

REVIEW

OPEN

A systematic review on automated clinical depression diagnosis

Kaining Mao¹, Yuqi Wu¹ and Jie Chen¹✉

Assessing mental health disorders and determining treatment can be difficult for a number of reasons, including access to healthcare providers. Assessments and treatments may not be continuous and can be limited by the unpredictable nature of psychiatric symptoms. Machine-learning models using data collected in a clinical setting can improve diagnosis and treatment. Studies have used speech, text, and facial expression analysis to identify depression. Still, more research is needed to address challenges such as the need for multimodality machine-learning models for clinical use. We conducted a review of studies from the past decade that utilized speech, text, and facial expression analysis to detect depression, as defined by the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), using the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guideline. We provide information on the number of participants, techniques used to assess clinical outcomes, speech-elicitng tasks, machine-learning algorithms, metrics, and other important discoveries for each study. A total of 544 studies were examined, 264 of which satisfied the inclusion criteria. A database has been created containing the query results and a summary of how different features are used to detect depression. While machine learning shows its potential to enhance mental health disorder evaluations, some obstacles must be overcome, especially the requirement for more transparent machine-learning models for clinical purposes. Considering the variety of datasets, feature extraction techniques, and metrics used in this field, guidelines have been provided to collect data and train machine-learning models to guarantee reproducibility and generalizability across different contexts.

npj Mental Health Research (2023)2:20; <https://doi.org/10.1038/s44184-023-00040-z>

INTRODUCTION

Major depressive disorder is one of the most common diseases globally. It is estimated that 3.8% of the population is impacted by depression, and ~280 million people suffer from depression. The economic impact of depression is significant, even when compared to other medical conditions, such as cancer, cardiovascular diseases, diabetes, and respiratory diseases¹. Conventional methods for assessing and monitoring depression involve semi-structured interviews between patients and healthcare workers, which can be subjective and affected by bias (i.e., deemphasized or exaggerated symptoms), cognitive limitations (i.e., memory errors) and social stigma. The need for objective depression diagnosis, routine symptom monitoring, and timely treatment is widely recognized in the medical community. However, many people with depression have difficulty accessing psychological healthcare services due to geographical and financial barriers. As the World Health Organization (WHO) has reported, over 75% of individuals with depression in low and mid-income countries do not receive qualified psychotherapy². The shortage of well-trained healthcare professionals and the social stigma surrounding depression are major barriers patients face when seeking help. An automated depression assessment tool would assist in objective diagnosis and remote care to improve the quality of mental healthcare.

One potential solution to improve the objectivity of depression assessments and the quality of mental health services is to enhance mental health-related data collection and analysis via sensors and advanced machine-learning algorithms. Sensors that measure biological signals such as electrocardiogram, electroencephalogram, heart rate, and skin conductance may also be used to monitor biomarkers of depression^{3–15}. Researchers have developed models to detect depression and other psychological disorders by extracting features from interview videos, or audio

recordings^{16–28}. Toolkits such as OpenFace can extract facial landmarks, action units, face orientation, and eye gaze²⁹. Other modalities, such as neuroimaging data, have been used to predict the presence of Schizophrenia³⁰. Still, here we will focus only on non-neuroimaging modalities, such as acoustic, semantic, and facial modalities. Lastly, features extracted from social media and transcribed audio recordings have also been used to detect depression and stress^{31,32}. These previous studies provide valuable insights into new approaches for improved depression assessment.

A machine-learning algorithm for automated depression assessment can provide several benefits, including (a) supporting clinicians in making accurate diagnoses and providing effective treatment; (b) identifying individuals at risk for depression before they seek treatment from mental healthcare clinics and (c) tracking symptoms over time, both during and after treatment. These benefits are further discussed in the following:

One of the primary advantages of automated depression assessment is that it can help overcome barriers preventing people from receiving proper diagnoses and treatment for mental health issues in a timely manner. Mohr et al. reported that the main barriers patients face when seeking help are stigma, lack of motivation, time or availability constraints, and cost^{33–35}. Automated diagnostic tools could allow depression patients who have not yet sought help from healthcare professionals to evaluate their mental states remotely and receive online support from healthcare workers. In addition, these tools can be designed to customize treatment based on an individual's specific symptoms, improving treatment efficacy^{36,37}. These systems can also be utilized for mental disorder screening in various settings, such as universities, the military, and basic healthcare facilities.

