Expert Systems With Applications 257 (2024) 125025

Contents lists available at ScienceDirect

Expert Systems With Applications

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

An End-to-End framework for extracting observable cues of depression from diary recordings

Izidor Mlakar[a], Umut Arioz[a], Urška Smrke[a], Nejc Plohl[b], Valentino Šafran[a], Matej Rojc[a,*]

[a] University of Maribor, Faculty of Electrical Engineering and Computer Science, Koroška cesta 46, 2000 Maribor, Slovenia
[b] University of Maribor, Faculty of arts, Department of Psychology, Koroška cesta 160, 2000 Maribor, Slovenia

ARTICLE INFO

Keywords:
Digital biomarkers of depression
Facial cues
Speech cues
Language cues
End-to-end pipeline
Artificial intelligence

ABSTRACT

Because of the prevalence of depression, its often-chronic course, relapse and associated disability, early detection and non-intrusive monitoring is a crucial tool for timely diagnosis and treatment, remission of depression and prevention of relapse. In this way, its impact on quality of life and well-being can be limited. Current attempts to use artificial intelligence for the early classification of depression are mostly data-driven and thus non-transparent and lack effective means to deal with uncertainties. Therefore, in this paper, we propose an end-to-end framework for extracting observable depression cues from diary recordings. Furthermore, we also explore its feasibility for automatic detection of depression symptoms using observable behavioural cues. The proposed end-to-end framework for extracting depression was used to evaluate 28 video recordings from the Symptom Media dataset and 27 recordings from the DAIC-WOZ dataset. We compared the presence of the extracted features between recordings of individuals with and without a depressive disorder. We identified several cues consistent with previous studies in terms of their differentiation between individuals with and without depressive disorder across both datasets among language (i.e., use of negatively valanced words, use of first-person singular pronouns, some features of language complexity, explicit mentions of treatment for depression), speech (i.e., monotonous speech, voiced speech and pauses, speaking rate, low articulation rate), and facial cues (i.e., rotational energy of head movements). The nature/context of the discourse, the impact of other disorders and physical/psychological stress, and the quality and resolution of the recordings all play an important role in matching the digital features to the relevant background. In this way, the work presented in this paper provides a novel approach to extracting a wide range of cues relevant to the classification of depression and opens up new opportunities for further research.

1. Introduction

Depression is a major public health problem today. It can affect both mental and physical health (Liu et al., 2020). It has even become a common chronic disease worldwide. 3.8 % of the world's population, or 280 million people, are already affected by depression. This includes 5.0 % of adults and 5.7 % of adults over the age of 60 (World Health Organization, 2021). This pandemic has therefore significantly increased the need to strengthen mental health systems and triggered a 25 % increase in the global prevalence of anxiety and depression (Santomauro et al., 2021, World Health Organization, 2022). Furthermore, depression is a highly prevalent comorbidity in patients with chronic diseases such as diabetes, heart disease and stroke, chronic obstructive pulmonary disease, and cancer (Li et al., 2019). Depression often interferes with normal functioning and causes depressive thoughts, which may even lead to suicide (Yang et al., 2021). Therefore, early identification of depressive symptoms is very important to control symptoms and the underlying condition. Treating depression in its early stages can successfully reduce its negative impact on well-being and health, as well as on economic, personal and social life (Harlé et al., 2010, Linder et al., 2020, Uddin et al., 2022).

In current practice, screening for depression is done reactively, when the first symptoms are observed, as episodic healthcare remains the norm (Solid, 2022) (i.e., symptom-driven). The problem, however, is that the overlapping clinical symptoms of depression and many chronic conditions make it very difficult to definitively identify symptoms at an

* Corresponding author.
E-mail addresses: izidor.mlakar@um.si (I. Mlakar), umut.arioz@um.si (U. Arioz), urska.smrke@um.si (U. Smrke), nejc.plohl1@um.si (N. Plohl), valentino.safran@um.si (V. Šafran), matej.rojc@um.si (M. Rojc).

https://doi.org/10.1016/j.eswa.2024.125025
Received 13 December 2023; Received in revised form 16 May 2024; Accepted 5 August 2024
Available online 8 August 2024

early stage (Craven et al., 2013, DeJean et al., 2013, Menear et al., 2015, Shirazian et al., 2017). Thus, unrecognised symptoms of depression, particularly in patients with chronic conditions in primary care, are of great concern (Egede, 2007), as untreated depression often leads to poorer quality of life, poorer treatment outcomes and greater disability (Brenes, 2007, Ambaw et al., 2018, Ploughman et al., 2020). In addition, screening often uses standardised tools such as the Patient Health Questionnaire (PHQ-8 and 9) (Kroenke et al., 2001, Kroenke et al., 2009). These tools rely on patients' subjective opinions about their past experiences and self-reports, which are affected by reporting and interpretation biases (Nahum et al., 2017) and are often underreported or exaggerated (Jacobson et al., 2020), depending on the individual (Sato et al., 2011, Al-Mosaiwi et al., 2018, Cole et al., 2018, Pietkiewicz et al., 2021). Therefore, clinicians' diagnostic ability is often hampered by confusion of depression with other disorders and suboptimal diagnostic accuracy due to response bias (Bickman et al., 2020, D'Alfonso, 2020).

Nowadays, machine learning (ML) and artificial intelligence (AI) also represent a very promising way for timely and accurate screening and diagnosis of depression (Bickman, 2020, D'Alfonso, 2020, Singh et al., 2022). For example, tools for better pre-diagnostic risk assessment and screening of an individual's predisposition or risk of developing mental illness can be developed using various AI and ML techniques (Graham et al., 2019, Su et al., 2020, Balcombe et al., 2021). However, a relatively narrow understanding of the interactions between these psychological, biological and social systems (Graham et al., 2019) makes the development of more advanced knowledge-based systems rather challenging. On the other hand, data-driven (i.e., non-knowledge-based) systems offer a unique opportunity to address this heterogeneity in the pathophysiology and diagnosis of mental illness by leveraging big data to discover novel complex patterns in high-dimensional data that are far beyond human comprehension (Bzdok et al., 2018, Su et al., 2020, Balcombe et al., 2021, Fazi, 2021, Wongkoblap et al., 2021).

In healthcare, any mistake can cost lives. Therefore, the unexplainable and insufficiently interpretable nature of AI results makes it less acceptable to regulators and clinicians for such a task (Shortliffe et al., 2018, Ramon et al., 2021). Therefore, the design of an explainable, knowledge-based system is indeed essential for widespread adoption in such practice (Devaraj et al., 2014). We have already analysed the perspective of behavioural cues as observable biomarkers of depression in our previous work (Smrke et al., 2021). And these cues are spontaneous. Thus, the influence of cognitive and other biases, such as social desirability, is significantly reduced (Low et al., 2020). In previous work, we have also found that increased use of absolutist words, increased use of first-person singular pronouns, shorter sentences, repetitive pitch inflections, reduced intonation, predominant use of sad, negative, and neutral expressions, and reduced emotional facial expressivity, etc. are also meaningful cues to depression (Smrke et al., 2021). Thus, the aim of this paper is to propose and evaluate an end-to-end (E2E) pipeline capable of extracting speech cues, both in their linguistic and paralinguistic content, with additional visual cues from diary recordings, and in this way to evaluate the contribution of these extracted cues to advanced depression classification when using knowledge-based and non-knowledge-based models in the process.

2. Related works

Deep learning (DL) models are increasingly used, for example, for early detection and screening of various diseases, for understanding, monitoring their progression and optimising possible treatments. In fact, as powerful models, they can lead to more automated diagnosis (Keane and Topol, 2018, Lee et al., 2021), etc. Although the number of studies is still relatively low, artificial intelligence also has great potential in the field of mental health (Cummins et al., 2015, Burdisso et al., 2019, Mezzi et al., 2022, Liu et al., 2022). Indeed, depressive disorders tend to be more studied in this topic (Graham et al., 2019). However, the existing solutions, based on AI algorithms and developed to recognise the observable cues of depression, are generally data driven. Furthermore, in most cases they also focus on a single modality or at most two modalities (i.e. unimodal or bimodal algorithms are used) (Arioz et al., 2022). This is also due to the availability of datasets. Thus, most of the research can focus on text (language), for example, in early risk detection and analysing data posted in social media streams (Burdisso et al., 2019, Liu et al., 2022, Mezzi et al., 2022, Uddin et al., 2022).

The classification of depression from speech is also quite common. In general, it is done by analysing acoustic (Cummins et al., 2015, Flanagan et al., 2021) or a combination of speech and acoustic cues (Lam et al., 2019, Lin et al., 2020, Sequeira et al., 2020, Sardari et al., 2022). In fact, they are similar to models that analyse visual cues or a combination of acoustic and visual cues (Maddage et al., 2009, Jan et al., 2017, Alghowinem et al., 2020). Furthermore, state-of-the-art algorithms operate on the assumption of continuity between facial and emotional expressivity. While these studies have reported results with high accuracy, they should be considered as an early proof of concept to address mental health issues (Graham et al., 2019, Pan et al., 2024). In fact, depression is expressed simultaneously through three modalities or trimodal approaches (Ray et al., 2019, Rohanian et al., 2019, Qureshi et al., 2019, Ceccarelli and Mahmoud, 2022, Nickels et al., 2021). Therefore, trimodal approaches have the potential to further improve the accuracy and, most importantly, the reliability of these predictions (Arioz et al., 2022). The main issue here is the awareness of the ethical challenges behind the use of 'black box' approaches (based on AI), especially when including already biased data (e.g. the expressive and subjective nature of clinical text, the association between culture and mental illness) (Chen et al., 2019, Habbal et al., 2024) and the explainability of the recommendations and generated classifications in healthcare settings (Lee et al., 2021, El-Sherif et al., 2022).

Furthermore, recent work has mainly focused on the identification/analysis of unique and specific features of the complex non-verbal and verbal production process in controlled environments. All studies have implemented feature extraction in their own way, either manually or fully automatically. Thus, there are no generic pipelines to extract specific features and generate a universal approach. In fact, to the best of our knowledge, none of these studies provide a set of observable parameters that can be directly linked to AI or used to explain/validate AI-based classification.

The main goal of this paper is to propose and evaluate a framework for automatic extraction and quantification of the observable cues of depression as presented in Smrke et al. 2021. These moderators can complement the explainability for AI algorithms for depression classification and can present a robust set of features for future development of explainable AI algorithms. The main contributions of the research in this paper are:

- An E2E pipeline that allows seamless extraction of observable behavioural cues from video recordings.
- Digitisation of observable behavioural cues as identified by relevant psychological/psychiatric background.
- a framework of multimodal and minimally language dependent classification features, and
- a case study of how digital features are represented in real video s of depressed and non-depressed individuals.

3. Methodology

3.1. Datasets

Currently, the DAIC-WOZ dataset has mostly been used for depression analysis (Arioz et al., 2022). Furthermore, work in this area still mostly uses unimodal algorithms for depression detection, while some papers process two communication channels simultaneously or all three modalities (text, audio, video). Only a few of them process audio and visual cues together with physiological biomarkers, or simultaneously process speech, language and visual cues. The BDC-WB dataset was used in (Ceccarelli and Mahmoud, 2022) and the AVID-C dataset in (Sato and Kawahara, 2011).

The audio −Visual Bipolar Disorder Corpus (BDC) was created for research on personalised treatment of BD patients and contains audio −visual recordings of patients with BD and healthy controls. In this dataset, 35 male and 16 female patients were recruited from the mental health department of a hospital. Clinical information includes age, identity, severity of illness and treatments used. These records are annotated for BD state (hypomania, mania and depression). They are also annotated for the Young Mania Rating Scale (YMRS) by psychiatrists (Ciftci et al., 2018). This dataset was prepared for Turkish language and in a semi-structured interview; therefore, this study was not suitable for our study.

Orton collected a non-clinical dataset to investigate the bodily modality for the detection of psychological distress, the Well-being (WB) dataset. Participants, recruited at the University of Cambridge, were interviewed in face-to-face sessions by a researcher. This dataset contains facial expressions, body movement information, gestures and audio recordings for 35 interviewed subjects. Labels were also assigned using self-report questionnaires. The questionnaires used were the GAD-7 for anxiety, the PHQ-8 for depression, the SSS-8 for somatic symptoms and the PSS for perceived stress (Orton, 2020). This dataset is not a clinical dataset and is still in development.

The AVEC2013 dataset is a subset of the audio −visual depressive language corpus (AviD corpus), which contains 340 video clips of subjects performing a human–computer interaction task while being recorded by a webcam and microphone. The video clips range in length from 50 to 20 min (with an average of 25 min). The behaviour within the video clips consisted of different tasks. video feature extraction involves the extraction of Local Phase Quantisation (LPQ) features. The audio features then represent a set of 42 functionals. Namely, statistical functionals (23), regression functionals (4), local minima/maxima related functionals (9), other (LP, LPC) (6)) (Valstar et al., 2013). The AVID corpus is a bimodal dataset. There is a download link, but it is not working and not publicly available.

There are also some recently published datasets for depression, but most of them are bi-modal or have different modalities rather than linguistic, speech and visual (Cai et al, 2022 –(MODMA dataset) EEG and audio; Thati et al. (2023) – smart phone usage indicators, facial and audio features; Shen et al. (2022) – audio and text; Dibeklioglu et al. (2017) – (Pittsburg dataset) audio and video; Yoon et al. (2022) –(Black Dog dataset) audio and video.

In this study we focus on multimodal datasets, i.e. datasets capturing three or more modalities. Namely, compared to unimodal and bi-modal the multimodal datasets, have the ability to capture a wider range of information from different sources, such as audio, visual, and text data. However, despite contributions to human emotion studies, there remains a scarcity of multi-modal datasets for comprehensive psychological, physiological, and behavioural analysis in mental health research (Li et al., 2022). For this reason, we choose DAIC-WOZ dataset (DW) and SymptomMedia (SM) dataset.

The DAIC-WOZ dataset (DW) is an integral component of the broader Distress Analysis Interview Corpus (DAIC) (Gratch et al., 2014). The DAIC comprises of chatbot-driven interviews designed to aid in the diagnosis of mental health conditions like anxiety, depression, and post-traumatic stress disorder. These interviews were collected in a project aimed at developing a computer agent capable of conducting interviews and discerning verbal and nonverbal cues indicative of mental illness (DeVault et al., 2014). The dataset encompasses audio audio and video recordings, along with detailed responses to questionnaires. This specific segment of the corpus is centred around “Wizard-of-Oz” interviews, where participants interacted with an animated virtual interviewer named Ellie. Notably, Ellie was controlled by a human interviewer located in a different setting. The data has undergone transcription and annotation to emphasize various verbal and nonverbal attributes. The dataset comprises a total of 189 sessions, divided into predefined training, development, and testing sets. These sets include patient IDs, PHQ-8 scores, binary labels indicating PHQ-8 scores greater than or equal to 10, and gender information. Each session provides the interaction transcript, an audio file, audio features, and facial features. DW was selected since it allows us to assess the consistency of the proposed framework and the moderators used to formulate the digital representation of observable behavioural cues. However, the discourse in the DW does contain clinical interviews and is chatbot-driven conversation featuring ordinary subjects, thus, may lead to the under-emphasis of critical features.

SymptomMedia (SM) constitutes a collection of video recordings emulating the symptoms associated with various mental health conditions, including depression, anxiety, and schizophrenia (Symptom Media Films, 2023). These video s serve educational and training purposes within fields like psychology, social work, and counselling, aiding professionals in recognizing and comprehending diverse mental health symptoms. The dataset is explicitly crafted to enhance the diagnostic and treatment skills of students and practitioners. Featuring over 600 films, these recordings showcase actors portraying symptoms rather than actual patients, with durations spanning from 4 to 14 min. The video s exclusively present the patient’s appearance, concealing the interviewer, and each film is accompanied by transcripts for both the interviewer and the patient. The information stored in SM dataset is in contrast to DM dataset, artificial of nature, enacted by actors and may overemphasize certain features, aligning more closely with theoretical expectations. We chose the SM dataset to compensate for the under-emphasis within DM recordings.

In this paper we used 56 recordings sessions from the DW and SM dataset. For the SM sessions we chose those that have a low probability of overlap with symptoms of major depression (Plana-Ripoll et al., 2019). From the DW dataset, therefore, we selected 12 recordings classified as having moderate and severe symptoms of depression (with a PHQ-8 score above 14), and 15 recordings without significant depressive symptoms (with a PHQ-8 score of 0). The dataset consists of audio recordings, transcriptions, and pre-extracted low-level features (e.g., gaze, 3D actions units, and 3D head position and rotation), which are then used to construct the observable visual cues. However, the facial features are the same as those extracted from the SM dataset. The complementary nature of the features and the accuracy of the pipeline have been previously evaluated in (Arioz et al., 2022).

3.2. Observable behaviour cues

According to Smrke et al, 2021, the observable behavioural cues are classified into three distinct categories, i.e., speech, language and facial expressions. In this paper, we answer the question of whether artificial intelligence (AI) can support the process of detecting depression in cancer survivors using feature extraction from face, audio, and transcriptions. In the next subsections, we present how the observable behavioural cues are digitised using several AI-driven moderators (i.e., features). We focus on language, speech cues and facial expression cues.

3.2.1. Digital representation of linguistic cues

Based on the study by Smrke et al. 2021, the following features of linguistic cues are proposed in this study: (A1) use of first-person singular pronouns, (A2) use of specific linguistic features, such as negative-valence words (e.g., words related to rumination, expressions of sadness and anger, pain and aggression), (A3) explicit mention of the treatment of depression (i.e., treatment, therapy, pills, psychotherapy), (A4) use absolutist words (i.e., use words such as “nothing”, “always”, “everything”, “never”, “all the time”), (A5) focus on the past, (A6) avoid complex sentences and instead use shorter sentences, especially those containing adverbial clauses (e.g., “even though” and “as soon as”). The observable cue (A6) to measure the estimation of the complexity of language consists of evaluating: (A6.1.1) sentence length, (A6.1.2) sentence complexity, (A6.2.1) lexical diversity, (A6.2.2) lexical sophistication, (A6.2.3) lexical density. For (A4) small grammar (words like “always”, “never”, “all the time”, “everything”, “nothing” from previous studies) and extended grammar (words like “absolutely” “all”, “always”, “complete”, “completely”, “constant”, “constantly”, “definitely”, “entire”, “ever”, “every”, “everyone”, “everything”, “full”, “must”, “never”, “nothing”, “totally”, “whole”) were used.

