEEE Rev. Biomed. Eng., p. 1, 2020, doi: 10.1109/rbme.2020.3006860.

[74] S. Ntalampiras, "*Language-agnostic speech anger identification,*", 2021 44th International Conference on Telecommunications and Signal Processing (TSP), Brno, Czech Republic, 2021, pp. 249-253, doi: 10.1109/TSP52935.2021.9522606.

[75] L. He *et al.,* "*Deep learning for depression recognition with audiovisual cues: A review*", Inf. Fusion, vol. 80, pp. 56–86, Apr. 2022, doi: 10.1016/j.inffus.2021.10.012.

[76] M. Maithri *et al.,* "*Automated emotion recognition: Current trends and future perspectives*", Comput. Methods Programs Biomedicine, vol. 215, p. 106646, Mar. 2022, doi: 10.1016/j.cmpb.2022.106646.

[77] A. P. Vogel *et al.,* "*Comparability of modern recording devices for speech analysis: smartphone, landline, laptop, and hard disc recorder*", Folia Phoniatr. Logop., vol. 66, no. 6, pp. 244–250, 2014, doi: 10.1159/000368227

[78] A. Vabalas *et al.,* "*Machine learning algorithm validation with a limited sample size*", PLOS ONE, vol. 14, no. 11, Nov. 2019, doi: 10.1371/journal.pone.0224365.

[79] S. Sardari *et al.,* "*Audio based depression detection using Convolutional Autoencoder*", Expert Syst. with Appl., vol. 189, p. 116076, Mar. 2022, doi: 10.1016/j.eswa.2021.116076.

[80] A. Vázquez-Romero and A. Gallardo-Antolín, "*Automatic detection of depression in speech using ensemble convolutional neural networks*", Entropy, vol. 22, no. 6, p. 688, Jun. 2020, doi: 10.3390/e22060688.

[81] J. C. Mundt *et al.,* "*Vocal acoustic biomarkers of depression severity and treatment response*", Biol. Psychiatry, vol. 72, no. 7, pp. 580–587, Oct. 2012, doi: 10.1016/j.biopsych.2012.03.015.

[82] B. Nakisa *et al.,* "*Automatic Emotion Recognition Using Temporal Multimodal Deep Learning*", in IEEE Access, vol. 8, pp. 225463-225474, 2020, doi: 10.1109/ACCESS.2020.3027026.

[83] S. K. Gaikwad *et al.,* "*A review on speech recognition technique*", Int. J. Comput. Appl., vol. 10, no. 3, pp. 16–24, Nov. 2010, doi: 10.5120/1462-1976.

[84] X. Ma *et al.,* "*DepAudioNet: an efficient deep model for audio based depression classification*", in:Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, 35–42, ACM, 2016.

[85] D. Bzdok and A. Meyer-Lindenberg, "*Machine learning for precision psychiatry: opportunities and challenges*", Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, v. 3, n. 3, p. 223-230, 2018.

[86] N. Seneviratne and C. Espy-Wilson, "*Speech based depression severity level classification using a multi-stage dilated CNN-LSTM model*", arXiv preprint arXiv:2104.04195, 2021.

[87] T. Baltrušaitis *et al.,* "*Multimodal Machine Learning: A Survey and Taxonomy*", in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423-443, 1 Feb. 2019, doi: 10.1109/TPAMI.2018.2798607.

[88] J. Gratch *et al.,* "*The distress analysis interview corpus of human and computer interviews*", In: LREC. 2014. p. 3123-3128.

[89] Y. Shen, H. Yang, and L. Lin, "*Automatic Depression Detection: an Emotional Audio-Textual Corpus and A Gru/Bilstm-Based Model*", in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore: IEEE, May 2022, pp. 6247–6251. doi: 10.1109/ICASSP43922.2022.9746569.

[90] M. Valstar *et al.,* "*Avec 2016: Depression, mood, and emotion recognition workshop and challenge*ii, in: Proceedings of the 6th international workshop on audio/visual emotion challenge. 2016. p. 3-10.

[91] H. Sun *et al.,* "*Multi-modal adaptive fusion transformer network for the estimation of depression level*", Sensors, vol. 21, no. 14, p. 4764, Jul. 2021, doi: 10.3390/s21144764.

[92] J. C. Mundt e*t al.,* "*Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology*", J. Neurolinguistics, vol. 20, no. 1, pp. 50–64, Jan. 2007, doi: 10.1016/j.jneuroling.2006.04.001.