Automated depression assessment systems can play an important role in helping doctors diagnose and make decisions

¹Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 2R3, Canada. ✉email: jc65@ualberta.ca

related to depression. Diagnosing depression can be difficult because symptoms may be episodic and multiple disorders may occur simultaneously, as demonstrated by the low inter-rater reliability³⁸ and test-retest reliability scores³⁹ in major depressive disorder diagnosis. In addition, it is also challenging to train a model to recognize any mental disorder, given that patients may have multiple disorders that present overlapping symptoms simultaneously. For example, over 50% of depression or anxiety cases co-occur with drug abuse or post-traumatic stress disorder (PTSD)⁴⁰. Because of these challenges, researchers have proposed prediction tools for suicidal thoughts and behaviors across many mental disorders (for a review, see ref. ⁴¹). The Research Domain Criteria, created by the National Institute of Mental Health, assists in distinguishing diagnoses and symptoms⁴². Consequently, we can train models to predict the probability of different mental disorders to assist in both diagnosis and early interventions. Furthermore, prediction tools can enable customized treatment plans based on multimodality features (genetic, behavioral, neuroimaging)^{43–45}. As a result, automated depression assessment models improve the efficiency of the healthcare systems, lower costs, and make treatment plans more customizable.

Automated depression assessment models can also enhance mental healthcare by allowing more frequent monitoring of symptoms, even in real time. Real-time monitoring enables individuals at risk for depression to be reminded to seek mental healthcare. Additionally, depression symptoms can change between healthcare appointments⁴⁶. Real-time monitoring can detect important signs related to suicidal or self-harm thoughts and conduct online psychotherapy. With real-time monitoring, patients and clinicians can monitor symptoms, conduct early intervention, and tailor treatment plans in a more personalized and timely manner.

Despite the potential benefits, these advantages have yet to be fully realized. Previous studies have relied on small, unrepresentative and non-clinical datasets, which are inadequate as they do not reflect real-world clinical situations. Most clinical datasets are limited in size and data type, for example, recorded in noise-controlled rooms, spoken in English, and restricted to adult participants. Moreover, models can be biased toward differences in the demographic characteristics of patients and healthy controls. For example, the model may tend to assign lower depression likelihood in male subjects because fewer male subjects have depression in the training set^{47–51}. Another significant issue in developing high-performance models is that their output depression probabilities are difficult to explain, limiting their clinical applications. This is because machine-learning models are often considered “black boxes”, meaning that their decision-making process is not easily interpretable by humans. While high-performing models can achieve high accuracy in depression prediction, the lack of transparency in how they made such decisions can be a barrier to clinical use. Thus, an explanation of the output probabilities of high-performance models is crucial for developing interpretable automated depression assessment models.

The study of speech patterns has been a research topic in identifying indicators of mental disorders since the 1920s. Emil Kraepelin, the founder of modern scientific psychiatry, reported that the voices of depressed patients were lower in pitch, sound intensity, and speech rate, and instead tend to be monotonous and hesitant, with shivering and whispering⁵². Moreover, acoustic features can be extracted across different languages, which is important for languages without pre-trained natural language processing models. In addition, speech recordings can be easily collected with smartphones and laptop computers instead of complex and costly equipment. With the advancements in speech recognition, especially its application for electronic medical records, speech recording will become more accessible for research purposes.

In research settings, depression is conventionally assessed using clinical depression rating scales. However, in clinical practice, the official psychiatric diagnosis is typically determined through a clinical interview, which may be semi-structured, with rating scales used as supplementary data to inform the diagnosis. However, these rating scales have limitations, as responses can be influenced by factors such as the patient’s emotional state, relationship with the clinician, and patient self-bias (e.g., participants may be more likely to exaggerate their symptoms)⁵³. With the advancement of machine-learning applied to text data from social media, new methods have emerged to address these limitations. Social media such as Twitter, Facebook, and Reddit provide a wealth of information about individuals’ feelings, thoughts and activities. Machine learning, especially text mining and sentiment analysis techniques, have become more accurate and intelligent, aiding mental healthcare providers in detecting depression. For instance, Pirina et al. and Yates et al. proposed machine-learning models to screen depression symptoms based on text data from Reddit^{54,55}. Researchers have recently developed models that utilize semantic features and other information extracted from social media platforms for detecting depression^{56–60}.