See Appendix A.1 for a detailed description of these digitalized features.

3.2.2. Digital representation of cues from speech

Based on the work of Smrke et al, 2021 the following cues are proposed in this study: (B1) speaking rate, (B2) engagement in verbal communication, (B3) voiced speech and pauses, when answering questions, during conversation and when initiating speech, (B4) low articulation rate with flatter trajectory (i.e., low articulatory movements), (B5) decreased voice quality, and (B6) monotonous speech. Observable cue B2 and B5 are more complex cues. Engagement in verbal communication (B2) is measured by analysing temporal and voice quality characteristics (pitch (B2.1), intensity (B2.2) and formants (B2.3)) and head movement amplitude (B2.4).

See Appendix A.2 for a detailed description of these digitalized features.

3.2.3. Digital representation visual cues

Based on the study by Smrke et al. 2021, the following features are proposed in this study: (C1) occurrence of facial expressions and facial mobility, (C2) intensity of facial expressions, (C3) occurrence and emotional variability, (C4) eyebrow movements, (C5) gaze aversion and downward gaze, (C6) frowns, (C7) turns and head movement, i.e., head looking away from the collocutor, (C8) rotational energy of head movements, (C9) mouth movements, (C10) lip press, (C11) down-angled mouth corners, and (C12) smiling.

Table Appendix A.3 provides a detailed description of these digitalized features.

As can be seen in Table 1, most of the studies define the low-level descriptors as hand-crafted features, although hand-crafted features are defined as the high-level features that are constructed from several combinations of low-level descriptors. In this study we define hand-crafted features for linguistic, speech and facial feature extraction, where most of these hand-crafted features, especially for speech and video, are novel for the extraction of depression features. These hand-crafted features were selected following the work of Smrke et al, 2021, who identified the observable features of depression that can be captured using artificial intelligence from the scoping *meta*-review study. In addition, subjective observable cues were mapped to objectively measurable traits A1-A6, B1-B6 and C1-C12. These traits constitute the main contribution to literature, and the novelty part of our study.

3.3. An E2E framework for extracting digital cues from the observable behavior

In order to extract these digital cues from video recordings, we propose and implement an E2E pipeline, which is outlined in Fig. 1.The pipeline consists of (1) data preprocessing component, which includes a validity check, (2) an audio / video signal extraction and a multilingual speech recognition engine, (3) a component for extracting *Digital Linguistic Cues*, (4) a component for extracting *Digital Speech Cues*, and (5) a component for extracting *Digital Visual Cues*.

As can be seen in Fig. 1, each input video file is split into video and audio streams. The presence of the face in the corresponding video stream is then checked. The presence of the face is monitored using the OpenCV library (Bradski, 2000). If the face is not present in the first 5 s of the video, the video is discarded. The audio stream is passed to the

Table 1
The comparison of features used for depression detection[1.]

Modality	Hand-Crafted Features in The Literature	Our Low-Level Descriptives	Our Hand-Crafted Features
text [Maupomé et al., 2020]	syntactic: freq. of select Parts of speech; lexical: freq. of most frequent word unigrams, bigrams; freq. of most frequent character unigrams, bigrams, trigrams; number of unique words; number of alphanumeric characters; number of alphanumeric characters; number of digits; number of non-ASCII characters; punctuation ratio; morphological: average length of words; number of long words; number of short words; number of uppercase words; number of uppercase characters; pragmatical: average length of sentences; number of hyperlinks	lexical: 300-sized vector representation	**A1-A6**
text [Trifan et al., 2020]	lexical category of a user’s text; use of self-related words; use of absolutist words; mentions of words related to mental disorders; use of specific words and their derivatives		
text [Diep et al., 202210]	discourse mapping; local coherence; lexical complexity and richness; syntactic complexity; utterance cohesion; sentiment; word finding difficulty		
speech [Biswas et al., 2021; Lang and Cui, 2018]	LLD, MRELBP	prosodic: F0; VUV; energy; duration.	**B1-B6**
speech [Li et al., 2022; Bailey and Plumbley, 2021; Rejaibi et al., 2022]	spectral: MFCC	spectral: MFCC; HMPDM0-24; HMPDD0-12; LPCC; LFPC; GFCC; formants	
speech [Tasnim and Novikova, 2023; Diep et al., 202210]	spectral: intensity; MFCC; ZCR; voice-related: F0; HNR; shimmer and jitter; durational features; pauses and fillers; phonation rate	voice quality: NAQ; QOQ; H1H2; PSP; MDQ; peakSlope; Rd; jitter; shimmer; harmonics-to-noise ratio; teager energy operator.	
speech [Li et al., 2018]	spectral: MFCC; ZCR; energy and formant frequencies		
speech [Wu et al., 2023]	prosodic: changes in pitch and loudness based on F0 and energy; changes in length of syllables, words, phrases; voice quality: normalized amplitude quotient;		

(*continued on next page*)

Table 1 (*continued*)

Modality	Hand-Crafted Features in The Literature	Our Low-Level Descriptives	Our Hand-Crafted Features
	quasi open quotient; harmonic difference; spectrum perturbation; amplitude perturbation; formants F1-F3; spectral: MFCC; LPCC.		
Speech [Vázquez and Gallardo, 2020]	voice-related: F0; formants; energy; normalized amplitude quotient; spectral: MFCC		
video [Jan et al., 2018]	local binary patterns; edge orientation histogram; local phase quantization	action units: AU01-02/04–06/09–10/12/14–15/17/20/25–26	
video [Fan and Tjahjadi, 2019]	Euclidean distances, median, magnitude for consecutive facial points;	eye gaze: vectors of both eyes (world and head coordinate space)	**C1-C12**
video [Giannakakis et al., 2022]	Action Units (AUs)	head pose:pose confidence; vectors of head (world position and rotation coordinate space) facial landmarks: 68 2D/3D points on face	

[1] Mel-Frequency Cepstrum Coefficient (MFCC); Fundamental frequency (F0); Binary decision of voicing/unvoicing part (VUV); harmonic model and phase distortion mean (HMPDM0-24);harmonic model and phase distortion deviations (HMPDD0-12); linear-prediction cepstral coefficients (LPCC); log-frequency power coefficients (LFPC); gammatone frequency cepstral coefficient (GFCC); normalized amplitude quotient (NAQ); quasi open quotient (QOQ); differentiated glottal source spectrum (H1H2); parabolic spectral parameter (PSP); maxima dispersion quotient (MDQ); spectral tilt/slope of wavelet responses (peakSlope); shape parameter of the Liljencrants-Fant model (Rd); Median Robust Extended Local Binary Pattern (MRELBP); Zero-Crossing Rate (ZCR); Harmonic-to-Noise Ratio (HNR).

multilingual automatic speech recognition (ASR) engine. If there is no speech in the audio stream, the audio is also ignored. If the video is suitable for further comprehensive processing, the video stream is sent to the Visual Features Extraction (VFE) component to extract digital visual cues (C1-C12), the audio stream is sent to the Speech Features Extraction (SFE) component to extract digital speech cues (B1-B6), and the text stream from the ASR is sent to the Linguistic Features Extraction (LFE) component to extract linguistic cues (A1-A6). Finally, the result is stored as a HL7 FHIR composition resource (JSON format) (Ayaz et al., 2021).

The complete feature extraction is shown in more detail in Fig. 2. As can be seen, the E2E feature extraction starts with the video input (X). First, face recognition is performed within a predefined time interval (m). The system continues only if there is a face suitable for facial feature extraction. Then, audio (a) and its transcription (T) are extracted for speech and linguistic feature extraction respectively. NLTK and Stanza are used for low-level linguistic features (TXT-LLD), MATLAB and PRAAT for low-level speech features (SPEECH-LLD) and The OPENFACE tool for low-level visual features (VISUAL-LLD). All low-level features are then mapped to manually created high-level features (TXT-HCFs, SPEECH-HCFs and VISUAL-HCFs). Finally, all manually created features are stored on the FHIR server.

Linguistic low-level descriptors were extracted using the NLTK toolkit (Loper and Bird, 2002) and the Stanza package (Qi et al., 2020) in the Python programming language. These descriptors included first person singular pronouns, words associated with depression, absolutist words, verbs in past tense, sentence length, sentence complexity, word diversity, token length, and lexical item ratio for feature mapping.

Speech low-level descriptors were obtained using MATLAB via the COVAREP repository (Degottex, et al., 2014). The COVAREP script extracted features related to glottal source and spectral envelope from sound files using a feature matrix of 35 features per frame. The Parselmouth library was utilized for the PRAAT toolbox, offering a Pythonic interface to PRAAT's internal code (Jadoul et al., 2018). This allowed for extraction of features such as speaking rate, pitch, speech intensity, formants, pause duration, voice percentage, silence percentage, articulation rate, jitter, shimmer, and harmonics to noise ratio for speech analysis.

For visual low-level descriptors, OpenFace 2.2.0 (a facial behavior analysis toolkit) was used to extract facial landmarks, head pose estimations, action units, and gaze directions from video inputs, providing a single file output for feature mapping.

3.3.1. *Linguistic feature extraction component (LFE)*

The natural language processing (NLP) pipeline (Fig. 3) was designed using Stanza and the NLTK natural language processing library. We used Stanza because, unlike other widely used toolkits, Stanza provides a language-agnostic, fully neural pipeline for text analysis.

This pipeline takes text streams generated on input, i.e. the 'transcription' generated by the SPREAD ASR (Mlakar et al., 2021). The text (document) is additionally annotated with 'low-level features' using predefined natural language processors. In the pipeline, we use the full set of processors available in Stanza and NLTK. The specific digital linguistic cues A1-A6 are then extracted by implementing the cue definition (Fig. 3) in Python. For example, we analyse the morphology, i.e., part-of-speech (POS) and morphological features of the word, of the annotated document to extract the number of first-person singular pronouns (Cue A1), as well as counting the number of verbs in the past tense (Cue A5), the use of nouns, verbs, adjectives and adverbs to measure lexical sophistication (Cue A6. 2.2), analyse the use and variety of different words and word types to measure lexical diversity (cue A6.2.1), lexical density (cue A6.2.3) and sentence complexity (cue A6.1.2).

3.3.2. *Speech recognition engine (SPREAD)*

SPREAD in Fig. 4 (Šafran et al., 2024) is an automatic speech recognition (ASR) system that supports multiple languages and employs neural acoustic model with a connectionist temporal classification (CTC) technique. The system utilizes a convolutional neural network (CNN) for both acoustic and pronunciation models, with preprocessing involving the calculation of mel filter bank features. The acoustic model follows a B×R architecture, comprising 10 blocks with 5 sub-blocks each, incorporating 1D-convolution, batch normalization, ReLU activation, and dropout. Residual connections with projection layers and batch normalization connect these blocks. The NovoGrad optimizer is utilized for second-moment computations per layer (Ginsburg et al., 2019). The SPREAD decoder transforms a probability distribution over characters into text, utilizing a beam search decoder with language model rescoring to identify and grade relevant decoding based on N-gram similarity, spell checking model, and model for predicting punctuations and capitalization. For training acoustic model for english language, CommonVoice and LibriSpeech datasets were used (approx. 2000 + hours), while for training language model, spell checker and punctuations with capitalization, Google corpus was used (Chelba et al., 2013). Evaluation on test data (320 h of speech) resulted in 2.9 %-word error rate (WER).

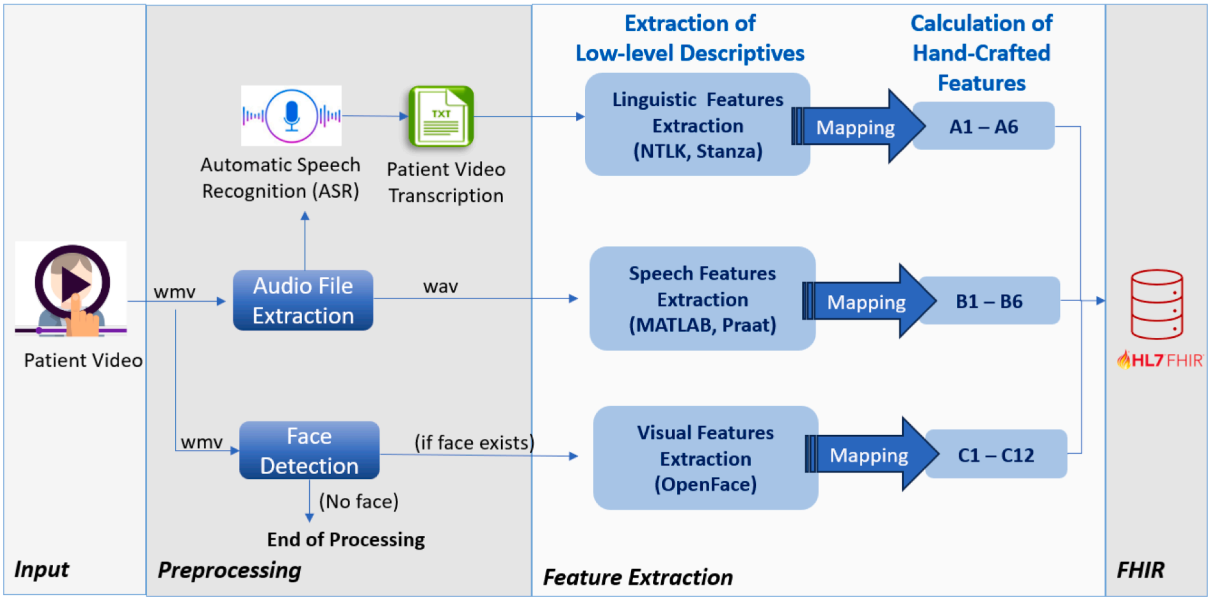
Input
Patient Video
Preprocessing
wmv
Audio File Extraction
wav
Automatic Speech Recognition (ASR)
TXT
Patient Video Transcription
wmv
Face Detection
(if face exists)
(No face)
End of Processing
Feature Extraction
Extraction of Low-level Descriptives
Calculation of Hand-Crafted Features
Linguistic Features Extraction (NTLK, Stanza)
Mapping
A1 – A6
Speech Features Extraction (MATLAB, Praat)
Mapping
B1 – B6
Visual Features Extraction (OpenFace)
Mapping
C1 – C12
FHIR
HL7 FHIR

Fig. 1. The proposed E2E framework for extracting digital cues.

```
Input → Video file X = frames[1..N],
                    m = face detection time interval,
                    fps = frame per second
  // face detection
  for n = 1 to (m x fps)
     if no face
        return
  // audio data extraction – a
  ffmpeg{X} → a
  // transcriptions extraction – T
  ASR{a} → T
  // feature extraction
  NLTK{T}, Stanza{T} → TXT-LLD (linguistic low-level descriptors)
  // calculation of linguistic hand-crafted features
  TextProcessor{ TXT-LLD } → TXT-HCFs (linguistic hand-crafted features)
  // feature extraction
  MATLAB{a},PRAAT{a} → SPEECH-LLD (speech low-level descriptors)
  // calculation of speech hand-crafted features
  SpeechProcessor{ SPEECH-LLD } → SPEECH-HCFs (speech hand-crafted features)
  // feature extraction
  OPENFACE{X} → VISUAL-LLD (visual low-level descriptors)
  // calculation of visual hand-crafted features
  VisualProcessor{ VISUAL-LLD } → VISUAL-HCFs (visual hand-crafted features)
  // FHIR
  json.store (TXT-HCFs, SPEECH-HCFs,VISUAL-HCFs) → JSON files → FHIR server
```

Fig. 2. Algorithm used by E2E multi-modal feature extraction framework.

3.3.3. *Speech feature extraction component (SFE)*

Speech features are extracted using the Python library Parselmouth (Boersma and Van Heuven, 2001) (Fig. 5). The subject's speech is extracted as an audio file. And then the audio is used as input to the Parselmouth library to extract audio features. The speech related handcrafted features are implemented by using the speech features described in Fig. 5.

3.3.4. *Visual feature extraction component (VFE)*

Visual features are extracted by using Openface (Baltrusaitis et al., 2016) (Fig. 6). This feature extraction pipeline consists of three main steps: face detection and feature extraction, post processing, and action unit calculation. Feature extraction is started by detecting the face in each video frame. Facial landmarks are detected next by using Conditional Local Neural Fields (CLNF) (Baltrušaitis et al., 2013), which provide 68 points on face (Ekman et al., 2002). Head pose and eye gaze are estimated with the help of 3D facial landmarks as explained in Fig. 6. Once all features have been extracted, a few post-processing steps are applied, such as face alignment, dimensionality reduction, normalisation and feature fusion. Action unit prediction uses linear kernel supported vector machines (SVM) for presence and support vector regression (SVR) for intensity. After extracting all action units, head pose vectors and eye gaze vectors, hand-crafted facial features are designed as explained in Fig. 6.