[93] M. Valstar *et al.,* "*AVEC 2014: 3d dimensional affect and depression recognition challenge*", in: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, pp.3–10. ACM, 2014.

[94] D. DeVault *et al.,* "*Sim Sensei Kiosk: a virtual human interviewer for healthcare decision support*", in: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems, 1061–1068. IFAAMAS, 2014.

[95] D. Zhou *et al.,* "*Tackling mental health by integrating unobtrusive multimodal sensing*", Proc. AAAI Conf. Artif. Intell., vol. 29, no. 1, Feb. 2015, doi: 10.1609/aaai.v29i1.9381.

[96] K.-Y. Huang *et al.,* "*Unipolar depression vs. Bipolar disorder: An elicitation-based approach to short-term detection of mood disorder*", in Interspeech 2016. ISCA, 2016, doi: 10.21437/interspeech.2016-620.

[97] E. Ciftci *et al.,* "*The turkish audio-visual bipolar disorder corpus*", in 2018 First Asian Conf. Affect. Comput. Intell. Interaction (ACII Asia), Beijing, May 20–22, 2018. IEEE, 2018, doi: 10.1109/aciia-sia.2018.8470362.

[98] H. Dibeklioglu *et al.,* "*Dynamic multimodal measurement of depression severity using deep autoencoding*", IEEE J. Biomed. Health Inform., vol. 22, no. 2, pp. 525–536, Mar. 2018, doi: 10.1109/jbhi.2017.2676878.

[99] H. Cai e*t al.,* "*MODMA dataset: a multi-modal open dataset for mental disorder analysis*", arXiv preprint arXiv:2002.09283, 2020.

[100] D. Reynolds *et al.,* "*The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition*", in IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 4. Hong Kong, China: IEEE, 2003, pp. 784–787, doi: 10.1109/icassp.2003.1202760.

[101] D. Reynolds *et al.,* "*Beyond cepstra: exploiting high-level information in speaker recognition*", in Workshop on Multimodal User Authentication, Santa Barbara, CA, 2003, pp. 223–229.

[102] F. Ringeval *et al.,* "*Avec 2017: Real-life depression, and affect recognition workshop and challenge*", Proceedings of the 7th annual workshop on audio/visual emotion challenge, 2017.

[103] P. Lopez-Otero *et al.,* "*Assessing speaker independence on a speech-based depression level estimation system*", Pattern Recogn. Lett. 68,343–350, 2015, doi: 10.1016/j.patrec.2015.05.017.

[104] S. Ntalampiras, "*A transfer learning framework for predicting the emotional content of generalized sound events*", The Journal of the Acoustical Society of America 141(3),1694–1701, 2017, doi: 10.1121/1.4977749.

[105] H. Jiang *et al.,* "*Detecting depression using an ensemble logistic regression model based on multiple speech features*", Comput. Math. Method. M.2018, 1–9, 2018, doi: 10.1155/2018/6508319.

[106] R. Hasan *et al.,* "*Speaker identification using Mel frequency cepstral coefficients*", in: 3rd International Conference on Electrical and Computer Engineering ICECE 2004, p. 565–8, 2004.

[107] A. Maxhuni *et al.,* "*Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients*", Pervasive Mobile Comput., vol. 31, pp. 50–66, Sep. 2016, doi: 0.1016/j.pmcj.2016.01.008.

[108] F. Eyben *et al.,* "*Recent developments in opensmile, the munich open-source multimedia feature extractor*", Proceedings of the 21st ACM International Conference on Multimedia, ACM, 2013, pp. 835–838.

[109] F. Simonetta, F. Certo, and S. Ntalampiras "*Joint Learning of Emotions in Music and Generalized Sounds*", In Proceedings of the 19th International Audio Mostly Conference: Explorations in Sonic Cultures. ACM, New York, NY, USA, 302–307, https://doi.org/10.1145/3678299.3678328.

[110] C. Busso, S. Lee, and S. S. Narayanan, "*Using neutral speech models for emotional speech analysis*", in Eighth Annual Conference of the International Speech Communication Association, 2007.

[111] S. Scherer *et al.,* "*Investigating voice quality as a speaker-independent indicator of depression and PTSD*", in Interspeech 2013. ISCA: ISCA, 2013, doi: 10.21437/interspeech.2013-240.