Several studies have been published on depression detection via facial image processing. The Audio/Visual Emotion Challenge and Workshop (AVEC) depression sub-challenge published papers with encouraging results on depression detection. Jan et al. achieved plausible results with a Motion History Histogram to extract Local Binary Pattern (LBP) and Edge Orientation Histogram (EOH) for depression identification⁶¹. Other features, such as head posture, blink rate, and eye gaze, are also reported to be effective in depression detection^{62,63}. Depressive individuals tend to display less nodding, avoid eye contact, and lower their heads more frequently than healthy individuals. Alghowinem et al. developed a support vector machine (SVM) to recognize depression and concluded that the head movements of depressed individuals are different from those of healthy controls⁶⁴.

This paper reviews recent research on the use of computational methods to predict major depressive disorder using acoustic, semantic, and facial features. This review is unique in that it uses the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines for extensive and rigorous evaluation of the latest research findings. By synthesizing and analyzing recently published data, this review offers important insights into the current state of the field and identifies key areas for future research. We would like to show how multimodal features could enhance mental healthcare, which provides insights into establishing connections between psychiatry and computing. Consequently, this review aims to (a) summarize results from previous publications using acoustic, semantic, and facial features to detect major depressive disorder; (b) characterize psychiatric disorders by identifying significant differences in acoustic, semantic, and visual features; (c) associate these multimodal features to depression symptoms and behaviors; and (d) summarize the challenges of automated depression assessment, make suggestions for future data collection and model training with better reproducibility and performance. Therefore, we propose that artificial intelligence may be a key tool for improving the assessment and treatment of depression through automated approaches.

METHODS

Inclusion criteria and literature search

The PRISMA guidelines were followed in this literature review, as shown in Fig. 1. Our goal was to search for articles published in the last ten years that included artificial intelligence methods for predicting the presence or severity of major depressive disorder by analyzing acoustic, semantic, and facial landmarks. Google