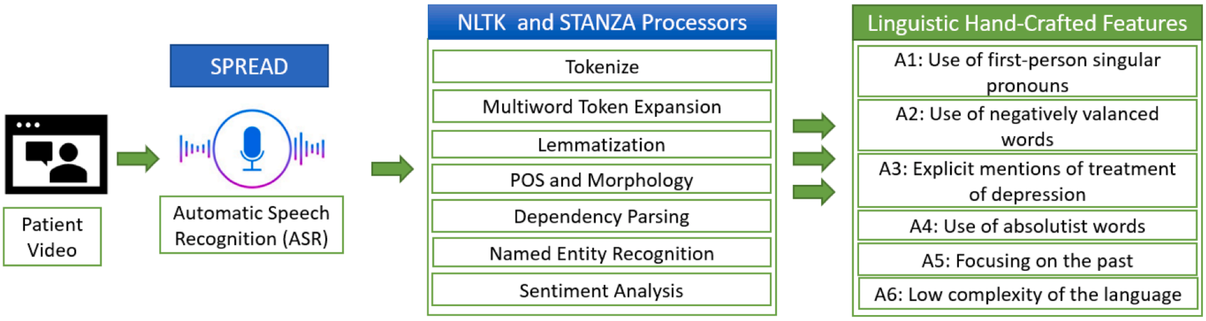
SPREAD
Patient Video
Automatic Speech Recognition (ASR)
NLTK and STANZA Processors
Tokenize
Multiword Token Expansion
Lemmatization
POS and Morphology
Dependency Parsing
Named Entity Recognition
Sentiment Analysis
Linguistic Hand-Crafted Features
A1: Use of first-person singular pronouns
A2: Use of negatively valanced words
A3: Explicit mentions of treatment of depression
A4: Use of absolutist words
A5: Focusing on the past
A6: Low complexity of the language

Fig. 3. Design of the proposed Linguistic Feature Extraction Pipeline.

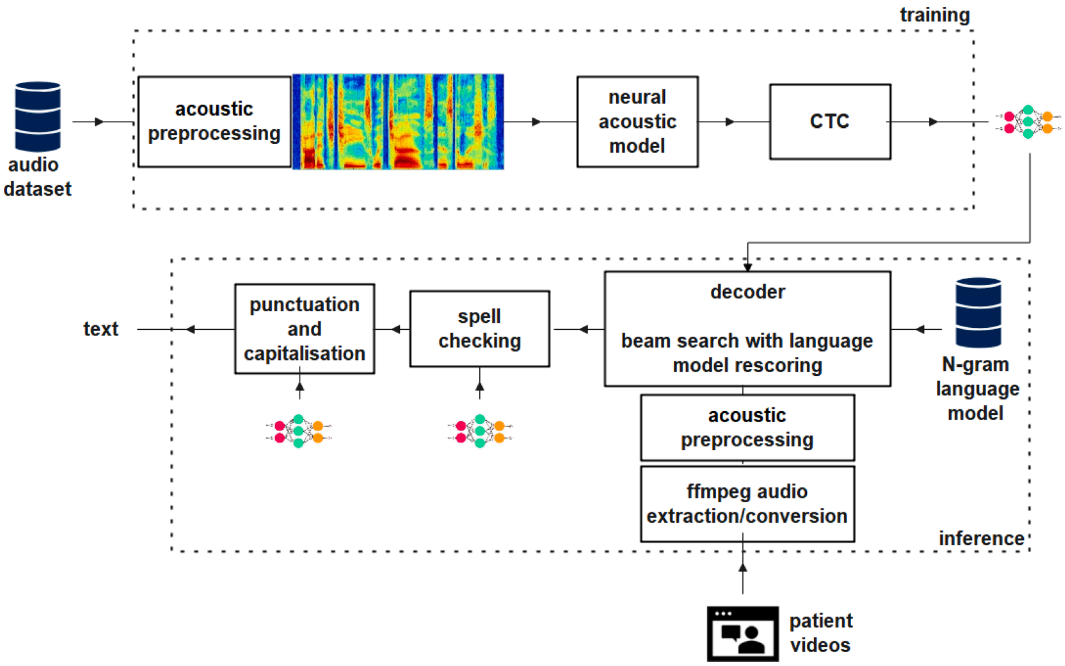

Fig. 4. ASR SPREAD: an end-to-end architecture.

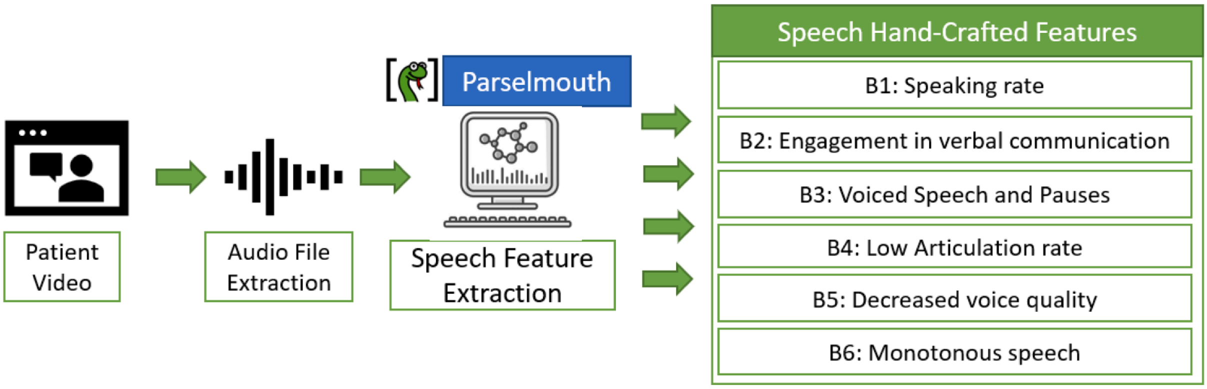
Parselmouth
Patient Video
Audio File Extraction
Speech Feature Extraction
Speech Hand-Crafted Features
B1: Speaking rate
B2: Engagement in verbal communication
B3: Voiced Speech and Pauses
B4: Low Articulation rate
B5: Decreased voice quality
B6: Monotonous speech

Fig. 5. Design of the proposed Speech Feature Extraction Pipeline.

4. Results and discussion

This section presents results from linguistic, speech, and visual features extraction pipelines for the SM and DW datasets, and comparisons of extracted features between those recordings portraying depressive disorder and those without. The *effect sizes* used are independent of the sample size. It is used to indicate the practical significance of the results, as suggested by the APA guidelines (Dauphin, 2020). Pearson's coefficient r in Eq. (1), or the correlation coefficient is used since it can measure the extent of a linear relationship between two variables:

$$r = \frac{n \bullet \sum x \bullet y - (\sum x) \bullet (\sum y)}{\sqrt{\left[n \bullet \sum x^2 - (\sum x)^2\right] \bullet \left[n \bullet \sum y^2 - (\sum y)^2\right]}} \qquad (1)$$

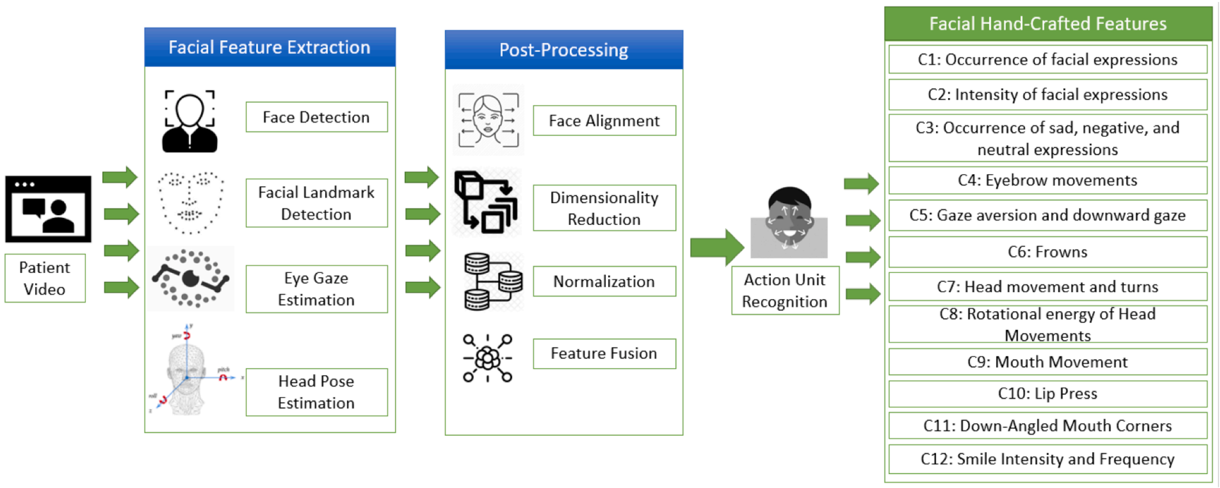
Patient Video
Facial Feature Extraction
Face Detection
Facial Landmark Detection
Eye Gaze Estimation
Head Pose Estimation
Post-Processing
Face Alignment
Dimensionality Reduction
Normalization
Feature Fusion
Action Unit Recognition
Facial Hand-Crafted Features
C1: Occurrence of facial expressions
C2: Intensity of facial expressions
C3: Occurrence of sad, negative, and neutral expressions
C4: Eyebrow movements
C5: Gaze aversion and downward gaze
C6: Frowns
C7: Head movement and turns
C8: Rotational energy of Head Movements
C9: Mouth Movement
C10: Lip Press
C11: Down-Angled Mouth Corners
C12: Smile Intensity and Frequency

Fig. 6. Design of the proposed Visual Feature Extraction Pipeline.

where r represents the strength of the correlation between variables x and y, and n is sample size. The main idea is to compute how much of the variability of one variable is determined by the variability of the other variable. This scale can be used to measure correlations between variables — that makes it unit-free, and to directly compare the strengths of all correlations with each other. Further, the *effect size* is categorized into *small (0.1 to 0.3 or* $-$*0.1 to* $-$*0.3), medium (0.3 to 0.5 or* $-$*0.3 to* $-$*0.5),* or *large* (*0.5 or greater or* $-$*0.5 or less*) according to Cohen's criteria (Lakens, 2013).

The closer r is to 0, the smaller the effect size, while a value of r closer to–1 or 1 indicates a larger effect size. Furthermore, a positive value of r means that both variables either increase or decrease together, while a negative value of r means that one variable increases, when the other decreases (or vice versa).

The aim is to evaluate the 'explanatory power' of the proposed E2E framework for extracting observable cues of depression from diary records in the SymptomMedia (SM) and DAIC-WOZ (DW) datasets. To achieve this goal, we used the proposed framework as outlined in Section 3.2, performed a feature extraction process on the datasets, and compared the extracted features between the datasets through a comprehensive analysis. In the following subsections, we discuss the results for each modality and compare the extracted features with the nature of observable cues of depression as defined by Smrke et al. 2021, which provides our theoretical background.

4.1. The linguistic feature extraction pipeline

Table 2 summarizes the linguistic nature of the inputs as extracted using the *Linguistic Feature Extraction Pipeline*.

The analysis in Table 2 shows that although the number of statements is similar, the statements are generally shorter in recordings from people with depressive disorders compared to those without. In addition, the words used are more varied in the recordings of people without depressive symptoms. These findings were consistent across both databases. However, the lower use of unique words per statement for people with depression applies to the SW dataset. In the DW dataset, the use of unique words is similar, even slightly higher in the recordings of people with depression.

This linguistic cue seems to be under-expressed by actors in the SM dataset and less so in the DW dataset, possibly due to the structured nature of the Q&A sessions. This phenomenon is well illustrated in Table 2, where it can be seen that the discourse in DAIC-WOZ is generated with significantly more responses that are quite short (in terms of words per paragraph). Moreover, most related studies, including Kim et al. (2021), carry out linguistic analysis on written text, whereas the proposed framework analyses the linguistic nature of spoken discourse, which differs both in structure and in contextual influences (Lu et al., 2019).

Table 2
The linguistic nature of the discourse expressed by the subjects in SM and DW recordings.

Observable cues	Symptom Media recordings		DAIC-WOZ recordings	
	Depressive disorder	Without	Depressive disorder	Without
Total number of words (average per recording)	5975 (427)	8234 (588)	14,363 (1196)	27,223 (1815)
Total number of unique use of words per paragraph*	2711	3667	4284	6847
Total number of sentences (average per recording)**	751 (54)	1076 (77)	–	–
Total number of paragraphs (average per recording)	409 (29)	480 (34)	1679 (140)	2789 (186)
Average number of words per paragraph	14.61	17.15	8.55	9.76
Average number of unique words per paragraph**	6.62	7.64	2.55	2.45

Notes: * under a paragraph, a complete answer (i.e., a line in transcription) is considered. ** DW dataset is not transcribed using correct punctuation, but rather in paragraphs, thus paragraph is considered as for comparison.

Table 3 then presents the results of the analysis focusing on *the complexity of language (A6),* i.e. *syntactic complexity (A6.1) (sentence length (A6.1.1), sentence complexity (A6.1.2)) and lexical complexity (A6.2) (lexical diversity (A6.2.1), lexical sophistication (A6.2.2) and lexical density (A6.2.3)).*

The language complexity features extracted by this pipeline show somewhat mixed results between the two datasets. The differences between recordings with and without depressive disorder generally show small effect sizes (i.e. around r = 0.10 (Ellis, 2010)), e.g. the features Sentence Complexity (A6.1.2) and Lexical Density (A6.2.3) show the same pattern of higher/lower mean values between categories of recordings consistently across both datasets. Similarly, the feature Lexical diversity (A6.2.1) is more pronounced in recordings without depressive disorder than with depressive disorder within both datasets, with large effect sizes of differences. For the characteristics Sentence length (A6.1.1) and Lexical sophistication (A6.2.2), mean scores are higher for

Table 3
Language complexity using the Linguistic Feature Extraction Pipeline.

Observable cues	Symptom Media recordings			DAIC-WOZ recordings		
	Depressive disorder	Without	r	Depressive disorder	Without	r
	M (SD)	M (SD)		M (SD)	M (SD)	
A6.1: Syntactic Complexity						
A6.1.1: Sentence length (avg = 24.9)*	8.08 (2.66)	7.70 (1.80)	0.09	9.16 (2.84)	9.92 (3.25)	−0.13
A6.1.2: Sentence complexity (>1.5 for high)*	1.32 (0.18)	1.28 (0.11)	0.14	1.22 (0.09)	1.21 (0.09)	0.08
A6.2: Lexical Complexity						
A6.2.1: Lexical diversity (>0.7 for high)*	0.13 (0.01)	0.14 (0.01)	−0.59	0.14 (0.00)	0.15 (0.01)	−0.41
A6.2.2: Lexical sophistication (>16.25 for high)*	2.77 (0.73)	2.47 (0.53)	0.24	3.29 (1.02)	3.51 (1.18)	−0.11
A6.2.3: Lexical density (>65 for high)*	36.72 (4.79)	36.06 (3.02)	0.09	35.99 (2.58)	35.49 (2.24)	0.11

Notes: M=mean, SD=standard deviation, r = effect size of the difference between recordings of depressive disorder, and those without, for each dataset. * = The values indicate expected result on general population (for more information see Appendix A.1).

recordings with depressive disorder compared to those without in the SM dataset, and vice versa in the DW dataset.

As reported by Smrke et al,2021 people with depression tend to avoid complex sentences in favour of shorter sentences, especially those containing adverbial clauses (e.g. ‘even though’ and ‘as soon as’), compared to those without depression.

The proposed pipeline produces results that are consistent with the expected complexity of language, as reported in Smrke et al. 2021, in only a few instances of the features extracted from the DW dataset. Specifically, individuals with depression tend to produce statements with shorter sentence length (A6.1.1), lower lexical diversity (A6.2.1), and lower lexical sophistication (A6.2.2) than individuals without depression. Table 4.

Other extracted features in Table 4 were not aligned with the theoretical background. We can observe that, regardless of the type of data set, subjects with depressive symptoms in recordings tend to produce discourse with below-average sentence length, low complexity, low diversity, low lexical sophistication and low lexical density; all values were well below the identified average or high values. However, the same is true when these characteristics are observed in recordings without symptoms of depressive disorders. The results in the SM dataset are particularly misaligned with the background. Namely, in the DW dataset, although the differences are small, subjects with symptoms of depressive disorders tend to express language with less complexity than subjects in recordings without depressive disorders. In the SM dataset, complexity is even slightly higher in recordings with depressive disorders, but the overall differences are of small effect size. The main reason for this incoherence could be found in the nature of the datasets, i.e. the discourse in the recordings in the SM dataset is of a ‘diary’ nature, generated by actors, and the recordings without symptoms of depressive disorders do not represent individuals without any psychological distress/condition, so the similarity to real-world data of individuals with depression and without significant psychological conditions is debatable. On the other hand, although the recordings of individuals without depression in the DW dataset were classified as having no symptoms of depression according to the PHQ-8, this does not mean that anxiety or other psychological distress is not present. In addition, the recordings in the DW dataset are of a Q&A nature, i.e. a chatbot guides and drives the interaction. This means that some of the linguistic cues may be under-expressed due to the structured nature of the conversation.

Table 4
Matching linguistic cues against theoretical background.

Observable cue	SM dataset	DW dataset
A6: Language Complexity		
A6.1: Syntactic Complexity		
A6.1.1: Sentence length	✘	✓
A6.1.2: Sentence complexity	✘	✘
A6.2: Lexical Complexity		
A6.2.1: Lexical diversity	✓	✓
A6.2.2: Lexical sophistication	✘	✓
A6.2.3: Lexical density	✘	✘

Notes. ✓ – aligned with background, ✓ – inconsistent with background, ✘ – misaligned with background.

Table 5 presents the results of the analysis focusing on semantic expressivity, i.e., *use of first-person singular pronouns (A1), use of negatively valued words (A2), explicit mention of treatment for depression (A3), use of absolutist words (A4), and focus on the past (A5)*.

The semantic expressivity features of the proposed pipeline show relatively consistent results across the two databases. The use of first-person singular pronouns (A1), the use of negatively valenced words (A2), and the explicit mention of treatment for depression (A3) are all more present in recordings with the depressive disorder than in recordings without the depressive disorder in both datasets, with the differences being in the range of medium to large effect sizes. Similarly consistent are the results for focusing on the past (A5), which is more present in recordings without depressive disorder than in those with depressive disorder in both datasets, with small effect sizes. The results for the use of absolutist words (A4) are somewhat mixed. In particular, in the case of Small grammar, the feature is more present in recordings with the depressive disorder than in those without in both datasets, but in the case of Extended grammar this is true for the SM dataset, while in the DW dataset the mean is slightly higher in recordings without the depressive disorder, although the effect size of this difference is very small.