[112] Y. Yang, C. Fairbairn, J.F. Cohn, "*Detecting depression severity from vocal prosody*", IEEE Trans. Affect. Comput., vol. 4, no. 2, pp. 142–150, Apr. 2013, doi: 10.1109/t-affc.2012.38.

[113] E. W. McGinnis *et al.,* "*Giving Voice to Vulnerable Children: Machine Learning Analysis of Speech Detects Anxiety and Depression in Early Childhood"*, in IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 6, pp. 2294-2301, Nov. 2019, doi: 10.1109/JBHI.2019.2913590.

[114] H. S. Alsagri and M. Ykhlef, "*Machine learning-based approach for depression detection in twitter using content and activity features*", IEICE Transactions on Information and Systems, vol. 103, no. 8, pp. 1825–1832, Aug. 2020, doi: 10.1587/transinf.2020EDP7023.

[115] Z. Liu *et al.,* "*A novel decision tree for depression recognition in speech"*, arXiv preprint arXiv:2002.12759, 2020.

[116] P. R. Parekh and M. M. Patil, "*Clinical depression detection for adolescent by speech features*", 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, India, 2017, pp. 3453-3457, doi: 10.1109/ICECDS.2017.8390102.

[117] L. Kerkeni *et al.,* "*Speech emotion recognition: methods and cases study*", ICAART 20(2) , 2018.

[118] H. Purwins *et al.,* "*Deep Learning for Audio Signal Processing," in IEEE Journal of Selected Topics in Signal Processing*", vol. 13, no. 2, pp. 206-219, May 2019, doi: 10.1109/JSTSP.2019.2908700.

[119] S. Jayawardena, J. Epps and E. Ambikairajah, "*Ordinal Logistic Regression With Partial Proportional Odds for Depression Prediction*", in IEEE Transactions on Affective Computing, vol. 14, no. 1, pp. 563-577, 1 Jan-March 2023, doi: 10.1109/TAFFC.2020.3031300.

[120] L. K. Xin and N. b. A. Rashid, "*Prediction of Depression among Women Using Random Oversampling and Random Forest*", 2021 International Conference of Women in Data Science at Taif University (WiDSTaif), Taif, Saudi Arabia, 2021, pp. 1-5, doi: 10.1109/WiDSTaif52235.2021.9430215.

[121] G. Saranya and A. Pravin, "*A comprehensive study on disease risk predictions in machine learning*", Int. J. Electr. Comput. Eng., vol. 10, no. 4, pp. 4217-4225, 2020.

[122] S. Sawangarreerak and P. Thanathamathee, "*Random forest with sampling techniques for handling imbalanced prediction of university student depression*", Information, vol. 11, no. 11, p. 519, Nov. 2020, doi: 10.3390/info11110519.

[123] M. Kumar *et al.,* "*Gray matter biomarkers for major depressive disorder and manic disorder using logistic regression*", J. Psychiatric Res., Jan. 2024, doi: 10.1016/j.jpsychires.2024.01.043.

[124] J. Zhao, X. Mao and L. Chen, "*Speech emotion recognition using deep 1D and 2D CNN LSTM networks*", Biomed. Signal Process. Control 47 (2019) 312–323, doi: 10.1016/j.bspc.2018.08.035.

[125] M. Muzammel *et al.,* "*AudVowelConsNet: A phoneme-level based deep CNN architecture for clinical depression diagnosis*", Mach. Learn. with Appl., vol. 2, p. 100005, Dec. 2020, doi: 10.1016/j.mlwa.2020.100005.

[126] Z. Huang, J. Epps and D. Joachim, "*Exploiting Vocal Tract Coordination Using Dilated CNNS For Depression Detection In Naturalistic Environments*", ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6549-6553, doi: 10.1109/ICASSP40776.2020.9054323.

[127] M. Niu *et al.,* "*A time-frequency channel attention and vectorization network for automatic depression level prediction*", Neurocomputing, vol. 450, pp. 208–218, Aug. 2021, doi: 10.1016/j.neucom.2021.04.056.

[128] L. Yang, D. Jiang and H. Sahli, "*Feature Augmenting Networks for Improving Depression Severity Estimation From Speech Signals*", in IEEE Access, vol. 8, pp. 24033-24045, 2020.

[129] M. Al Jazaery and G. Guo, "*Video-Based Depression Level Analysis by Encoding Deep Spatiotemporal Features*", in IEEE Transactions on Affective Computing, vol. 12, no. 1, pp. 262-268, 1 Jan.-March 2021, doi: 10.1109/TAFFC.2018.2870884.