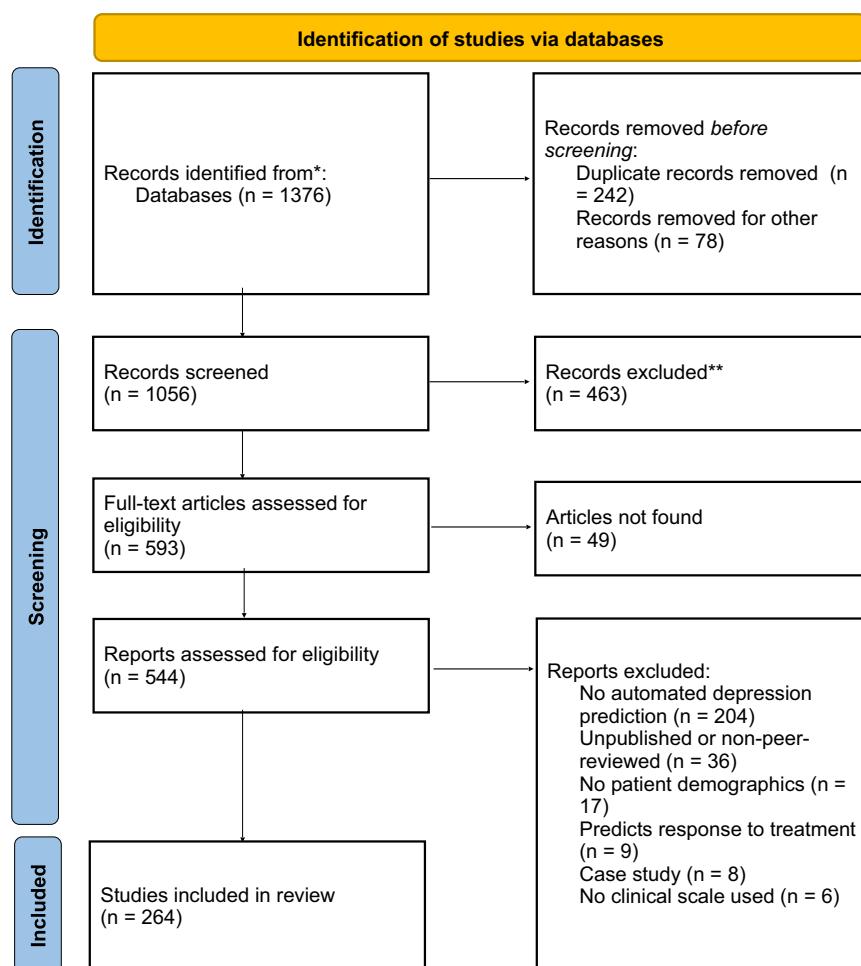


Fig. 1 PRISMA flow diagram of study inclusion and exclusion criteria in this review.

Scholar was used as the search engine for articles from 2012 to the present, queried between July 20, 2022, and May 20, 2023, excluding case studies, studies that solely used perceptual evaluation of speech, studies without a control group or clinical depression rating scales, non-peer-reviewed preprint and theses, and articles published before 2022 and having fewer citations than years of publication (e.g., articles published in 2019 with three citations, or articles published in 2017 with five citations would be included). We excluded certain articles that lacked comprehensive methodology or detailed results. In addition, we encountered cases where articles were published in both journals and conference proceedings, covering similar topics, methods, and results. Furthermore, some articles only focused on proposing methods for feature extraction without incorporating the training of models for depression detection. The search terms used to find relevant articles were: “allintitle: (“depression” OR “major depressive disorder”) + (acoustic OR acoustical OR speech OR voice OR vocal OR audio OR pitch OR prosody OR prosodic OR vowel) + (automated OR behavioral OR measures OR diagnosis)).” Articles related to depression caused by Parkinson’s Disease, autism, and substance overdose disorders were excluded. Replacement of the acoustic feature with the semantic feature and facial landmarks in the command resulted in the following search term with associated features: “allintitle: (“depression” OR “major depressive disorder”) + (semantic OR text OR interview OR transcript OR social media) + (automated OR behavioral OR measures OR diagnosis))” and “allintitle: (“depression” OR “major depressive disorder”) + (facial expression OR visual OR facial features OR

facial landmarks OR facial muscles)+ (automated OR behavioral OR measures OR diagnosis)).”

Information extraction was performed by reading the title, abstract, and conclusion. The following information was synthesized from each article: mental disorders, number of subjects, age range, optimal model, best metrics, type of validation, and predictive features.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

RESULTS

Summary of results

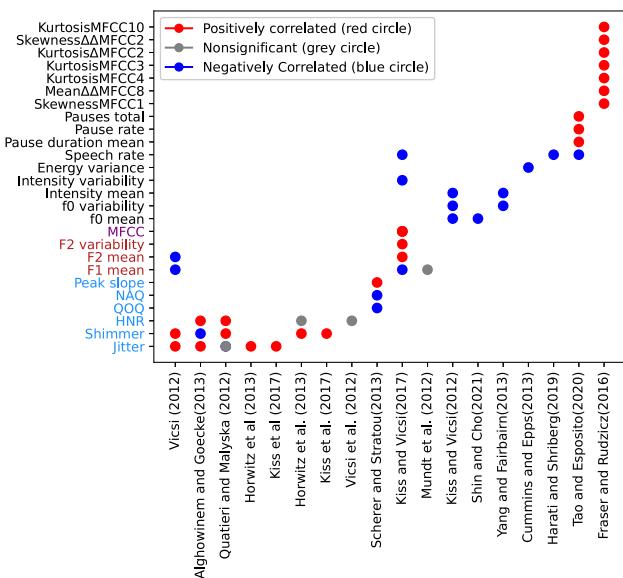
In total, 264 studies were included in the review. Table 1 summarizes the search results. Synthesized information can be found online <https://bit.ly/3DBQtZk>, <https://bit.ly/43Q6Yyy>, and <https://bit.ly/44IKaPv>, which can be extended by adding new studies on a blank row or fields on a blank column. Previous review and datasets-only articles were included in this study but are not included in Table 1.

Predictive acoustic features in major depressive disorder

With the publicly available code provided by Low et al.⁶⁵, we created Fig. 2, which provides a synthesis of the acoustic features investigated using machine learning. The table shows the acoustic

Table 1. Summary of literature review results.