As reported by Smrke et al.,2021 people with depression tend to: (A1) focus more on themselves and use first-person singular pronouns regardless of age, gender or conversational context, (A2) use more specific linguistic features, such as words with negative valence (e.g. words related to rumination, expressions of sadness and anger, pain and aggression), (A3) explicitly mention the treatment of depression (i.e., (i.e., use words such as “medication”, “side effects”, “treatment”, “therapy”, “pills”, “psychotherapy”), (A4) tend to use absolutist words (i.e., use words such as “nothing”, “always”, “everything”, “never”, “all the time”), (A5) tend to focus on the past.

The pipeline generates linguistic cues that are generally consistent with the theoretical background. Namely, subjects with depression tend to use more first-person singular pronouns than those without depression (Dwyer et al., 2021). Regardless of the dataset and the type of discourse (i.e., diary-style reporting in the SM dataset and Q&A-style reporting in the DW dataset), the use of first-person singular pronouns (cue A1) is higher in recordings with symptoms of depression, as reported in Table 5. A similar trend, consistent with the findings of Smrke et al. 2021, is also observed in explicit mentions of treatment for depression (cue A3), which are more present in subjects with depression compared to those without (Table 6).

Furthermore, the AI-driven analysis performed with the proposed framework is also consistent with the theoretical background on the use of negatively valenced words (cue A2) (Dwyer et al., 2021, Kim et al., 2021). Indeed, we found that subjects with symptoms of depression produced more negatively valanced words than subjects without symptoms in both datasets.

Table 5
Analysing semantic expressivity using the Linguistic Feature Extraction Pipeline.

Observable cues	Symptom Media recordings			DAIC-WOZ recordings		
	Depressive disorder	Without	r	Depressive disorder	Without	r
	M (SD) in %	M (SD) in %		M (SD) in %	M (SD) in %	
A1: Use of first-person singular pronouns	10.55 (2.48)	9.26 (1.76)	0.31	9.59 (1.54)	8.31 (1.56)	0.40
A2: Use of negatively valanced words	5.49 (1.25)	4.64 (1.24)	0.33	3.75 (0.89)	2.42 (0.47)	0.76
A3: Explicit mentions of treatment of depression	0.58 (0.46)	0.15 (0.17)	0.63	0.41 (0.21)	0.13 (0.06)	0.78
A4: Use of absolutist words						
Small grammar*	1.02 (0.49)	0.72 (0.66)	0.27	0.49 (0.36)	0.44 (0.23)	0.11
Extended grammar**	1.85 (0.81)	1.29 (0.72)	0.36	0.94 (0.50)	0.98 (0.34)	−0.05
A5: Focusing on the past	5.44 (3.54)	5.58 (2.67)	−0.02	4.57 (1.71)	4.90 (1.45)	−0.11

Notes. M=mean, SD=standard deviation, r = effect size of the difference between recordings of depressive disorder and those without, within each dataset. *Small grammar = words including “always”, “never”, “all the time”, “everything”, “nothing” from previous studies. Extended grammar = words including “absolutely” “all”, “always”, “complete”, “completely”, “constant”, “constantly”, “definitely”, “entire”, “ever”, “every”, “everyone”, “everything”, “full”, “must”, “never”, “nothing”, “totally”, “whole”^{73}.

Table 6
Matching linguistic cues against theoretical background.

Observable cue	SM dataset	DW dataset
A1: Use of first-person singular pronouns	✓	✓
A2: Use of negatively valanced words	✓✓	✓
A3: Explicit mentions of treatment of depression	✓	✓
A4: Use of absolutist words	✓	?
A5: Focusing on the past	✘	✘

Notes. ✓ – aligned with background, ✓ – inconsistent with background, ✓- misaligned with background.

Other cues extracted by the proposed pipeline show somewhat mixed results, depending on the dataset. For the use of absolutist words (cue A4), we would expect it to be higher in the recordings of subjects with depression (Smrke et al., 2021), which was clearly observed only in the SM dataset for both variations of this cue (i.e. small and extended grammar). In the DW dataset, the use of absolutist words was higher in subjects with depression only in the small grammar variation, whereas in the extended grammar version, subjects without depression used more such words, although the effect size of this difference is very small (Ellis, 2010).

Finally, the limitations related to the contextual nature of the generated discourse may well explain the inconsistency in the use of words focusing on the past (cue A5). Although Kim et al, 2021 report a higher focus on the past as an important linguistic cue for depression, this is not supported by our extracted features. The focus on the past is even slightly more present in recordings of individuals without depression. This linguistic cue appears to be under-expressed by actors in the SM dataset and less so in the DW dataset, possibly due to the structured nature of the Q&A sessions.

4.2. Speech feature extraction pipeline

Table 7 summarizes the results for spectral cues related to the Engagement in verbal communication (B2).

The Speech Feature Extraction Pipeline’s spectral cue features show relatively consistent results for the F0 and F2 formant range features, for which averages are higher in recordings without depressive disorder in both datasets, with small to medium effect sizes, and for the fourth formant (F4) and amplitude difference H1-A3 (F0-F3) features, for which averages are higher in recordings with depressive disorder, again with small to medium effect sizes. For the first (F1), second (F2) and third (F3) formants the results are not consistent across datasets. In the SM dataset, means for these features are higher in recordings without depression, with small to medium effect sizes, whereas in the DW dataset the results are reversed, i.e. means are higher in recordings with depressive disorder compared to those without, again with small to medium effect sizes of differences.

As reviewed by Smrke et al., 2021, people with depression tend to be less engaged in verbal communication compared to those without depression (B2).

Previous reviews and studies suggest that speech is a more favourable modality for classifying depression (Guohou et al., 2020), as it is generally an even richer source of information than written language. Formants have been suggested as promising observable cues for the classification of psychological distress. Namely, formants are the primary resonances of the vocal tract and can shift due to numerous conscious and unconscious processes, depending on one’s speaking style (Shahin et al., 2001). The change in F0 (i.e. the fundamental frequency) and the shifts in formant F1 and formant F2 have been shown to be the more significant features. Overall, research indicates a universal trend towards an increase in F0 in stressed subjects (Giddens et al., 2013, Kappen et al., 2022). When analysing recordings with the proposed pipeline, we found mixed results (Table 8). F0 was higher in subjects without depressive disorder in both datasets. F1, F2 and F3 were higher in individuals with depressive disorder, but only in the DW dataset, whereas in the SM dataset they were higher in individuals without symptoms of depression. While it would be expected that F1 and F2 would be higher in those with depression, there is no consensus about

Table 7
Spectral cues related to the Engagement in verbal communication using the Speech Feature Extraction Pipeline.

Observable cues	Symptom Media recordings			DAIC-WOZ recordings		
	Depressive disorder	Without	r	Depressive disorder	Without	r
	M (SD)	M (SD)		M (SD)	M (SD)	
B2: Engagement in verbal communication						
F0 (Hz)	144.54 (33.79)	166.77 (41.21)	−0.30	159.80 (38.49)	147.05 (37.52)	0.18
First formant (F1) (Hz)	499.90 (43.17)	534.39 (60.49)	−0.34	543.79 (39.12)	536.37 (72.82)	0.07
Second formant (F2) (Hz)	1547.83 (59.56)	1588.17 (77.27)	−0.30	1561.29 (86.80)	1526.15 (83.98)	0.21
Third formant (F3) (Hz)	2509.99 (67.95)	2525.05 (81.57)	−0.11	2527.63 (113.34)	2476.59 (107.16)	0.24
Fourth formant (F4) (Hz)	3555.17 (73.28)	3548.23 (77.59)	0.05	3643.82 (220.75)	3506.52 (191.66)	0.34
Amplitude difference H1-A3 (F0-F3) (Hz)	2365.45 (46.60)	2358.28 (64.57)	0.07	2527.63 (113.34)	2476.59 (107.16)	0.24
Range of F2 formant (Hz)	2866.28 (156.60)	2887.91 (265.80)	−0.06	2631.46 (437.14)	2905.84 (358.26)	−0.35

Notes. M=mean, SD=standard deviation, r = effect size of the difference between recordings of depressive disorder and those without, within each dataset.

Table 8
Matching spectral cues related to the Engagement in verbal communication against theoretical background.

Observable cue	SM dataset	DW dataset
B2: Engagement in verbal communication	?	✓

Notes. ✓ – aligned with background, ✓ – inconsistent with background, ✘ – misaligned with background.

their 'universal' nature. The effects of psychological stress on F1 and F2 appear to be influenced by speech style and individual context rather than a global trend (Sigmund, 2012). The remaining sub-features, i.e. F4, amplitude difference H1-A3 (F0-F3) and F2 formant range, were consistent across the datasets, with the first two being higher in individuals with depression and the last being higher in individuals without depression.

Table 9 shows the prosodic observable cues, i.e. engagement in verbal communication (B2; pitch, jitter, intensity) and voice quality cues, i.e. reduced voice quality (B5; absolute jitter, shimmer and harmonics-to-noise ratio (HNR)).

For the feature Engagement in verbal communication (B2), the results extracted via the Speech Feature Extraction Pipeline are consistent only for Intensity, i.e., the means are higher for recordings with depressive disorder than for those without in both datasets with a small effect size of difference in DW and with a medium effect size in SM data. For pitch, the means are higher for recordings without depressive disorder in the SM dataset with a medium effect size of the difference, and higher for recordings with depressive disorder in the DW dataset with a small effect size. The opposite is true for local jitter, where the means are higher for recordings with depressive disorder in the SM dataset, and higher for recordings without depressive disorder in the DW dataset, with a medium effect size of differences in both cases.

The results for Decreased voice quality (B5) are more consistent. For Local shimmer (in % and dB versions) the means are higher for recordings without depressive disorder in both datasets, with small to medium effect sizes of differences in both cases. Harmonic-to-noise ratio averages higher in recordings with depressive disorder than in those without in both datasets, again with small to medium effect sizes of differences. The results for local absolute jitter differ between the datasets, i.e. the average is higher for recordings with depressive disorder in the SM dataset and for recordings without depressive disorder in the DW dataset, while the differences are in the medium effect size range in both cases.

As reviewed by Smrke et al. 2021, people with depressive disorder tend to produce harsher speech with higher vocal tension and lower quality compared to people without depressive disorder.

Studies on the effects of distress on voice production have primarily focused on voice quality, with results showing that depression is associated with reduced voice quality (e.g., aspiration, jitter, shimmer, and breathy phonation) and that the voices of depressed individuals are generally harsher (Cummins et al., 2015). In terms of voice quality (cue B5), we observed jitter, shimmer, and harmonic-to-noise ratio (Table 9). While jitter shows the fundamental frequency perturbation, shimmer shows the amplitude perturbation of the voice. Higher values of jitter and shimmer have been observed in depressed individuals (Silva et al., 2021). While all three observed sub-features showed slightly higher absolute values in the SM dataset compared to the DW dataset, both absolute local jitter (cue B5) and local jitter (cue B2) were higher in individuals with depressive disorder compared to those without in the SM dataset, and vice versa in the DW dataset. The results of the remaining sub-features are more consistent between the datasets, i.e. shimmer is lower in individuals with depressive disorder in both cases, and harmonics-to-noise ratio is higher in this group compared to individuals without depressive disorder. Overall, harmonics-to-noise ratio, jitter and shimmer have not been shown to be universal (Silva et al., 2021) and have reproduced mixed results in the context of cognitive load/psychological stress. Furthermore, there are many acoustic factors, such as recording parameters and quality, distance of the subject from the microphone, loudness of the speech, etc., that have a significant impact on highly sensitive features such as jitter, shimmer and HNR, and as such are less robust than, for example, features such as voiced/unvoiced speech and pauses during speech (Table 10).

Table 11 summarises the observable cues related to some prosodic features (voiced-unvoiced-silent parts, pauses, duration), i.e., speaking rate (B1), voiced speech and pauses (B3), low articulation rate (B4) and monotonous speech (B6).

The results regarding speech cues in terms of quantity and duration of vocal elements are very consistent in both datasets. Speech rate (B1; Speech rate) is higher in recordings without depressive disorder, with medium to large effect sizes of differences. Similarly, for Voiced speech and pauses (B3), recordings without depressive disorder have a higher number of voiced syllables, number of pauses (silence), and balance of speech between voiced and unvoiced segments than the recording without, with predominantly medium effect sizes of differences. Similarly, for the feature Low articulation rate (B4), the articulation rate is higher in recordings without depressive disorder in both SM and DW data sets, with large and small effect sizes of differences, respectively.

For the feature Monotonous speech (B6), recordings with depressive disorders have more pauses (silences) (both in sec and % versions) and longer duration of spoken syllables than recordings without depressive disorders in both datasets, with medium to large effect sizes of

Table 10
Matching prosodic and voice quality cues related to the Engagement and Quality of verbal communication against theoretical background.

Observable cue	SM dataset	DW dataset
B2: Engagement in verbal communication	?	✓
B5: Decreased voice quality	?	?

Notes. ✓ – aligned with background, ✓ – inconsistent with background, ✘ – misaligned with background.

Table 9
Prosodic and voice quality cues related to Engagement and Quality of verbal communication using the Speech Feature Extraction Pipeline.

Observable cues	Symptom Media recordings			DAIC-WOZ recordings		
	Depressive disorder	Without	r	Depressive disorder	Without	r
	M (SD)	M (SD)		M (SD)	M (SD)	
B2: Engagement in verbal communication						
Pitch (Hz)	150.51757 (35.16333)	175.06171 (45.55554)	−0.31	161.71342 (38.62093)	149.71620 (37.78544)	0.16
Local jitter (%)	0.02236 (0.00356)	0.02071 (0.00216)	0.30	0.01450 (0.00350)	0.01660 (0.00540)	−0.24
Intensity (Db)	40.55071 (4.29640)	36.28843 (7.91860)	0.37	49.46458 (4.97600)	49.22040 (4.37559)	0.03
B5: Decreased voice quality						
Local absolute jitter (sec)	0.00017 (0.00006)	0.00013 (0.00004)	0.35	0.00010 (0.00005)	0.00013 (0.00006)	−0.24
Local shimmer (%)	0.12571 (0.01542)	0.13114 (0.01123)	−0.21	0.07208 (0.01574)	0.07927 (0.01738)	−0.22
Local dB shimmer (dB)	1.19400 (0.12285)	1.23007 (0.09239)	−0.18	0.67625 (0.16521)	0.74633 (0.17961)	−0.21
Harmonics-to-noise ratio (HNR)	10.71571 (2.15857)	10.35021 (1.22188)	0.12	13.83333 (2.28963)	12.80000 (2.27408)	0.23

Notes. M=mean, SD=standard deviation, r = effect size of the difference between recordings of depressive disorder and those without, within each dataset.

Table 11
Speech cues related to quantity and duration of vocal elements using the Speech Feature Extraction Pipeline.

Observable cues	Symptom Media recordings			DAIC-WOZ recordings		
	Depressive disorder	Without	r	Depressive disorder	Without	r
	M (SD)	M (SD)		M (SD)	M (SD)	
B1: Speaking rate						
Speech rate	1.56 (0.48)	1.99 (0.43)	−0.44	1.28 (0.64)	1.63 (0.63)	−0.28
B3: Voiced speech and pauses						
Number of voiced syllables	46.32 (14.33)	58.72 (13.15)	−0.43	38.28 (19.20)	48.74 (19.02)	−0.28
Number of pauses (silence)	8.95 (1.94)	10.62 (2.45)	−0.37	8.11 (1.86)	8.74 (2.51)	−0.15
Speaking balance	0.31 (0.10)	0.38 (0.09)	−0.35	0.39 (0.09)	0.47 (0.11)	−0.36
B4: Low articulation rate						
Articulation rate	4.10 (0.31)	4.36 (0.19)	−0.50	2.84 (1.35)	3.04 (0.86)	−0.11
B6: Monotonous speech						
Speaking duration (without pauses) (sec)	9.25 (3.06)	11.21 (2.70)	−0.33	11.72 (2.76)	13.96 (3.28)	−0.36
Pausing (silence) duration (sec)	20.28 (3.05)	18.14 (2.75)	0.36	18.16 (2.60)	15.97 (3.30)	0.36
Speaking duration (%)	31.30 (10.30)	38.21 (9.06)	−0.35	39.17 (9.09)	46.66 (10.99)	−0.36
Pausing (silence) duration (%)	68.70 (10.30)	61.79 (9.06)	0.35	60.84 (9.09)	53.34 (10.99)	0.36
Duration of voiced syllables (sec)	0.25 (0.02)	0.23 (0.11)	0.47	0.77 (1.09)	0.38 (0.21)	0.33

Notes. M=mean, SD=standard deviation, r = effect size of the difference between recordings of depressive disorder and those without, within each dataset.

differences. Recordings without depressive disorders tend to have longer speaking duration (excluding pauses and in %) than recordings with depressive disorders in both datasets, with medium effect sizes of differences.

As reviewed by Smrke et al. 2021, people with depressive disorders, compared to those without, tend to: (B1) speak at a slower rate, (B3) have longer pauses and unvoiced speech when answering questions, during conversation, and when initiating speech, (B4) articulate with lower articulation rates and flatter trajectories (i.e., fewer articulatory movements), and (B6) generate less pitch variability, resulting in a lack of linguistic stress and reduced loudness variation.

As shown in Table 12, we can observe that in both datasets, subjects with depressive symptoms in recordings tend to produce speech with a lower speaking rate (cue B1) and a lower articulation rate (cue B4) compared to subjects without depressive symptoms in recordings, supporting the findings of previous studies (Smrke et al., 2021). In addition, individuals with depression tend to incorporate longer, even inappropriate pauses and produce more monotonous speech (i.e., cues B3 and B6) when answering questions, in conversation, and at speech onset (Cummins et al., 2015, Puyvelde et al., 2018). As shown in Table 6-1, in both datasets, subjects in recordings with depressive symptoms tend to produce longer pauses than subjects in recordings without depressive symptoms. In addition, spoken segments are generally shorter, and pauses tend to make up a higher proportion of speech in recordings of people with depression than in recordings of people without symptoms. In addition, regardless of the type of discourse (i.e., SM vs. DW dataset), subjects with depressive symptoms in recordings produce fewer voiced syllables compared to subjects without symptoms, but the voiced syllables tend to be longer. In general, subjects with depressive disorders produced more monotonous speech (cue B6) than subjects without depression in both datasets, which was already found in the previous studies. That is, depression has been associated with a lack of linguistic stress (e.g., the relative emphasis given to a particular syllable or word), reduced intonation (Cummins et al., 2015), repetitive pitch inflections and stress patterns (Guohou et al., 2020), and poorer articulation (particularly in terms of diphthong production, which refers to a sound produced by combining two vowels) (Yamamoto et al., 2020).