[130] Y. Zhang, W. Hu and Q. Wu, "*"Autoencoder based on cepstrum separation to detect depression from speech*", in ICITEE2020: 3rd Int. Conf. Inf. Technol. Elect. Eng., Changde City Hunan China. New York, NY, USA: ACM, 2020, doi: 10.1145/3452940.3453038.

[131] Z. Zhao *et al.,* "*Automatic Assessment of Depression From Speech via a Hierarchical Attention Transfer Network and Attention Autoencoders*", in IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 2, pp. 423-434, Feb. 2020, doi: 10.1109/JSTSP.2019.2955012.

[132] A. Othmani *et al.,* "*Towards robust deep neural networks for affect and depression recognition from speech*", Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II. Springer International Publishing, 2021.

[133] Q. Deng, L. Saturnino and G. F. de Sofia, "*Hierarchical attention interpretation: an interpretable speech-level transformer for bi-modal depression detection*", arXiv preprint arXiv:2309.13476, 2023.

[134] C. Rudin, "*Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*", Nature Mach. Intell., vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

[135] M. Squires *et al., "Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment*", Brain Inform., vol. 10, no. 1, Apr. 2023, doi: 10.1186/s40708-023-00188-6.

[136] A. Chatziagapi *et al.,* "*Data Augmentation Using GANs for Speech Emotion Recognition*", Interspeech. 2019.

[137] D. Park *et al.,* "*SpecAugment: a simple data augmentation method for automatic speech recognition. In: Proceedings of the Interspeech 2019*", 2019 Presented at: Interspeech 2019; Sep 15-19, 2019; Graz, Austria. doi: 10.21437/interspeech.2019-2680.

[138] A. Y. Hussenbocus, M. Lech and N. B. Allen, "*Statistical differences in speech acoustics of major depressed and non-depressed adolescents*", 2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS), Cairns, QLD, Australia, 2015, pp. 1-7.

[139] A. Pampouchidou *et al.,* "*Depression assessment by fusing high and low level features from audio, video, and text*", in MM '16: ACM Multimedia Conf., Amsterdam The Netherlands. New York, NY, USA: ACM, 2016, doi: 10.1145/2988257.2988266.

[140] B. Stasak, D. Joachim and J. Epps, "*Breaking Age Barriers With Automatic Voice-Based Depression Detection*", in IEEE Pervasive Computing, vol. 21, no. 2, pp. 10-19, 1 April-June 2022.

[141] M. Pandharipande *et al.,* "*An unsupervised frame selection technique for robust emotion recognition in noisy speech*", in 2018 26th Eur. Signal Process. Conf. (EUSIPCO), Rome, Sep. 3–7, 2018. IEEE, 2018, doi: 10.23919/eusipco.2018.8553202.

[142] R. Chakraborty *et al.,* "*Front-end feature compensation and denoising for noise robust speech emotion recognition*", Proc. Interspeech, pp. 3257-3261, 2019.

[143] L. Yang, D. Jiang and H. Sahli, "*Feature Augmenting Networks for Improving Depression Severity Estimation From Speech Signals*", in IEEE Access, vol. 8, pp. 24033-24045, 2020.

[144] Z. Huang *et al.,* "*Natural Language Processing Methods for Acoustic and Landmark Event-Based Features in Speech-Based Depression Detection*", in IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 2, pp. 435-448, Feb. 2020, doi: 10.1109/JSTSP.2019.2949419.

[145] M. Gerczuk *et al.,* "*Noise Robust Recognition of Depression Status and Treatment Response from Speech via Unsupervised Feature Aggregation*", 2023 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Sydney, Australia, 2023, pp. 1-4, doi: 10.1109/EMBC40787.2023.10340985.

[146] H. Long *et al.,* "*Detecting depression in speech: Comparison and combination between different speech types*", 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, MO, USA, 2017, pp. 1052-1058, doi: 10.1109/BIBM.2017.8217802.

[147] V. Nemes *et al.,* "*A feasibility study of speech recording using a contact microphone in patients with possible or probable Alzheimer's disease to detect and quantify repetitions in a natural setting*", Alzheimer's and Dementia 8.4 (2012): P490-P491.

[148] M. Faurholt-Jepsen *et al.,* "*Voice analysis as an objective state marker in bipolar disorder*", Transl. Psychiatry, vol. 6, no. 7, Jul. 2016, Art. no. e856-e856, doi: 10.1038/tp.2016.123.