Modality	Articles	Median dataset size (range)	Clinical assessment	Predictive models
Acoustic	140	189	36	140
Semantic	99	1046	2	81
Facial landmarks	25	49	16	21

**Fig. 2 Synthesis of acoustic feature analysis in major depressive disorder.** Acoustic features are sorted, such as vocal fold source features (blue), vocal tract filter features (red), spectral features (purple), and features related to prosody or melody (black). Features that are significantly higher in a psychiatric group than healthy controls or that correlate positively with the depression level receive a score of 1 (red), features that are lower or correlate negatively receive a score of -1 (blue), and nonsignificant or contradicting findings receive a score of 0 (gray). Features not studied in any studies are blank.

features found to be statistically different between the group with a mental disorder and the healthy control group or highly correlated with a diagnostic rating scale. Each cell in Fig. 2 represents the correlation between a specific acoustic feature and depression. For example, an acoustic feature that correlates positively with the disorder severity would be marked with a red dot, a negative correlation with a blue dot, and a nonsignificant feature with a gray dot.

Table 2 provides an overview of the key findings from previous studies on automated depression detection using acoustic features. One common finding among the studies is the relationship between acoustic volume and depression. Cummins et al.^{66,67} found that as the level of depression increases, the acoustic volume significantly decreases, indicating a potential acoustic marker for depression. Similarly, Harati et al.⁶⁸ reported that individuals with depression tend to have lower voices compared to the control group.

Another notable finding is the influence of gender on acoustic features related to depression. Cummins et al.⁶⁹ proposed gender-dependent formant features that outperformed acoustic-only features in depression detection. This suggests that gender-specific acoustic characteristics may play a role in accurately

Table 2. Predictive acoustic features in prior research publications.

Study	Key finding
Cummins et al. ⁶⁶	Decreased acoustic volume
	More concentrated MFCC space
Cummins et al. ⁶⁹	Gender-dependent formant features
Morales et al. ⁷⁰	Fundamental frequency
	Pronoun use and negatively valenced words
Scherer et al. ¹¹⁸	Tenser voice
Vicsi et al. ⁷¹	Jitter and shimmer values of vowels
	First and second formant frequencies
Marmor et al. ¹⁵⁰	Seeking care for voice problem
Cummins et al. ⁶⁷	Acoustic volume
	Probabilistic acoustic volume slope
Harati et al. ⁶⁸	Lower voices
	Variance in voice pitch
Kiss et al. ⁷²	Articulation rate
	Speech rate
	Pause lengths
	Formant frequency
Stasak et al. ⁷³	Use phonemes that require less effort
	Articulatory precision

detecting depression. Fundamental frequency (f_0) was found to be negatively correlated with depression level by Morales et al.⁷⁰. This finding indicates that changes in f_0 could serve as an indicator of depression. In addition, Vicsi et al.⁷¹ discovered that depressed individuals exhibited higher jitter and shimmer values in vowel production, along with lower first and second formant frequencies. These acoustic features may serve as potential biomarkers for depression detection. Kiss et al.⁷² highlighted the importance of speech rate, articulation rate, pause lengths, and formant frequency in detecting depression. They found that these acoustic features differed between individuals with depression and the control group, suggesting potential utility in automated depression detection. Stasak et al.⁷³ proposed that depressed individuals tend to use phonemes that require less effort and demonstrate decreased articulatory precision. These findings indicate that analyzing articulatory characteristics could provide valuable insights for depression detection.

It is important to note that these findings are based on previous studies and should be interpreted within the context of their respective methodologies and limitations. Further research is needed to validate and refine the use of these acoustic features for automated depression detection.