Table 12
Matching observable cues from speech against theoretical background.

Observable cue	SM dataset	DW dataset
B1: Speaking rate	✓	✓
B3: Voiced speech and pauses	✓	✓
B4: Low articulation rate	✓	✓
B6: Monotonous speech	✓	✓

Notes. ✓ – aligned with background, ✓ – inconsistent with background, ✘ – misaligned with background.

4.3. Results for visual extraction pipeline

Table 13 summarize the visual cues related to *Facial emotional expressivity,* i.e., *Occurrence of facial expressions (C1), Intensity of facial expressions (C2),* and *Occurrence and emotional variability (C3)* for the SM and DW datasets.

The results for visual cues are related to facial emotional expressivity and are generally inconsistent between the two datasets. The consistent result is, for example, for the occurrence of facial expressions (C1), i.e., for the sub-features Emotion – Anger and Emotion – Disgust, where a higher percentage of these emotions are detected in the recordings with depressive disorders than in the recordings without in both datasets, with small to medium effect sizes of the differences. Other emotions within C1, i.e., Surprise, Fear, Happiness and Sadness, are more present in the recordings without depressive disorders than in those with depressive disorders for the SM dataset, and vice versa for the DW dataset, again with small to medium effect sizes of differences for both datasets. Similar results are observed for the intensity of facial expressions (B2), where all emotions tend to be perceived as more intense in recordings without depressive disorders than in those with in the SM dataset, and exactly the opposite in the DW dataset, i.e., they are perceived as more intense in recordings with depressive disorders, again with small to medium effect sizes of differences for both datasets.

Similarly, inconsistencies persist in the feature “occurrence and emotional variability” (C3). Positive and negative emotions, i.e., their percentage of occurrence and intensity, show higher averages for recordings without depression in the SM dataset and higher averages for recordings with depressive disorders in the DW dataset. The same is true for the total number of emotion variabilities, for which the mean is higher for the recordings without depressive disorders in the SM dataset, and higher for the recordings with depressive disorders in the DW dataset. All effect sizes of differences between categories of recordings for C3 traits range from small to medium.

As reviewed by Smrke et al., 2021 people with depression, compared to those without, tend to: (C1) express fewer facial expressions with reduced facial mobility, (C2) express emotions with reduced facial expressivity, (C3) present more of negative, sad, and neutral expressions.

The least conclusive results are generated by the pipeline for the extraction of visual cues (Table 14). Namely, 5 out of 12 features generated by the proposed pipeline observed in the DW dataset were completely and 4 of them partially misaligned with the trends observed in the literature. When evaluating the pipeline on the recordings in the SM dataset, the results are only slightly better, i.e. one feature was completely misaligned and 7 of them partially misaligned with previous

Table 13
Visual cues related to facial emotional expressivity using the Visual Extraction Pipeline.

Observable cues	Symptom Media recordings			DAIC-WOZ recordings		
	Depressive disorder	Without	r	Depressive disorder	Without	r
	M (SD)	M (SD)		M (SD)	M (SD)	
C1: Occurrence of facial expressions (Frame (%))						
Emotion – Surprise	1.23 (1.36)	1.77 (1.31)	−0.20	1.31 (1.85)	0.55 (0.62)	0.36
Emotion – Anger	0.053 (0.08)	0.049 (0.07)	0.03	1.52 (3.96)	0.82 (1.15)	0.16
Emotion – Fear	2.18 (1.62)	2.84 (2.10)	−0.19	12.33 (9.55)	10.71 (7.06)	0.11
Emotion – Happiness	5.80 (9.33)	10.67 (17.40)	−0.20	11.66 (9.62)	9.34 (6.39)	0.17
Emotion – Sadness	8.12 (6.38)	10.82 (5.76)	−0.23	20.62 (6.91)	18.11 (9.41)	0.16
Emotion – Disgust	25.00 (17.51)	23.46 (24.84)	0.04	20.49 (14.60)	15.71 (8.68)	0.24
C2: Intensity of facial expressions (Intensity (0 – 5 point))						
Emotion – Surprise	1.79 (0.64)	1.81 (0.22)	−0.03	1.81 (0.35)	1.79 (0.22)	0.04
Emotion – Anger	0.67 (0.81)	0.87 (0.69)	−0.14	0.86 (0.36)	0.78 (0.26)	0.15
Emotion – Fear	1.42 (0.24)	1.55 (0.20)	−0.31	1.50 (0.31)	1.36 (0.12)	0.37
Emotion – Happiness	1.36 (0.44)	1.37 (0.46)	−0.01	1.45 (0.19)	1.37 (0.16)	0.23
Emotion – Sadness	1.61 (0.18)	1.66 (0.18)	−0.14	1.77 (0.35)	1.58 (0.20)	0.39
Emotion – Disgust	0.88 (0.13)	0.97 (0.18)	−0.30	0.83 (0.12)	0.78 (0.13)	0.23
C3: Occurrence and emotional variability						
Positive emotions (Frame (%))	5.80 (9.33)	10.67 (17.40)	−0.20	11.66 (9.62)	9.34 (6.39)	0.17
Negative emotions (Frame (%))	36.57 (18.77)	38.93 (30.05)	−0.05	56.26 (27.00)	45.90 (20.93)	0.24
Positive emotions (Intensity (0 – 5 point))	1.36 (0.44)	1.37 (0.46)	−0.01	1.45 (0.19)	1.37 (0.16)	0.24
Negative emotions (Intensity (0 – 5 point))	1.14 (0.29)	1.26 (0.23)	−0.24	1.36 (0.16)	1.26 (0.11)	0.39
Total number of emotion variability	4.43 (0.94)	4.93 (1.00)	−0.26	4.33 (0.49)	4.27 (0.46)	0.08

Notes. M=mean, SD=standard deviation, r = effect size of the difference between recordings of depressive disorder and those without, within each dataset.

Table 14
Matching visual cues related to facial emotional expressivity against theoretical background.

Observable cue	SM dataset	DW dataset
C1: Occurrence of facial expressions	?	✘
C2: Intensity of facial expressions	✓	✘
C3: Occurrence and emotional variability	?	?

Notes. ✓ – aligned with background, ✓ – inconsistent with background, ✘ – misaligned with background.

studies. In the remaining cases, our pipeline produces results consistent with previous findings (Smrke et al. 2021).

The intensity of facial expressions (cue C2) produces inconsistent results, i.e. valid, or at least partially valid, for the ‘artificial’ actor-based dataset (i.e. the SM dataset) and mostly not valid for the more natural dataset (i.e. the DW dataset). Namely, in line with Bylsma and colleagues (Bylsma et al., 2008), positive and negative emotions in both datasets (for both categories of recordings) tend to be slightly more pronounced (i.e. generated with a higher overall intensity). Within the SM dataset, all emotions tend to be more intense in recordings from individuals without depression, whereas in the DW dataset they are all more intense in recordings from individuals with depressive disorder.

In general, individuals with depression showed less intense facial expressions (cue C2) in the SM dataset (but not in the DW dataset) compared to those without symptoms of depression. Also, although we would expect the occurrence of facial expressions (cue C1) and their emotional variability (cue C3) to be lower in individuals with depression than in those without (Smrke et al., 2021), the results extracted by our proposed pipeline were mixed or inconsistent with the previous literature.

Table 15 summarises the visual cues related to gaze and head movement, i.e., *gaze aversion and downward gaze (C5), head movement and turns (C7), and rotational energy of head movements (C8).*

These results are more consistent across databases than the results for facial emotional expressivity. For the feature Head movement and turns (C7), there are more Head moving frames and Head facing away (left/right and z direction) frames extracted from the recordings without depressive disorder compared to those with depressive disorder in both datasets, with small (for Head facing away (left/right) frames) to even large effect sizes (for Head facing away (z direction) frames in the DW dataset). The exceptions for the C7 feature are Head down frames and Head facing away (up/down) frames, which are more present in the nondepressed recordings in the SM dataset and more present in the

Table 15
Visual cues related to gaze and head movement using the Visual Extraction Pipeline.

Observable cues	Symptom Media recordings			DAIC-WOZ recordings		
	Depressive disorder	Without	r	Depressive disorder	Without	r
	M (SD)	M (SD)		M (SD)	M (SD)	
C5: Gaze aversion and downward gaze						
Looking down frames	49.02 (1.15)	48.02 (1.90)	0.34	47.70 (1.49)	48.24 (1.63)	−0.18
C7: Head movement and turns						
Head moving frames	36.97 (10.36)	43.60 (15.51)	−0.27	36.44 (16.51)	40.64 (15.99)	−0.14
Head down frames	2.04 (3.71)	7.69 (16.62)	−0.31	13.44 (27.09)	5.74 (11.52)	0.24
Face moving frames	45.41 (13.91)	38.59 (21.14)	0.21	10.45 (7.37)	11.61 (7.86)	−0.08
Head facing away (up/down) frames	62.77 (23.17)	69.25 (26.81)	−0.13	4.30 (7.69)	2.01 (1.83)	0.28
Head facing away left/right) frames	50.20 (25.16)	51.74 (24.34)	−0.03	23.62 (27.18)	26.88 (32.79)	−0.06
Head facing away (z direction) frames	8.88 (8.96)	14.99 (15.78)	−0.27	0.97 (0.90)	3.33 (5.16)	−0.41
C8: Rotational energy of head movements						
Frames below rotational threshold (<50 J)	79.24 (11.51)	69.99 (15.26)	0.35	64.94 (17.22)	61.74 (17.40)	0.10
Low rotational energy frames (>= 50 J, < 500 J)	16.17 (7.14)	21.17 (8.40)	−0.32	26.85 (11.17)	27.94 (9.60)	−0.06
Moderate rotational energy frames (>= 500 J, < 5000 J)	3.99 (4.07)	7.06 (6.29)	−0.31	7.45 (6.19)	8.93 (7.27)	−0.11
High rotational energy frames (>= 5000 J)	0.44 (0.86)	1.04 (1.23)	−0.29	0.76 (0.78)	1.39 (2.07)	−0.24

Notes. M=mean, SD=standard deviation, r = effect size of the difference between recordings of depressive disorder and those without within each dataset.

depressed recordings in the DW dataset, and Face moving frames, which are more present in the depressed recordings in the SM dataset and more present in the nondepressed recordings in the DW dataset. Similarly inconsistent is the sub-feature Gaze aversion and downward gaze (C5) of Looking down frames, which is more present in the depressive disorder recordings in the SM dataset, and more present in the without depressive disorder recordings in the DW dataset.

On the other hand, all sub-features of rotational energy of head movements (C8) are expressed in the same pattern in both data sets. Frames below the rotation threshold are more present in the recordings with depressive disorder than in those without, with small (for the DW dataset) and medium (for the SM dataset) effect sizes of the differences. Low, moderate and high rotational energy frames are more present in the recordings without depressive disorders than in those with depressive disorders, with small to medium effect sizes of differences in both datasets.

In general, people with depression tend to look less at each other and limit eye contact. This results in (C5) gaze aversion and downward gaze with increased turning and head movements (C7), i.e. head looking away from the collocutor, and (C8) increased large head movements with lower rotational energy.

Head movements and turns (cue C7) produce inconsistent results, i. e., valid or at least partially valid for the 'artificial' actor-based dataset (i.e. the SM dataset) and mostly not valid for the more natural dataset (i. e. the DW dataset). Similarly, the results are inconclusive for the notion of head movement (C7), i.e., people with depression tend to turn their head away and are more likely to hold their head in a downward position (Guohou et al., 2020, Smrke et al., 2021).The results generated by the proposed pipeline highlight that in both datasets (Table 16), consistent with the notion of reduced expressiveness, subjects with depressive symptoms tend to generate less head movement. The most consistent differences in movement in both datasets can be observed in the direction of the X-axes (i.e. left–right) and in the z-direction, where those with depression tend to show less such movement than those without. However, other sub-features of head movement are less conclusive as they are not consistent across the datasets.

A similar pattern can also be observed for gaze aversion (cue C5). Namely, in the SM dataset the proportion of frames in which gaze aversion is observed tends to be higher in recordings with individuals with depressive symptoms, whereas in the DW dataset the proportion of such frames is higher in recordings with individuals without depressive symptoms. However, these results also support the issue of the controlled context and controlled environment of the recordings, which has been taken into account in most research on the classification of depression and psychological distress (Kappen et al., 2022). In other words, in recordings of vocal actors performing stressful monologues, such as the SM dataset, controlled and enacted features are more pronounced, more universal, and more consistent with the theoretical context than in cues generated by individuals truly experiencing psychological stress (e.g., the DW dataset) (Giddens et al., 2013). Another clear example of this phenomenon is smiling. The results in actor-generated video s (i.e., SM dataset) are clearly consistent with the general notion that depression is associated with less smiling (Cummins et al., 2015, Pampouchidou et al., 2019), whereas in the DW dataset, it was observed that subjects with symptoms of depression smile more than those without symptoms. However, the average intensity of smiles tends to be slightly higher in both datasets, which is consistent with the notion that depressed individuals use smiles as an external I to mask sadness and depression (Draucker, 2005).

Table 16
Matching visual cues related to gaze and head movement against theoretical background.

Observable cue	SM dataset	DW dataset
C5: Gaze aversion and downward gaze	✓	?
C7: Head movement and turns	?	?
C8: Rotational energy of head movements	✓	✓

Notes. ✓ – aligned with background, ✓ – inconsistent with background, ✘ – misaligned with background.

It has already been shown that individuals with symptoms of depression also produce less energetic movements with a higher frequency of small (energetically low) movements (Sandmeir et al., 2021), further emphasising the reduced expressivity and variability as already observed in speech (both acoustic and linguistic patterns), which is consistent with the results of the rotational energy of head movements (cue C8) by our proposed pipeline in both datasets.

Finally, Table 17 summarises the results of the comprehensive analysis of facial expressions and movements.

Overall, in the SM dataset, cues occur with lower intensity in recordings with symptoms of depressive disorders. In the DW dataset, however, the intensity is slightly higher in recordings with symptoms of depressive disorders.

The results for facial expression and movement cues are again of mixed consistency between the two datasets. While the percentages of frames with eyebrow movements (C4) and frowns (C6) are on average higher in recordings with depressive disorders in both datasets, with generally small effect sizes of differences, the intensity of these same features is higher in recordings without depressive disorders in the SM dataset, and higher in recordings with depressive disorders in the DW dataset, again with generally small effect sizes of differences. Also consistent in both datasets is the intensity of smiling (C12), which is higher in recordings with depressive disorder, with medium effect sizes of differences, while the average percentages of frames for this feature are higher for recordings without depression in the SM dataset, and for those with depression in the DW dataset, with generally small effect sizes of differences for both datasets, except for the effect size for frame (0.31-medium effect size) in the SM dataset.

Furthermore, the average percentage of frames and the intensity of Mouth movement (C9) and Down-angled mouth corners (C11) are higher for recordings without depressive disorder in the SM dataset, whereas they are higher for recordings with depressive disorder in the DW dataset.

The Lip Press (C10) feature was not available for extraction using the proposed pipeline from the DW dataset, and in the SM dataset this feature was present in less than 0.1 % of frames with almost no variability between video s, so the comparison between recording categories was not performed in this study. The intensity of lip presses (C10) was slightly higher in the recordings without depression compared to those with depression.

As reviewed by Smrke et al. 2021, people with depression tend to show fewer eyebrow movements (C4) compared to those without depression. Associations between depression and facial expressions, such as (C9) fewer mouth movements, (C10) frequent lip pressing, (C11) frequent shapes with the corners of the mouth angled downwards, and (C12) less intense smiling, have also been reported at lower frequencies and for shorter durations. People with depression generally tend to look at each other less and limit eye contact. This results in a downward gaze with more (C6) frowning, with the head turned away from the collocutor.

In addition, it has been observed that individuals with symptoms of depression produce more frowning (cue C6) (Smrke et al., 2021), which is only fully supported by the features extracted in the DW dataset.

Some of the features, such as the intensity of smiling expressions (cue C12) and mouth movement (cue C9), produce inconsistent results, i.e., valid or at least partially valid for the 'artificial' actor-based dataset (i.e., the SM dataset) and mostly not valid for the closer-to-nature dataset (i. e., the DW dataset). Namely, in line with Bylsma and colleagues (Bylsma et al., 2008), positive emotions and negative emotions tend to be slightly more pronounced (i.e., generated with a higher overall intensity) in both datasets (for both categories of recordings). Within the SM dataset, all emotions tend to be more intense in recordings from individuals without

Table 17
Visual cues related to facial expressions and movements in SymptomMedia Dataset using the Visual Extraction Pipeline.