[149] A. Kumar and S. K. Agarwal, "*Spoken web: Using voice as an accessibility tool for disadvantaged people in developing regions*", ACM SIGACCESS Accessibility Comput., no. 104, pp. 3–11, 2012.

[150] T. Hodgson, F. Magrabi, and E. Coiera, "*Evaluating the usability of speech recognition to create clinical documentation using a commercial electronic health record*", Int. J. Med. Inform., vol. 113, pp. 38–42, May 2018, Doi: 10.1016/j.ijmedinf.2018.02.011.

[151] K. Cortiñas-Lorenzo and G. Lacey, "*Toward Explainable Affective Computing: A Review*", in IEEE Transactions on Neural Networks and Learning Systems, Doi: 10.1109/TNNLS.2023.3270027.

[152] B. M. L. Srivastava *et al.,* "*Privacy-preserving adversarial representation learning in ASR: Reality or illusion?*", in Interspeech 2019. ISCA: ISCA, 2019, Doi: 10.21437/interspeech.2019-2415.

[153] M. Jaiswal and E. M. Provost, "*Privacy enhanced multimodal neural representations for emotion recognition*", 2019, arXiv:1910.13212.

[154] R. Shokri and V. Shmatikov, "*Privacy-preserving deep learning*", in Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur., 2015, pp. 1310–1321.

[155] A. Ahmed *et al.,* "*A review of mobile chatbot apps for anxiety and depression and their self-care features*", Comput. Methods Programs Biomedicine Update, vol. 1, p. 100012, 2021, Doi: 10.1016/j.cmpbup.2021.100012.

[156] L. Martinengo *et al.,* "*Suicide prevention and depression apps' suicide risk assessment and management: a systematic assessment of adherence to clinical guidelines*", BMC Med., vol. 17, no. 1, Dec. 2019, Doi: 10.1186/s12916-019-1461-z.

[157] L. Martinengo *et al.,* "*Self-guided cognitive behavioral therapy apps for depression: systematic assessment of features, functionality, and congruence with evidence*", J. Med. Internet Res., vol. 23, no. 7, Jul. 2021, Art. no. e27619, Doi: 10.2196/27619.

[158] G. Teepe *et al.,* "*Just-in-time adaptive mechanisms of popular mobile apps for individuals with depression: systematic app search and literature review*", J. Med. Internet Res., Apr. 2021, Doi: 10.2196/29412.

[159] Weninger, *et al.,* "*Recognition of Nonprototypical Emotions in Reverberated and Noisy Speech by Nonnegative Matrix Factorization*". EURASIP Journal on Advances in Signal Processing, 2011(1). Doi: https://doi.org/10.1155/2011/838790

[160] F. Feng *et al.,* "*Test-retest reliability of acoustic and linguistic measures of speech tasks*". Comput. Speech and Lang., vol. 83, p. 101547, Oct. 2023. Doi: https://doi.org/10.1016/j.csl.2023.101547

[161] B. A. Yawer, J. Liss and V. Berisha, "*Reliability and validity of a widely-available AI tool for assessment of stress based on speech*". Scientific Rep., vol. 13, n. 1, nov. 2023. Doi: https://doi.org/10.1038/s41598-023-47153-1

[162] D. M. Low, V. Rao, G. Randolph, P. C. Song and S. S. Ghosh, "*Identifying bias in models that detect vocal fold paralysis from audio recordings using explainable machine learning and clinician ratings*". PLOS Digit. Health, vol. 3, n. 5, mag. 2024, art. n. e0000516. Doi: https://doi.org/10.1371/journal.pdig.0000516

[163] E. Fried, "*Moving forward: how depression heterogeneity hinders progress in treatment and research*". Expert Rev. Neurotherapeutics, vol. 17, n. 5, pp. 423–425, mar. 2017. Doi: https://doi.org/10.1080/14737175.2017.1307737

[164] E. I. Fried *et al.,* "*Measuring depression over time . . . Or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression*". Psycholog. Assessment, vol. 28, n. 11, pp. 1354–1367, nov. 2016. Doi: https://doi.org/10.1037/pas0000275

[165] E. I. Fried and R. M. Nesse, "*Depression sum-scores don't add up: why analyzing specific depression symptoms is essential*". BMC Med., vol. 13, n. 1, apr. 2015. Doi: https://doi.org/10.1186/s12916-015-0325-4