Observable cues	Symptom Media recordings			DAIC-WOZ recordings		
	Depressive disorder	Without	r	Depressive disorder	Without	r
	M (SD)	M (SD)		M (SD)	M (SD)	
C4: Eyebrow movements						
Frame (%)	39.39 (27.12)	31.75 (24.01)	0.15	38.72 (15.11)	33.38 (13.60)	0.20
Intensity (0 – 5 point)	0.71 (0.12)	0.77 (0.13)	−0.21	0.83 (0.19)	0.81 (0.13)	0.05
C6: Frowns						
Frame (%)	32.20 (31.47)	23.29 (24.92)	0.16	19.66 (11.83)	15.62 (10.41)	0.19
Intensity (0 – 5 point)	1.55 (0.27)	1.56 (0.27)	−0.02	1.61 (0.27)	1.51 (0.44)	0.14
C9: Mouth movement						
Frame (%)	33.35 (14.23)	41.42 (21.37)	−0.24	54.19 (12.28)	52.59 (12.83)	0.07
Intensity (0 – 5 point)	0.47 (0.09)	0.51 (0.10)	−0.21	0.54 (0.17)	0.45 (0.07)	0.39
C10: Lip press						
Frame (%)	< 0.1 %	< 0.1 %	/	/	/	/
Intensity (0 – 5 point)	0.07 (0.03)	0.08 (0.04)	−0.11	/	/	/
C11: Down-angled mouth corners						
Frame (%)	4.12 (3.56)	4.95 (4.40)	−0.11	13.76 (10.80)	11.97 (7.66)	0.11
Intensity (0 – 5 point)	0.82 (0.12)	0.87 (0.15)	−0.20	0.79 (0.15)	0.76 (0.08)	0.15
C12: Smiling						
Frame (%)	1.47 (2.05)	5.68 (13.10)	−0.31	4.88 (5.94)	3.27 (3.40)	0.20
Intensity (0 – 5 point)	2.22 (3.12)	1.32 (0.74)	0.27	1.50 (0.13)	1.38 (0.42)	0.24

Notes. M=mean, SD=standard deviation, r = effect size of the difference between recordings of depressive disorder and those without within each dataset.

depression, whereas in the DW dataset they are all more intense in recordings from individuals with depressive disorder.

We observed a large deviation from previous literature in the case of cues that require the AI to analyse images at a high level of detail, i.e., eyebrow movements (cue C4), lip pressing (cue C10), and down-slanting corners of the mouth (cue C11; which was in line with previous literature in DW recordings, but with small effect sizes of the expected differences). As highlighted in Table 18, the results generated by our pipeline are not consistent with the theoretical background. Lip pressing was observed in less than 0.1 % of the SM frames. Due to the lack of actual video recordings and relevant features, we were unable to extract this behavioural cue in the DW dataset. The inconsistency may be best explained by the quality of the video recordings, as the features considered relevant in previous literature are highly dependent on the context and the quality of the video recording.

4.4. Validation of the pipeline by human observers

By harnessing multiple modalities, a more holistic understanding of an individual's mental state can be achieved, facilitating a more precise classification of mental distress. Scientific studies have demonstrated that amalgamating features from diverse modalities outperforms the utilization of any singular modality alone (Bayoudh et al., 2022). However, the scarcity of available multimodal datasets poses a substantial obstacle in optimizing the performance of the AI models for the classification of depression. The initial assessment of the overall pipeline's performance was conducted in Arioz et al. 2022, utilizing a limited set of 56 recordings, thereby constraining the generalizability of the pipeline. Moreover, the divergence between the SM dataset, enacted by actors without genuine psychological stress, possibly exaggerates certain features, while the DW datasets, controlled by a chatbot and featuring everyday subjects, may downplay critical features, introducing a layer of complexity. To address this, an observational study had been undertaken to scrutinize how human experts interpret cues and to compare their observations with those made by the pipeline, aiming to enhance the robustness and interpretability of the classification model.

Table 18
Matching visual cues related to gaze and head movement against theoretical background.

Observable cue	SM dataset	DW dataset
C4: Eyebrow movements	?	✘
C6: Frowns	?	✓
C9: Mouth movement	✓	✓
C10: Lip press	?	✘
C11: Down-angled mouth corners	✘	✓
C12: Smiling	?	✘

Notes. ✓ – aligned with background, ✓ – inconsistent with background, ✘ – misaligned with background.

4.4.1. Method

Raters. In the human observation part of the study, five raters participated. All were female, had completed bachelor's degree in psychology and were at the time of the study students of masters's program of Psychology at Faculty of Arts, University of Maribor. They were invited to participate in the study via e-mail.

Materials and procedure. For feature extraction via AI pipeline, presented in this paper, as well as for human observation part, a set of 20 SymptomMedia video s was selected. 15 of them were video s of individuals portraying one of the diagnoses of depression (i.e., major depressive disorder (MDD), MDD with recurrent severe episode, MDD with anxious distress, MDD with melancholic features, MDD with peripartum onset, MDD with seasonal pattern, premenstrual depressive disorder, and general assessment of depression), and 5 of individuals portraying personality disorders (i.e., antisocial personality disorder (PD), histrionic PD, narcissistic PD, schizoid PD, schizotypical PD) due to their relatively low (compared to other psychiatric diagnoses) overlap with depressive disorders in the population.

Raters in the human observation part had a task observing video s and marking their observations on a checklist. Checklist for marking the observations was constructed according to the scoping *meta*-review of features of depression (Smrke et al., 2021) and the human ability to detect such features. The checklist consisted of two parts, part A for observing features via event recording method, and part B for observing features via time sampling method. In part A, frequencies of occurrence of 8 features (i.e., person on the video smiled) had to be observed, each occurrence had to be marked in the checklist. In part B, raters observed the presence or absence of some feature in predetermined time intervals (30 s), i.e., after each interval, video had to be paused, and their impression of the video regarding 17 features had to be marked as present or not present (i.e., person on the video talked slowly). Exact items that were assessed in the checklist are presented in tables 19 - 21. Raters were instructed to watch each video thrice; first for filling in part A, second for part B, and third to complete part B if any feature was left unrated. Before rating the set of video s, participants rated a practice video. The number of frequencies and the number of events observed

Table 19
Correlations of AI extracted and human observation features with depression: Linguistic cues.

ALGORITHM DATA			consistent result	HUMAN OBSERVATION DATA			
Features	DEPRESSION (1 = yes, 0 = no)			DEPRESSION (1 = yes, 0 = no)		Inter-rater reliability	Features
	r			r		ICC	Person in the video …
A1: Use of first-person singular pronouns	0.33			0.02		0.93	… used a personal pronoun (“I”, “Me”, “My”, “Mine”).[A]
A2: Use of negatively valanced words				0.41		0.96	… used a word with explicitly negative valence (“Hurt”, “Tears”, “Hate”, “Sleep”, “Worry”, “Pain”, “Alone”, “Sad”, “Sadness”, “Angry”, “Anger”, “Depressed”, “Depression”).[A]
… percentage of negatively valanced words	0.17		X				
… percentage of negatively valanced sentences	0.30		X				
A3: Explicit mentions of treatment of depression	0.45	*	X	0.35		0.96	… used word(s) about the treatment (“Therapy”, “Psychotherapy”, “Treatment”, “Pills”, “Medication”, “Side effects”).[A]
A4: Use of absolutist words – small grammar	0.47	*	X	0.54	*	0.94	… used absolutistic word(s) (“Always”, “Never”, “All the time”, “Everything”, “Nothing”).[A]
A4: Use of absolutist words – extended grammar	0.47	*	X				
A5: Focusing on the past	−0.26			0.06		0.61	… talked about the past.[B]
A6.1: Syntactic complexity				0.13		0.72	… used short, simple sentences.[B]ˉ
A6.1.1: Sentence length	−0.14		X				
A6.1.2: Sentence complexity	−0.07		X				
A6.2: Lexical complexity							
A6.2.1: Lexical diversity	−0.23		X				
A6.2.2: Lexical sophistication	0.04						
A6.2.3: Lexical density (normalized)	−0.47	*	X				

Notes. ICC=Inter-rater variability. * $p < 0.05$. X marks consistent result between AI extracted and human observed feature. Type of observation in human observation part: A – event recording, B time sampling method. ˉFeature in human observation part was observed in a way that higher score means corresponds to the lower score in AI extracted feature.

were summed for each feature and averaged to account for variability in the length of the video s. AI observation features were extracted using the pipeline presented in this paper.

Statistical analyses. Statistical analyses were performed using R 4.0.3. First, for the human observation part, inter-rater reliability was evaluated via intraclass correlation coefficients (R package *psych*) by the two-way random effects, absolute agreement and multiple raters/measurements method and interpreted as < 0.50 pointing to poor reliability, 0.50 – 0.75 to moderate reliability, 0.75 – 0.90 good reliability, and > 0.90 pointing to excellent reliability (Koo & Li, 2016). Bivariate correlation coefficients between features, both AI extracted and observed by raters, and presence of depression in a video were calculated (R package *Hmisc*). Correlations for AI extraction part and human observation part were compared for their consistency.

4.4.2. Results

Tables 19 - 21 present correlations of AI extracted features and depression, and of features observed by raters and depression, where inter-rater reliability is reported for each observed feature. Inter-rater reliability for all features was at least moderate, i.e., moderate for 6 features (25.0 %), good for 11 features (44.0 %), and excellent for 8 features (32.0 %).

As AI extracted features can be importantly more specific than human observations, various of them are used to be compared to more general observations done by raters, e.g., while raters observed, whether a person on a video used short, simple sentences, AI extracted features indicating to such characteristic of a speech, are defined as features of syntactic complexity (A6.1; sentence length and complexity) and lexical complexity (A6.2; lexical diversity, sophistication and density).

Table 19 presents correlations of AI extracted features and features observed by raters with depression for linguistic cues. By examining the consistency of correlations between the AI and human observation of features with depression, they are especially aligned (of similar size and direction) for explicit mentions of depression (A3) and use of absolutist words (A4), i.e., their observations were in both types of features more often observed in video s portraying depression. Somewhat different correlation values but in the same direction, are observed in use of negatively valanced words (A2), syntactic (A6.1) and lexical complexity (A6.2), i.e., negatively valanced words and higher complexity in video s portraying depression in both types of features. Use of first-person singular pronouns (A1) and focusing on the past (A5) were positively and negatively, respectively, related to video s portraying depression only in the AI extracted features.

As outlined in Table 19, the use of negatively valanced words was measured using specific words with explicitly negative valence by the human observers. This allows us to minimize bias in human perception of bias (Trofimova, 2014) and exploiting word and sentence sentiment. In our AI-based pipeline, we address bias in human perception by incorporating sentiment analysis at both word and sentence levels, aiming to minimize subjective interpretation. Leveraging sentiment analysis in depression classification proves beneficial, with word-level analysis aiding in the identification of specific terms associated with depression. This includes recognizing keywords or phrases commonly used by individuals experiencing depression. However, word-level analysis may lack context, leading to potential misinterpretation. On the other hand, sentence-level sentiment analysis offers a more comprehensive view of the emotional tone in a statement, providing a deeper understanding of the writer’s intent. This approach is generally considered more robust and less susceptible to errors caused by ambiguous words, thanks to its consideration of the entire sentence structure. In practical terms, the combination of both word-level and sentence-level sentiment analysis enhances the accuracy and reliability of depression classification (Obagbuwa et al., 2023).

Table 20 presents correlations of AI extracted features and features

Table 20
Correlations of AI extracted and human observation features with depression: Speech cues.

ALGORITHM DATA Features	DEPRESSION (1 = yes, 0 = no) r		consistent result	HUMAN OBSERVATION DATA DEPRESSION (1 = yes, 0 = no) r		Inter-rater reliability ICC	Features Person in the video …
B1: Speaking rate							
Speech rate	−0.17		X	0.15		0.71	… talked slowly.[Bˇ]
B2: Engagement in verbal communication							
F0 (Hz)	−0.32		X	−0.26		0.74	… seemed not interested in the conversation.[B]
First formant (F1) (Hz)	−0.17		X				
Second formant (F2) (Hz)	−0.24		X				
Third formant (F3) (Hz)	−0.05		X				
Fourth formant (F4) (Hz)	0.36						
Amplitude difference H1-A3 (F0-F3) (Hz)	0.16						
Range of F2 formant (Hz)	0.19						
Pitch (Hz)	−0.30		X				
Local jitter (%)							
Intensity (Db)	0.51	*					
B3: Voiced speech and pauses							
Number of voiced syllables	−0.16						
Number of pauses (silence)	−0.12						
Speaking balance	−0.04						
B4: Low articulation rate							
Articulation rate	−0.51	*		−0.04		0.86	… talked clearly and understandably.[B]
B5: Decreased voice quality							
Local absolute jitter (sec)	0.25			0.56	*	0.89	… had voice problems (e.g. tremors, strained, breathy voice).[B]
Local shimmer (%)	−0.48	*	X				
Local dB shimmer (dB)	−0.40		X				
Harmonics-to-noise ratio (HNR)	0.32						
B6: Monotonous speech				−0.06		0.8	… spoke monotonously.[Bˇ]
Speaking duration (without pauses) (sec)	−0.02						
Speaking duration (%)	−0.03						
Duration of voiced syllables (sec)	0.46	*					
Pausing (silence) duration (%)	0.03			0.24		0.9	… took long pauses during speech.[B]
Pausing (silence) duration (sec)	0.05						

Notes. ICC=Inter-rater variability. * $p < 0.05$. X marks consistent result between AI extracted and human observed feature. Type of observation in human observation part: A – event recording, B – time sampling method.ˇFeature in human observation part was observed in a way that higher score means corresponds to the lower score in AI extracted feature.

observed by raters with depression for speech cues. Examining the consistency of correlations between the AI and human observation of features with depression, they are aligned in the following features: Speech rate (B1) as an AI extracted feature was less present in video s portraying depression, which aligns with 'talking slowly' as more often observed by raters in such video s. Several AI features also align with a person in the video that 'seemed not interested in the conversation' more often in video s portraying as observed by raters, i.e., some features of engagement in verbal communication (B2; F0, first, second, and third formant, pitch). Similarly, while raters observed more often in the video s portraying depression, that a person 'had voice problems', this was similarly observed in AI extracted features pointing to decreased voice quality (B5; local (dB) shimmer. Other features were less aligned, in most cases more often observed in AI features in relation to the type of video than observed by raters.

Table 21 presents correlations of AI extracted features and features observed by raters with depression for visual cues. Examining the consistency of correlations between the AI and human observation of features with depression, they are aligned in the following features: Raters less often observed a display of positive emotional facial expression in video s portraying depression, which was also extracted via AI features, which holds true also for observation by raters that a person 'often changed his/her facial expression' and the number of emotional variability (both features C3: Occurrence and emotional variability), and neutral emotional expressions. As outlined in Table 21, we did not measure negative emotions as an overall concept, but extract emotions from facial expressions based on Ekman's 8 basic emotions. Namely, interpreting negative emotions requires a nuanced understanding of subtle cues and contexts, making it more difficult to automate than positive or neutral emotions (Machová et al., 2023). Focusing on negative emotions on a more granular level offers several advantages over considering them solely through valence (Tan et al., 2022). Furthermore, in several AI features, Frame (%) and Intensity were extracted, of which only Intensity variant aligned with the results of human observations for eyebrow movements (C4), frowns (C6), but both variants aligned with the results of human observations for down-angled mouth corners (C11) and smiling (C12). Aligned are also the correlations of: features with the portray of depression for gaze aversion and downward gaze (C5) as an AI extracted feature and observation that a person 'was looking down'; some features related to head movements and turns (C7), more specifically, head facing away in up/down and z direction frames with observation that a person 'had his/her head moved away from the person conducting the interview'. Head moving frames (C7) and low, moderate, and high rotational energy frames (C8) were aligned to an observation that a person in the video 'moved his/her head' less often in the video s portraying depression.

5. Study limitations

While the proposed framework provides a unified and consistent approach to automatically describe symptoms of depression using observable cues, it is not without limitations. First, as already emphasized, the main limitation of this study is the nature and context of the material within which the pipeline is evaluated. Firstly, there is a

Table 21
Correlations of AI extracted and human observation features with depression: Visual cues.

ALGORITHM DATA Features	DEPRESSION (1 = yes, 0 = no) r		consistent result	HUMAN OBSERVATION DATA DEPRESSION (1 = yes, 0 = no) r		Inter-rater reliability ICC	Features Person in the video …
C1: Occurrence of facial expressions (Frame (%)) – Other emotions							
Emotion – Surprise	−0.46	*					
Emotion – Anger	0.07						
Emotion – Fear	−0.35						
Emotion – Happiness	−0.33						
Emotion – Sadness	−0.43						
Emotion – Disgust	−0.10						
C2: Intensity of facial expressions (Intensity 0—5 point))							
Emotion – Surprise	−0.34						
Emotion – Anger	−0.10						
Emotion – Fear	−0.31						
Emotion – Happiness	−0.10						
Emotion – Sadness	−0.30						
Emotion – Disgust	−0.46	*					
C3: Occurrence and emotional variability							
Positive emotions (Frame (%))	−0.33		X	−0.53	*	0.86	… displayed positive emotional facial expression.^{B}
Positive emotions (Intensity (0—5 point))	−0.10		X				
Negative emotions (Frame (%))	–			0.65	*	0.91	… displayed negative emotional facial expression.^{B}
Negative emotions (Intensity (0—5 point))	–						
Total number of emotional variabilities	−0.24		X	−0.42		0.88	… often changed his/her facial expressions.^{B}
Neutral expressions (Frame (%)) [*Additional feature]	−0.70	*	X	−0.57	*	0.87	… displayed neutral emotional facial expression.^{B}
C4: Eyebrow movements							
Frame (%)	0.02			−0.39		0.87	… moved his/her eyebrows.^{A}
Intensity (0—5 point)	−0.11		X				
C5: Gaze aversion and downward gaze							
Looking down frames	0.34		X	0.23		0.97	… was looking down.^{B}
C6: Frowns							
Frame (%)	0.10			−0.46	*	0.66	… frowned.^{A}
Intensity (0—5 point)	−0.25		X				
C7: Head movement and turns							
Head facing away (up/down) frames	−0.33		X	−0.26		0.83	… had his/her head moved away from the person conducting the interview.^{B}
Head facing away (left/right) frames	0.06						
Head facing away (z direction) frames	−0.34		X				
Head down frames	−0.59	*		0.35		0.81	… had his/her head tilted down.^{B}
Face moving frames	−0.12						
Head moving frames	−0.35		X	−0.31		0.82	… moved his/her head.^{B}
C8: Rotational energy of head movements							
Frames below rotational threshold (<50 J)							
Low rotational energy frames (>= 50 J, < 500 J)	−0.32		X				
Moderate rotational energy frames (>= 500 J, < 5000 J)	−0.27		X				
High rotational energy frames (>= 5000 J)	−0.17		X				
C9: Mouth movement							
Frame (%)	−0.38						
Intensity (0—5 point)	−0.29						
C10: Lip press							
Frame (%)	−0.04			0.36		0.6	… pressed his/her lips.^{A}
Intensity (0—5 point)	0.01						
C11: Down-angled mouth corners							
Frame (%)	−0.24		X	0.59	*	0.86	… had the mouth corners turned down.^{B}
Intensity (0—5 point)	−0.38		X				
C12: Smiling							
Frame (%)	−0.22		X	−0.33		0.94	… smiled.^{A}
Intensity (0—5 point)	−0.10		X				

Notes. ICC=Inter-rater variability. * $p < 0.05$. X marks consistant result between AI extracted and human observed feature. Type of observation in human observation part: A – event recording, B – time sampling method.

significant lack of available multimodal dataset, correlating depression with how it is expressed through language, facial expressions, and speech. Secondly, the discourse in the SM dataset is, on the one hand, generated by actors acting out a monologue describing their experiences (i.e., unstructured interview), but presumably without experiencing psychological stress. The discourse in the DW dataset, on the other hand, is arguably more natural and generated by 'ordinary' subjects rather than professionals. However, it is scripted and controlled by a chatbot (i.e., a structured interview) and focused on citizens. This means that some features may be over-emphasised in the SM dataset and thus more in line with the theoretical background, whereas some features may be significantly under-emphasised in the DW dataset due to the structured nature of the discourse. This may well explain the 'smaller' differences in observable behavioural cues between recordings with and without depressive symptoms in the DW dataset. To increase the generalizability, further limited by evaluation of the pipeline with only a small-subset of recordings, we also compare the AI-extracted features and how they are perceived by human observer. As expected, the features are not observed the same, however, the results clearly show that AI-pipeline extraction is aligned with human perception. Second, neither dataset allows for a broader psychological analysis of the subjects. In fact, in the SM dataset, non-depressed subjects present with symptoms of other psychological disorders, and in the DW dataset, assessments of disorders that may co-occur with depression (e.g., anxiety, conduct disorder, dysthymic disorder, chronic fatigue) are not known. Furthermore, the cultural bias stemming from datasets consisting solely of English-language discourse with participants from developed countries underscores the need for more socio-culturally diverse subjects to ensure the broader applicability of your findings. Consequently, contextual factors and the lack of comprehensive data on individuals' mental health may bias our results. This may well explain some of the 'smaller' differences in observable behavioural cues between recordings of people with depressive symptoms and those without, as well as, subtle differences in the perception of human observes. Third, some linguistic, but even more so acoustic and visual cues, tend to be individual-specific and depend on the individual's speaking style (Liu et al., 2017). As our results show, this leads to relatively high variability for some features. Thus, without understanding the baselines of depressed individuals, it is difficult to generalise the results for behavioural cues. Fourth, although the accuracy of algorithms measuring the presence of an observable cue was evaluated in Arioz et al. 2022, the availability of only 56 relevant recordings in both datasets limits the generalizability of the results. Finally, as both datasets consist only of English-language discourse with Anglophone participants from developed countries, our conclusions may be culturally biased. Further research with more socio-culturally diverse subjects is therefore needed to understand if and how our findings can be generalised. Finally, the reliance on hand-crafted features, although referenced from a previous *meta*-review, raises questions about the adaptability and relevance of these features across diverse populations. The lack of existing datasets and lack of socio-cultural diversity in the datasets further undermines the representatives of this and similar studies and clinical feasibility of such tools to be used in psychiatric practice.

6. Conclusion

In this paper, we have presented and evaluated an E2E framework for extracting observable cues of depression from diary recordings. To the best of our knowledge, this is the first digital framework that consistently and automatically detects depressive disorders using observable cues. The framework can be used to add a layer of explicability to general data-driven AI models for depression classification, as well as a baseline for designing explicable knowledge-based models for depression classification. We identify cues A1-A3, B1-B4, B6 and C8 as the most salient and consistent features for describing depressive disorders using digital observable cues and explaining the decisions behind AI. The digital observable cues A4-6, B3, B5, and most of the C features require further research (i.e., comparative analysis using expert annotation of the cues) to analyse the influence of the nature/context of the discourse, as well as the influence of other disorders and physical/psychological stress. In addition, this paper is the first, to our knowledge, to provide calculations of effect sizes of differences between features extracted from recordings of individuals with and without depressive disorder, with implications for further (clinical) research on depression cue recognition. Overall, linguistic and speech features tend to be less person-specific than visual cues. Furthermore, depression was found to be a disorder that is expressed using highly interrelated cues from all three modalities (i.e., speech, language, and facial expressions). Thus, future research, using longitudinal and experimental studies, should analyse and model the interference and causal relationships between observable behavioural cues.

The most apparent barrier for AI in mental health is the lack of multimodal datasets. Namely, multimodal data incorporation is crucial for improving the accuracy and effectiveness of AI in mental health diagnostics and treatments, yet the availability of such extensive and diverse datasets remains insufficient. As part of our future efforts, we are preparing an observational controlled trial, through which we intend to collect audio −visual responses from a large cohort of subjects from multiple countries, during clinical sessions and controls. The activity is planned to be carried out in the second quarter of 2024 (a study protocol is being formulated). Furthermore, considering growing body of research in the field of artificial intelligence applications in psychiatry, we will consider how to harmonize different datasets in mental health research using tools and frameworks such as Harmony (Moltrecht and McElroy, 2022). Given, one of the main barriers for existence of such multimodal dataset being the sensitivity of video recordings containing personal information, the pipeline proposed in the paper will be used to make, a privacy preserving dataset with features and observable cues, publicly available for research. Finally, to improve applicability in clinical settings, a generalisation method should be investigated to generate standardised scores that would indicate the risk of depression without considering or understanding the healthy baseline of individual subjects. Moreover, depression severity is a key dimension that informs clinicians about the intensity of symptoms and guides the selection of appropriate interventions, ranging from psychotherapy to pharmacotherapy. Given the abundance and diversity of information that will be collected in the clinical study, we will continue with the analysis of the more nuanced understanding of the observable cues in depression. On one hand this will involve the analysis of the discriminatory power between psychiatric disorders. On the other it will involve analysis of the observable cues related to severity depression and aligning our work with DSM-6 criteria (O'Connor et al., 2009).

Funding

This work has been funded by the European Union Horizon Europe Research and Innovation Programme projects SMILE (grant number 101080923) and the Slovenian Research and Innovation Agency, Advanced methods of interaction in telecommunication research programme (grant number P2-0069). The content of this paper does not reflect the official position of the European Union or any other institution. The information and views expressed are the sole responsibility of the authors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eswa.2024.125025.

References

Alghowinem, S., Gedeon, T., Goecke, R., Cohn, J. F., & Parker, G. (2020). Interpretation of depression detection models via feature selection methods. *IEEE transactions on affective computing, 14*(1), 133–152. https://doi.org/10.1109/TAFFC.2020.3035535

Al-Mosaiwi, M. &Johnstone, T. (2018). In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation. Clinical Psychological Science; 6. Epub ahead of print 2018. DOI: 10.1177/2167702617747074.

Ambaw, F., Mayston, R., Hanlon, C., Medhin, G. & Alem, A. (2018). Untreated depression and tuberculosis treatment outcomes, quality of life and disability, Ethiopia. Bull World Health Organ; 96. Epub ahead of print 2018. DOI: 10.2471/BLT.17.192658.

Arioz, U., Smrke, U., Plohl, N., & Mlakar, I. (2022). Scoping Review on the Multimodal Classification of Depression and Experimental Study on Existing Multimodal Models. *Diagnostics, 2022*(12), 2683.

Ayaz, M., Pasha, M. F., Alzahrani, M. Y., Budiarto, R., & Stiawan, D. (2021). The Fast Health Interoperability Resources (FHIR) standard: Systematic literature review of implementations, applications, challenges and opportunities. *JMIR medical informatics, 9*(7), e21929.

Bailey, A., & Plumbley, M. D. (2021). In *August). Gender bias in depression detection using audio features* (pp. 596–600). IEEE.

Balcombe, L. & de Leo, D. (2021). Digital mental health challenges and the horizon ahead for solutions. JMIR Mental Health; 8. Epub ahead of print 2021. DOI: 10.2196/26811.

Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016). OpenFace: An open source facial behavior analysis toolkit. In: 2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016. 2016. Epub ahead of print 2016. DOI: 10.1109/WACV.2016.7477553.

Baltrušaitis, T., Robinson, P., & Morency, L. P. (2013). Constrained local neural fields for robust facial landmark detection in the wild. In: Proceedings of the IEEE International Conference on Computer Vision. 2013. Epub ahead of print 2013. DOI: 10.1109/ICCVW.2013.54.

Bayoudh, K., Knani, R., Hamdaoui, F., & Mtibaa, A. (2022). A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets. *The Visual Computer, 38*(8), 2939–2970.

Bickman, L. (2020). Improving Mental Health Services: A 50-Year Journey from Randomized Experiments to Artificial Intelligence and Precision Mental Health. Administration and Policy in Mental Health and Mental Health Services Research; 47. Epub ahead of print 2020. DOI: 10.1007/s10488-020-01065-8.

Biswas, A., Sandhya, P & Saravanan, T.R. (2021). Depression Detection from Facial Behaviour through Deep Learning. Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, Pages. 5341 – 5349.

Boersma, P., & Van Heuven, V. (2001). Speak and unSpeak with PRAAT. *Glot International, 5*(9/10), 341–347.

Bradski, G. (2000). The OpenCV Library. *Dr Dobb's Journal of Software Tools, 120*, 122–125.

Brenes, G.A. (2007). Anxiety, depression, and quality of life in primary care patients. Prim Care Companion J Clin Psychiatry; 9. Epub ahead of print 200. DOI : 10.4088/PCC.v09n0606.

Burdisso, S. G., Errecalde, M., & Montes-y-Gómez, M. (2019). A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications, 133*, 182–197. https://doi.org/10.1016/j.eswa.2019.05.023

Bylsma, L. M., Morris, B. H., & Rottenberg, J. (2008). A meta-analysis of emotional reactivity in major depressive disorder. *Clinical psychology review, 28*(4), 676–691. https://doi.org/10.1016/j.cpr.2007.10.001

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 3*(3), 223–230. https://doi.org/10.1016/j.bpsc.2017.11.007

Cai, H., Yuan, Z., Gao, Y., Sun, S., Li, N., Tian, F., & Hu, B. (2022). A multi-modal open dataset for mental-disorder analysis. *Scientific Data, 9*(1), 178. https://doi.org/10.1038/s41597-022-01211-x

Ceccarelli, F., & Mahmoud, M. (2022). Multimodal temporal machine learning for Bipolar Disorder and Depression Recognition. *Pattern Analysis and Applications, 25*(3), 493–504. https://doi.org/10.1007/s10044-021-01001-y

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2013). One billion word benchmark for measuring progress in statistical language modeling. arXiv preprint arXiv:1312.3005. Chen, I. Y., Szolovits, P., & Ghassemi, M. (2019). Can AI help reduce disparities in general medical and mental health care?. AMA journal of ethics, 21(2), 167-179. DOI: 10.1001/amajethics.2019.167.

Cole, B. P., & Davidson, M. M. (2019). Exploring men's perceptions about male depression. *Psychology of Men & Masculinities, 20*(4), 459. https://doi.org/10.1037/men0000176

Çiftçi, E., Kaya, H., Güleç, H., & Salah, A. A. (2018, May). The turkish audio-visual bipolar disorder corpus. In 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia) (pp. 1-6). IEEE. Craven, M. A., & Bland, R. (2013). Depression in primary care: current and future challenges. The Canadian Journal of Psychiatry, 58(8), 442-448. DOI: 10.1177/070674371305800802.

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

D'Alfonso, S. (2020). AI in mental health. *Current Opinion in Psychology, 36*, 112–117. https://doi.org/10.1016/j.copsyc.2020.04.005

Dauphin, V. B. (2020). A critique of the American Psychological Association Clinical Practice Guideline for the Treatment of Posttraumatic Stress Disorder (PTSD) in Adults. *Psychoanalytic Psychology, 37*(2), 117.

Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S. (2014). "COVAREP – A collaborative voice analysis repository for speech technologies", In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy.

DeJean, D., Giacomini, M., Vanstone, M., & Brundisini, F. (2013). Patient experiences of depression and anxiety with chronic disease: a systematic review and qualitative meta-synthesis. Ontario health technology assessment series. 2013 Sep 1;13(16):1-33. PMID: 24228079; PMCID: PMC3817854.

Devaraj, S., Sharma, S. K., Fausto, D. J., Viernes, S., & Kharrazi, H. (2014). Barriers and facilitators to clinical decision support systems adoption: A systematic review. *Journal of Business Administration Research, 3*(2), 36. https://doi.org/10.5430/jbar.v3n2p36

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., ... & Morency, L. P. (2014, May). SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems (pp. 1061-1068). Dibeklioglu, H., Hammal, Z. & Cohn, J. F. (2017) Dynamic multimodal measurement of depression severity using deep autoencoding. 2017 IEEE Journal of Biomedical and Health Informatics, 22(2): 525–536.

Diep, B., Stanojevic, M., & Novikova, J. (2022). Multi-modal deep learning system for depression and anxiety detection. arXiv preprint arXiv:2212.14490. Draucker, C. B. (2005). Interaction patterns of adolescents with depression and the important adults in their lives. Qualitative Health Research, 15(7), 942-963. DOI: 10.1177/1049732305277859.

Dwyer, A., de Almeida Neto, A., Estival, D., Li, W., Lam-Cassettari, C., & Antoniou, M. (2021). Suitability of text-based communications for the delivery of psychological therapeutic services to rural and remote communities: Scoping review. *JMIR mental health, 8*(2), e19478.

Egede, L. E. (2007). Failure to recognize depression in primary care: Issues and challenges. *Journal of General Internal Medicine, 22*, 701–703. https://doi.org/10.1007/s11606-007-0170-z

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *The Facial Action Coding system: A Technique for the Measurement of Facial Movement*. San Francisco, A: Consulting Psychologists Press.

Ellis, P.D. (2010). The Essential Guide to Effect Sizes: An Introduction to Statistical Power, Meta-Analysis and the Interpretation of Research Results. Cambridge: Cambridge University Press. Doi:10.1017/CBO9780511761676.

El-Sherif, D. M., Abouzid, M., Elzarif, M. T., Ahmed, A. A., Albakri, A., & Alshehri, M. M. (2022, February). Telehealth and Artificial Intelligence insights into healthcare during the COVID-19 pandemic. In Healthcare (Vol. 10, No. 2, p. 385). MDPI. DOI: 10.3390/healthcare10020385.

Fan, X., & Tjahjadi, T. (2019). Fusing dynamic deep learned features and handcrafted features for facial expression recognition. *Journal of Visual Communication and Image Representation, 65*, Article 102659.

Fazi, M. B. (2021). Beyond human: Deep learning, explainability and representation. *Theory, Culture & Society, 38*(7–8), 55–77. https://doi.org/10.1177/0263276420966386

Flanagan, O., Chan, A., Roop, P., & Sundram, F. (2021). Using acoustic speech patterns from smartphones to investigate mood disorders: Scoping review. *JMIR mHealth and uHealth, 9*(9), e24352.

Giannakakis, G., Koujan, M. R., Roussos, A., & Marias, K. (2022). Automatic stress analysis from facial videos based on deep facial action units recognition. *Pattern Analysis and Applications, 1–15*. https://doi.org/10.1007/s10044-021-01012-9

Giddens, C. L., Barron, K. W., Byrd-Craven, J., Clark, K. F., & Winter, A. S. (2013). Vocal indices of stress: a review. Journal of voice, 27(3), 390-e21. DOI: 10. /j.jvoice.2012.12.010.

Ginsburg, B., Castonguay, P., Hrinchuk, O., Kuchaiev, O., Lavrukhin, V., Leary, R., ... & Cohen, J. M. (2019). Stochastic gradient methods with layer-wise adaptive moments for training of deep networks. arXiv preprint arXiv:1905.11286.

Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H. C., & Jeste, D. V. (2019). Artificial intelligence for mental health and mental illnesses: An overview. *Current psychiatry reports, 21*, 1–18.

Gratch, J., Artstein, R., Lucas, G. M., Stratou, G., Scherer, S., Nazarian, A., ... & Morency, L. P. (2014, May). The distress analysis interview corpus of human and computer interviews. In LREC (pp. 3123-3128).

Guohou, S., Lina, Z., & Dongsong, Z. (2020). What reveals about depression level? The role of multimodal features at the level of interview questions. *Information & Management, 57*(7), Article 103349.

Habbal, A., Ali, M. K., & Abuzaraida, M. A. (2024). Artificial Intelligence Trust, risk and security management (AI trism): Frameworks, applications, challenges and future research directions. *Expert Syst Appl, 240*, Article 122442.

Harlé, K. M., Allen, J. J., & Sanfey, A. G. (2010). The impact of depression on social economic decision making. *Journal of abnormal psychology, 119*(2), 440. https://doi.org/10.1037/a0020075

Jacobson, N. C., Summers, B., & Wilhelm, S. (2020). Digital biomarkers of social anxiety severity: Digital phenotyping using passive smartphone sensors. *Journal of medical Internet research, 22*(5), e16875.

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics, 71*, 1–15.

Jan, A., Meng, H., Gaus, Y. F. B. A., & Zhang, F. (2017). Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. *IEEE Transactions on Cognitive and Developmental Systems, 10*(3), 668–680. https://doi.org/10.1109/TCDS.2017.2721552

Kappen, M., Hoorelbeke, K., Madhu, N., Demuynck, K., & Vanderhasselt, M. A. (2022). Speech as an indicator for psychosocial stress: A network analytic approach. *Behavior Research Methods, 1–12*. https://doi.org/10.3758/s13428-021-01670-x

Keane, P. A., & Topol, E. J. (2018). With an eye to AI and autonomous diagnosis. *NPJ Digital Medicine, 1*(1), 40. https://doi.org/10.1038/s41746-018-0048-y

Kim, J., Uddin, Z. A., Lee, Y., Nasri, F., Gill, H., Subramanieapillai, M., & McIntyre, R. S. (2021). A systematic review of the validity of screening depression through Facebook, Twitter, Instagram, and Snapchat. *Journal of Affective Disorders, 286*, 360–369. https://doi.org/10.1016/j.jad.2020.08.091

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of chiropractic medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of general internal medicine, 16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of affective disorders, 114*(1–3), 163–173. https://doi.org/10.1016/j.jad.2008.06.026

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in psychology, 4*, 863.