[166] S. Fara, O. Hickey, A. Georgescu, S. Goria, E. Molimpakis and N. Cummins, "*Bayesian Networks for the robust and unbiased prediction of depression and its symptoms utilizing speech and multimodal data*". In INTERSPEECH 2023. ISCA: ISCA, 2023. Doi: https://doi.org/10.21437/interspeech.2023-1709

[167] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps and T. F. Quatieri, "*A review of depression and suicide risk assessment using speech analysis*". Speech Commun., vol. 71, pp. 10–49, lug. 2015. Doi: https://doi.org/10.1016/j.specom.2015.03.004

[168] D. M. Low, V. Rao, G. Randolph, P. C. Song and S. S. Ghosh, "*Identifying bias in models that detect vocal fold paralysis from audio recordings using explainable machine learning and clinician ratings*". PLOS Digit. Health, vol. 3, n. 5, mag. 2024, art. n. e0000516. Doi: https://doi.org/10.1371/journal.pdig.0000516

[169] M. Du et al., "*Depression recognition using a proposed speech chain model fusing speech production and perception features*". J. Affect. Disorders, nov. 2022. Doi: https://doi.org/10.1016/j.jad.2022.11.060

[170] H. Ghasemzadeh, R. E. Hillman and D. D. Mehta. "*Toward Generalizable Machine Learning Models in Speech, Language, and Hearing Sciences: Estimating Sample Size and Reducing Overfitting*". J. Speech, Lang., Hearing Res., pp. 1–29, feb. 2024.

[171] B. Barz and J. Denzler. "*Deep Learning on Small Datasets without Pre-Training using Cosine Loss*". in 2020 IEEE Winter Conf. Appl. Comput. Vis. (WACV), Snowmass Village, CO, USA, 1–5 mar. 2020. IEEE, 2020. Doi: https://doi.org/10.1109/wacv45572.2020.9093286

[172] X. Zhang, X. Zhang, W. Chen, C. Li e C. Yu, "*Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments*". Scientific Rep., vol. 14, n. 1, apr. 2024. Doi: https://doi.org/10.1038/s41598-024-60278-1

[173] N. Facchinetti, F. Simonetta, and S. Ntalampiras, "*A Systematic Evaluation of Adversarial Attacks against Speech Emotion Recognition Models*", Intelligent Computing, vol. 3, jan. 2024. Doi: 10.34133/icomputing.0088

[174] V. Berisha et al., "*Digital medicine and the curse of dimensionality*", npj Digit. Medicine, vol. 4, n. 1, ott. 2021. Doi: https://doi.org/10.1038/s41746-021-00521-5

[175] S. Aleem, N. u. Huda, R. Amin, S. Khalid, S. S. Alshamrani e A. Alshehri, "*Machine Learning Algorithms for Depression: Diagnosis, Insights, and Research Directions*", Electronics, vol. 11, n. 7, p. 1111, mar. 2022. Doi: https://doi.org/10.3390/electronics11071111

[176] P. A. Vöhringer and R. H. Perlis, "*Discriminating Between Bipolar Disorder and Major Depressive Disorder*", Psychiatr. Clin. North Am., vol. 39, no. 1, pp. 1–10, Mar. 2016. Doi: https://doi.org/10.1016/j.psc.2015.10.001

[177] Z. Wu et al., "*Clinical distinctions in symptomatology and psychiatric comorbidities between misdiagnosed bipolar I and bipolar II disorder versus major depressive disorder*", BMC Psychiatry, vol. 24, no. 1, May 2024. Doi: https://doi.org/10.1186/s12888-024-05810-3

[178] S. M. Alghowinem, T. Gedeon, R. Goecke, J. Cohn, and G. Parker, "*Interpretation of Depression Detection Models via Feature Selection Methods*", IEEE Trans. Affect. Comput., p. 1, 2020. Doi: https://doi.org/10.1109/taffc.2020.3035535

[179] A. Pampouchidou et al., "*Automatic Assessment of Depression Based on Visual Cues: A Systematic Review*", IEEE Trans. Affect. Comput., vol. 10, no. 4, pp. 445–470, Oct. 2019. Doi: https://doi.org/10.1109/taffc.2017.2724035

[180] J. W. Weeks et al., "*The Sound of Fear": Assessing vocal fundamental frequency as a physiological indicator of social anxiety disorder*", J. Anxiety Disorders, vol. 26, no. 8, pp. 811–822, Dec. 2012. Doi: https://doi.org/10.1016/j.janxdis.2012.07.005

[181] I. R. Titze, "*Principles of voice production*", Englewood Cliffs, N.J: Prentice Hall, 1994.