Lam, G., Dongyan, H., & Lin, W. (2019). Context-aware deep learning for multi-modal depression detection. In *ICASSP 2019–2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 3946–3950). IEEE. https://doi.org/10.1109/ICASSP.2019.8683027.

Lang, H., & Cui, C. (2018). Automated depression analysis using convolutional neural networks from speech, Journal of Biomedical Informatics, Volume 83. *ISSN, 103–111*, 1532–10464. https://doi.org/10.1016/j.jbi.2018.05.007

Lee, E. E., Torous, J., De Choudhury, M., Depp, C. A., Graham, S. A., Kim, H. C., & Jeste, D. V. (2021). Artificial intelligence for mental health care: Clinical applications, barriers, facilitators, and artificial wisdom. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 6*(9), 856–864. https://doi.org/10.1016/j.bpsc.2021.02.001

Li, Y., Niu, M., Zhao, Z., & Tao, J. (2022). Automatic depression level assessment from speech by long-term global information embedding. In *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8507–8511). IEEE. https://doi.org/10.1109/ICASSP43922.2022.9747292.

Li, H., Ge, S., Greene, B., & Dunbar-Jacob, J. (2019). Depression in the context of chronic diseases in the United States and China. *International journal of nursing sciences, 6*(1), 117–122. https://doi.org/10.1016/j.ijnss.2018.11.007

Li, J., Fu, X., Shao, Z., & Shang, Y. (2018). Improvement on speech depression recognition based on deep networks. In *In 2018 Chinese Automation Congress (CAC)* (pp. 2705–2709). IEEE. https://doi.org/10.1109/CAC.2018.8623055.

Li, W., Tan, R., Xing, Y., Li, G., Li, S., Zeng, G., & Cao, D. (2022). A multimodal psychological, physiological and behavioural dataset for human emotions in driving tasks. *Scientific Data, 9*(1), 481. https://doi.org/10.1038/s41597-022-01557-2

Lin, L., Chen, X., Shen, Y., & Zhang, L. (2020). Towards automatic depression detection: A BiLSTM/1D CNN-based model. *Applied Sciences, 10*(23), 8701. https://doi.org/10.3390/app10238701

Linder, A., Gerdtham, U. G., Trygg, N., Fritzell, S., & Saha, S. (2020). Inequalities in the economic consequences of depression and anxiety in Europe: A systematic scoping review. *European journal of public health, 30*(4), 767–777. https://doi.org/10.1093/eurpub/ckz127

Liu, D., Feng, X. L., Ahmed, F., Shahid, M., & Guo, J. (2022). Detecting and measuring depression on social media using a machine learning approach: Systematic review. *JMIR Mental Health, 9*(3), e27244.

Liu, Q., He, H., Yang, J., Feng, X., Zhao, F., & Lyu, J. (2020). Changes in the global burden of depression from 1990 to 2017: Findings from the Global Burden of Disease study. *Journal of psychiatric research, 126*, 134–140. https://doi.org/10.1016/j.jpsychires.2019.08.002

Liu, Z., Hu, B., Li, X., Liu, F., Wang, G., & Yang, J. (2017). Detecting depression in speech under different speaking styles and emotional valences. In Brain Informatics: International Conference, BI 2017, Beijing, China, November 16-18, 2017, Proceedings (pp. 261-271). Springer International Publishing. DOI: 10.1007/978-3-319-70772-3_25.90.

Loper, E., & Bird, S. (2002). The natural language toolkit NLTK: The Natural Language Toolkit. Proceedings of the ACL-02 Workshop on Effective tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, July 2002, Philadelphia, Pennsylvania, USA, pp. 63-70. DOI: 10.3115/1118108.1118117.

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope investigative otolaryngology, 5*(1), 96–116. https://doi.org/10.1002/lio2.354

Lu, C., Bu, Y., Wang, J., Ding, Y., Torvik, V., Schnaars, M., & Zhang, C. (2019). Examining scientific writing styles from the perspective of linguistic complexity. *Journal of the Association for Information Science and Technology, 70*, 462–475. https://doi.org/10.1002/asi.24126

Lu, C., Bu, Y., Dong, X., Wang, J., Ding, Y., Larivière, V., & Zhang, C. (2019). Analyzing linguistic complexity and scientific impact. *Journal of Informetrics, 13*(3), 817–829.

Machová, K., Szabóova, M., & Paralič, J. (2023). Detection of emotion by text analysis using machine learning. *Frontiers in Psychology, 14*, 1190326.

Maddage, N. C., Senaratne, R., Low, L. S. A., Lech, M., & Allen, N. (2009). Video-based detection of the clinical depression in adolescents. In *In 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 3723–3726). IEEE. https://doi.org/10.1109/IEMBS.2009.5334815.

Maupomé, D., Armstrong, M. D., Belbahar, R. M., Alezot, J., Balassiano, R., Queudot, M., ... & Meurs, M. J. (2020, September). Early Mental Health Risk Assessment through Writing Styles, Topics and Neural Models. In CLEF (Working Notes).

Menear, M., Dore, I., Cloutier, A. M., Perrier, L., Roberge, P., Duhoux, A., & Fournier, L. (2015). The influence of comorbid chronic physical conditions on depression recognition in primary care: A systematic review. *Journal of psychosomatic research, 78*(4), 304–313. https://doi.org/10.1016/j.jpsychores.2014.11.016

Mezzi, R., Yahyaoui, A., Krir, M. W., Boulila, W., & Koubaa, A. (2022). Mental health intent recognition for Arabic-speaking patients using the mini international neuropsychiatric interview (MINI) and BERT model. *Sensors, 22*(3), 846. https://doi.org/10.3390/s22030846

Mlakar, I., Šafran, V., Hari, D., Rojc, M., Alankuş, G., Pérez Luna, R., & Ariöz, U. (2021). Multilingual conversational systems to drive the collection of patient-reported outcomes and integration into clinical workflows. *Symmetry, 13*(7), 1187. https://doi.org/10.3390/sym13071187

Moltrecht, B., & McElroy, E. (2022). *Harmony Project: A new AI data tool that facilitates global mental health research. Centre for Longitudinal Studies (CLS)*. London: Social Research Institute IOE, UCL's Faculty of Education and Society.

Nahum, M., Van Vleet, T. M., Sohal, V. S., Mirzabekov, J. J., Rao, V. R., Wallace, D. L., & Chang, E. F. (2017). Immediate mood scaler: Tracking symptoms of depression and anxiety using a novel mobile mood scale. *JMIR mHealth and uHealth, 5*(4), e6544.

Nickels, S., Edwards, M. D., Poole, S. F., Winter, D., Gronsbell, J., Rozenkrants, B., & Fromer, M. (2021). Toward a mobile platform for real-world digital measurement of depression: User-centered design, data quality, and behavioral and clinical modeling. *JMIR mental health, 8*(8), e27589.

Obagbuwa, I. C., Danster, S., & Chibaya, O. C. (2023). Supervised machine learning models for depression sentiment analysis. *Frontiers in Artificial Intelligence, 6*.

Orton Indigo, J.D. (2020) Vision based body gesture meta features for affective computing. arXiv preprint arXiv:2003.00809.

Pampouchidou, A., Simos, P. G., Marias, K., Meriaudeau, F., Yang, F., Pediaditis, M., & Tsiknakis, M. (2017). Automatic assessment of depression based on visual cues: A systematic review. *IEEE Transactions on Affective Computing, 10*(4), 445–470. https://doi.org/10.1109/TAFFC.2017.2724035

Pan, Y., Shang, Y., Liu, T., Shao, Z., Guo, G., Ding, H., & Hu, Q. (2024). Spatial–temporal attention network for depression recognition from facial videos. *Expert Syst Appl, 237*, Article 121410.

Pietkiewicz, I. J., Bańbura-Nowak, A., Tomalski, R., & Boon, S. (2021). Revisiting false-positive and imitated dissociative identity disorder. *Frontiers in Psychology, 12*, Article 637929. https://doi.org/10.3389/fpsyg.2021.637929

Plana-Ripoll, O., Pedersen, C. B., Holtz, Y., Benros, M. E., Dalsgaard, S., De Jonge, P., & McGrath, J. J. (2019). Exploring comorbidity within mental disorders among a Danish national population. *JAMA psychiatry, 76*(3), 259–270. https://doi.org/10.1001/jamapsychiatry.2018.3658

Ploughman, M., Wallack, E. M., Chatterjee, T., Kirkland, M. C., Curtis, M. E., Lifestyle, T. H., & Aging with MS Consortium. (2020). Under-treated depression negatively impacts lifestyle behaviors, participation and health-related quality of life among older people with multiple sclerosis. *Multiple Sclerosis and Related Disorders, 40*, Article 101919. https://doi.org/10.1016/j.msard.2019.101919

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. arXiv preprint arXiv: 2003.07082. DOI: 10.18653/v1/2020.acl-demos.14.

Qureshi, S. A., Saha, S., Hasanuzzaman, M., & Dias, G. (2019). Multitask representation learning for multimodal estimation of depression level. *IEEE Intelligent Systems, 34*(5), 45–52. https://doi.org/10.1109/MIS.2019.2925204

Van Puyvelde, M., Neyt, X., McGlone, F., & Pattyn, N. (2018). Voice stress analysis: A new framework for voice and effort in human performance. *Frontiers in psychology, 9*, 1994. https://doi.org/10.3389/fpsyg.2018.01994

Ramon, Y., Farrokhnia, R. A., Matz, S. C., & Martens, D. (2021). Explainable AI for psychological profiling from behavioral data: An application to big five personality predictions from financial transaction records. *Information, 12*(12), 518. https://doi.org/10.3390/info12120518

Ray, A., Kumar, S., Reddy, R., Mukherjee, P., & Garg, R. (2019, October). Multi-level attention network using text, audio and video for depression prediction. In Proceedings of the 9th international on audio/visual emotion challenge and workshop (pp. 81-88). DOI: 10.1145/3347320.3357697.

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

Rohanian, M., Hough, J., & Purver, M. (2019, September). Detecting Depression with Word-Level Multimodal Fusion. In Interspeech (pp. 1443-1447). DOI: 10.21437/Interspeech.2019-2283.

Šafran, V., Lin, S., Nateqi, J., Martin, A. G., Smrke, U., Ariöz, U., & Mlakar, I. (2024). Multilingual Framework for Risk Assessment and Symptom Tracking (MRAST). *Sensors, 24*(4), 1101.

Sandmeir, A., Schoenherr, D., Altmann, U., Nikendei, C., Schauenburg, H., & Dinger, U. (2021). Depression severity is related to less gross body movement: A motion energy analysis. *Psychopathology, 54*(2), 106–112. https://doi.org/10.1159/000512959

Sardari, S., Nakisa, B., Rastgoo, M. N., & Eklund, P. (2022). Audio based depression detection using Convolutional Autoencoder. *Expert Syst Appl, 189*, Article 116076.

Sato, H., & Kawahara, J. (2011). Selective bias in retrospective self-reports of negative mood states. *Anxiety Stress Coping, 24*, 359–367. https://doi.org/10.1080/10615806.2010.543132

Sequeira, L., Perrotta, S., LaGrassa, J., Merikangas, K., Kreindler, D., Kundur, D., & Strauss, J. (2020). Mobile and wearable technology for monitoring depressive symptoms in children and adolescents: A scoping review. *Journal of affective disorders, 265*, 314–324. https://doi.org/10.1016/j.jad.2019.11.156

Shahin, I., & Botros, N. (2001, April). Modeling and analyzing the vocal tract under normal and stressful talking conditions. In Proceedings. IEEE SoutheastCon 2001 (Cat. No. 01CH37208) (pp. 213-220). IEEE. DOI: 10.1109/SECON.2001.923118.

Shen, Y., Yang, H., & Lin, L. (2022, May). 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) (pp. 6247-6251). IEEE. DOI: 10.1109/ICASSP43922.2022.9746569.

Shirazian, S., Grant, C. D., Aina, O., Mattana, J., Khorassani, F., & Ricardo, A. C. (2017). Depression in chronic kidney disease and end-stage renal disease: Similarities and differences in diagnosis, epidemiology, and management. *Kidney international reports, 2*(1), 94–107. https://doi.org/10.1016/j.ekir.2016.09.005

Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. *Jama, 320*(21), 2199–2200. https://doi.org/10.1001/jama.2018.17163

Sigmund, M. (2012). Influence of psychological stress on formant structure of vowels. Elektronika ir Elektrotechnika; 18. Epub ahead of print 2012. DOI: 10.5755/j01.eee.18.10.3059.

Silva, W. J., Lopes, L., Galdino, M. K. C., & Almeida, A. A. (2021). Voice acoustic parameters as predictors of depression. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2021.06.018

Singh, P., Srinivas, K. K., Peddi, A., Shabarinath, B., Neelima, I., & Bhagavathi, K. A. (2022). In *March). Artificial Intelligence based Early Detection and Timely Diagnosis of Mental Illness-A Review* (pp. 282–286). IEEE.

Smrke, U., Mlakar, I., Lin, S., Musil, B., & Plohl, N. (2021). Language, speech, and facial expression features for artificial intelligence–based detection of cancer survivors' depression: Scoping meta-review. *JMIR Mental Health, 8*(12), e30439.

Solid, C.A. (2022). Practical Strategies to Assess Value in Health Care. 1st ed. Cham: Springer International Publishing, 2022. Epub ahead of print 9 March 2022. DOI: 10.1007/978-3-030-95149-8.

Su, C., Xu, Z., Pathak, J., & Wang, F. (2020). Deep learning in mental health outcome research: A scoping review. *Translational Psychiatry, 10*(1), 116. https://doi.org/10.1038/s41398-020-0780-3

Symptom Media Films. (2023). https://symptommedia.com/film-library/ (accessed 23 January 2023).

Tan, T. Y., Wachsmuth, L., & Tugade, M. M. (2022). Emotional nuance: Examining positive emotional granularity and well-being. *Frontiers in psychology, 13*, Article 715966.

Thati, R. P., Dhadwal, A. S., Kumar, P., et al. (2023). A novel multi-modal depression detection approach based on mobile crowd sensing and task-based mechanisms. *Multimed Tools Appl, 82*, 4787–4820. https://doi.org/10.1007/s11042-022-12315-2

Tasnim, M., & Novikova, J. (2022). In *December). Cost-effective Models for Detecting Depression from Speech* (pp. 1687–1694). IEEE.

Trofimova, I. (2014). Observer bias: An interaction of temperament traits with biases in the semantic perception of lexical material. *PloS one, 9*(1), e85677.

Uddin, M. Z., Dysthe, K. K., Følstad, A., & Brandtzaeg, P. B. (2022). Deep learning for prediction of depressive symptoms in a large textual dataset. *Neural Computing and Applications, 34*(1), 721–744. https://doi.org/10.1007/s00521-021-06426-4

Wongkoblap, A., Vadillo, M. A., & Curcin, V. (2021). Deep learning with anaphora resolution for the detection of tweeters with depression: Algorithm development and validation study. *JMIR Mental Health, 8*(8), e19824.

World Health Organization. (2022). COVID-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide. https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide, https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide (2022, accessed 23 January 2023).

Depression, (2021), 2021, accessed 23 January 2023 https://www.who.int/news-room/fact-sheets/detail/depression, https://www.who.int/news-room/fact-sheets/detail/depression.

Wu, P., Wang, R., Lin, H., Zhang, F., Tu, J., & Sun, M. (2023). Automatic depression recognition by intelligent speech signal processing: A systematic survey. *CAAI Transactions on Intelligence Technology, 8*(3), 701–711.

Vázquez-Romero, A., & Gallardo-Antolín, A. (2020). Automatic Detection of Depression in Speech Using Ensemble Convolutional Neural Networks. *Entropy, 22*, 688. https://doi.org/10.3390/e22060688

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., ... & Pantic, M. (2013, October). Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge (pp. 3-10).

Yamamoto, M., Takamiya, A., Sawada, K., Yoshimura, M., Kitazawa, M., Liang, K. C., & Kishimoto, T. (2020). Using speech recognition technology to investigate the association between timing-related speech features and depression severity. *PloS one, 15*(9), e0238726.

Yang, Z., Jian, L., Qiu, H., Zhang, C., Cheng, S., Ji, J., & Li, K. (2021). Understanding complex functional wiring patterns in major depressive disorder through brain functional connectome. *Translational psychiatry, 11*(1), 526. https://doi.org/10.1038/s41398-021-01646-7

Yoon, J., Kang, C., Kim, S., & Han, J. (2022). D-vlog: Multimodal Vlog Dataset for Depression Detection. *Proceedings of the AAAI Conference on Artificial Intelligence, 36* (11), 12226–12234. https://doi.org/10.1609/aaai.v36i11.21483

Further reading

American Psychiatric Association. (2015). Depressive disorders: DSM-5® selections. American Psychiatric Pub.

Lee, D., & Yoon, S. N. (2021). Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. *International Journal of Environmental Research and Public Health, 18*(1), 271.

O'Connor, E. A., Whitlock, E. P., Gaynes, B., & Beil, T. L. (2010). Screening for depression in adults and older adults in primary care: an updated systematic review. Evidence Synthesis No. 75. AHRQ Publication No. 10-05143-EF-1. Rockville, Maryland: Agency for Healthcare Research and Quality, December 2009.

Oliveira, L. (2020). September). BioInfo@ UAVR at eRisk 2020: On the use of psycholinguistics features and machine learning for the classification and quantification of mental diseases. In *In Proceedings of the CEUR Workshop Proceedings* (pp. 22–25).