[182] S. Alghowinem, R. Goecke, J. F. Cohn, M. Wagner, G. Parker, and M. Breakspear, "*Cross-cultural detection of depression from nonverbal behaviour*", in 2015 11th IEEE Int. Conf. Workshops Autom. Face

Gesture Recognit. (FG), Ljubljana, May 4–8, 2015. IEEE, 2015. Doi: https://doi.org/10.1109/fg.2015.7163113
[183] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and D. M. Wilkes, "*Acoustical properties of speech as indicators of depression and suicidal risk*", in IEEE transactions on bio-medical engineering, vol. 47, no. 7, pp. 829–37, jul 2000.
[184] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes, "*Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk*", Vanderbilt University, Department of Biomedical Informatics, Nashville, TN 87215 USA. asli.ozdas@vanderbilt.edu, Tech. Rep. 9, 2004.
[185] K. Mao, Y. Wu, and J. Chen, "*A systematic review on automated clinical depression diagnosis*", npj Mental Health Res., vol. 2, no. 1, Nov. 2023. Doi: https://doi.org/10.1038/s44184-023-00040-z
[186] K. A. Britto, D. T. Varnikaa, R. Sathishkumar, and E. Anita Dolorosa, "*ML Model For Detection of Depression in Menopause Women*", in 2024 3rd Int. Conf. Artif. Intell. Internet Things (AIIoT), Vellore, India, May 3–4, 2024. IEEE, 2024. Doi: https://doi.org/10.1109/aiiot58432.2024.10574652
[187] M. M. Ali et al., "*Development and performance analysis of machine learning methods for predicting depression among menopausal women*", Healthc. Anal., p. 100202, May 2023. Doi: https://doi.org/10.1016/j.health.2023.100202

Samara Soares Leal received her BSc in Computer Science in 2012, followed by an MSc and PhD in Mathematical and Computational Modelling in 2016 and 2020, respectively, from the Federal Centre for Technological Education of Minas Gerais (CEFET-MG), Brazil. She is currently a postdoctoral researcher at the University of Milan, working on the project Paralinguistic Speech Signal Processing for Depression Assessment. In 2023, she completed an MBA in Artificial Intelligence. For the past four years, she has worked as a national project manager at Ânima Educação, Brazil. She has been a university teacher at higher education institutions in Brazil for six years. Her research interests include artificial intelligence, audio pattern recognition, software, game, and simulator development for medical education. Additionally, she works on content recommendation based on students' psychometric profiles and traffic light optimization.

Stavros Ntalampiras is an Associate Professor at the Department of Computer Science, University of Milan, Italy. He received the engineering and Ph.D. degrees from the Department of Electrical and Computer Engineering, University of Patras, Greece, in 2006 and 2010, respectively. He has carried out research and/or didactic activities at Politecnico di Milano, the Joint Research Center of the European Commission, the National Research Council of Italy, and Bocconi University. Currently, he is an Associate Editor of IEEE TNNLS, PLOS One, IET Signal Processing and CAAI Transactions on Intelligence Technology, as well as member of the IEEE Computational Intelligent Society Task Force on Computational Audio Processing. His research interests include signal processing, machine learning, audio pattern recognition, bioacoustics, medical acoustics, and cyber-physical systems.

Roberto Sassi is Full Professor of Computer Science at the University of Milan, Italy, where he teaches or have taught courses on digital signal and image processing, biomedical signal processing, intelligent systems and statistics. He graduated in electronic engineering in 1996 and received a Ph.D. in biomedical engineering in 2001, both from Politecnico di Milano, Italy. Prof. Sassi is the coordinator of the Ph.D. programme in Computer Science (2021–) and the director of the laboratory "Biomedical image and Signal Processing" (BiSP, 2008-), at the University of Milan. He is an IEEE senior member (2012–) and the vice-president of the IEEE EMB Italy Section Chapter (EMB18). He participated in the coordination of the activities of EU (MY-ATRIA, NESTORE, INSIDE-HEART) and national (SMARTA, COVIDSQUARED, SOLITAIRE) funded projects. His research interests have been mostly in the fields of biomedical signal and image processing, mainly to address related challenges in computer science, as biometrics and digital health.