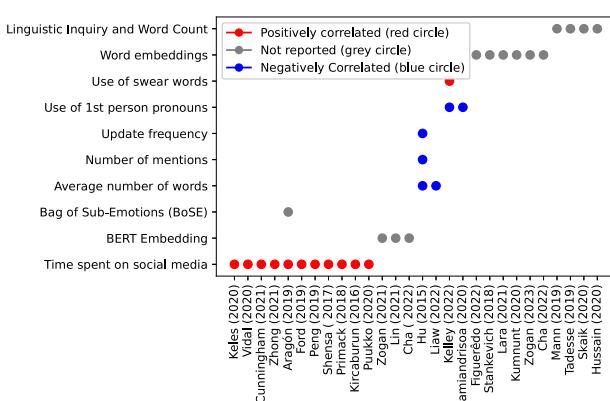
Predictive semantic features in major depressive disorder

This summary reviews studies on automated depression detection using language cues. Previous studies identified depression based on clinician diagnosis, patient self-reported mental status and online forum memberships. Clinician diagnosis means that depression levels were determined by a clinician based on interview transcripts or online posts. Among the 99 studies using language cues, only two identified depression based on the clinician diagnosis^{74,75}, while 77 studies used self-reported depression rating scales. The remaining 20 studies did not report which criteria were used to determine depression levels.

Social media and depression. Table 3 provides a summary of key findings from various studies examining the relationship between social media usage and depression. These studies employ a range

Table 3. Exploring the predictive relationship between social media usage and depression.

Author	Main findings
Hartanto et al. ¹⁵¹	Social media usage increases among depressive individuals
Figueiredo et al. ⁷⁸	Semantic mapping of emoticons improves the performance.
Stankevich et al. ⁷⁹	Future work needs to involve applying semantic role labeling to obtain better results.
Lara et al. ⁷⁷	DeepBoSE outperforms conventional Bag-of-Features(BoF) representations.
Hussain et al. ¹⁵²	Proposed depression lexicons that distinguish depressive individuals.
Ramiandrisoa et al. ⁸⁶	Analyzing users' social signals could be considered for further analysis.
Liaw et al. ¹⁵³	Topic modeling features such as liked tweets can be useful.
Guo et al. ⁸⁴	Fused the lexical features using a correlation-based metric to enhance prediction effectiveness.
Cui et al. ⁸³	Capture deep emotional information from the input embeddings with a pre-trained TextCNN.
Zogan et al. ⁸⁷	The model captures semantic features from user timelines for depression detection.
Tlatelpa et al. ⁸⁰	User characteristics and sentiment analysis improved depression detection performance.
Cha et al. ⁸²	Proposed lexicon features for depression detection.
Primack et al. ¹⁵⁴	Using multiple social media platforms is associated with depression.
Primack et al. ¹⁵⁵	Social media use is associated with the development of depression.
Vedula et al. ¹⁵⁶	Depressed users exhibit reduced online activities, increased negative sentiment, and self-focused pronoun usage.
Nesi et al. ¹⁵⁷	More frequent negative emotional reactions to social media are linked to more severe depression symptoms, especially among female subjects.
Thorisdottir et al. ¹⁵⁸	Time spent on social media has a stronger relationship with emotional distress among female subjects.
Ghosh et al. ¹⁵⁹	Depressed users frequently use negative words and mostly post late at night, in addition to increased use of personal pronouns and sharing personal events.
Aragon et al. ¹⁶⁰	Representations based on fine-grained emotions can more comprehensively capture users experiencing depression.
Puukko et al. ¹⁶¹	Depressed individuals increasingly use active social media during early and late adolescence.
Robinson et al. ¹⁶²	Depressed individuals are more likely to compare themselves to others and dislike being tagged in self-perceived unflattering pictures.
Choudhury et al. ¹⁶³	Social media can provide valuable indicators of depression onset, including decreased social activities, increased negative emotions, focus on personal and medical issues, and more frequent expressions of religious involvement.

**Fig. 3 Synthesis of semantic feature analysis in major depressive disorder.** Features that are significantly higher in a psychiatric group than healthy controls or that correlate positively with the depression level receive a score of 1 (red), features that are lower or correlate negatively receive a score of -1 (blue), and findings without reporting their changes receive a score of 0 (grey). Features not studied in any studies are blank.

of techniques and models to improve the detection and understanding of depression based on social media data. Several studies highlight increased social media usage among individuals with depression. Various advanced models, such as GRU models with knowledge-aware dot-product attention⁷⁶ and DeepBoSE⁷⁷, demonstrate improved performance in depression detection compared to conventional methods. In addition, semantic mapping of emoticons⁷⁸ and the application of semantic role labeling⁷⁹

are proposed as techniques to enhance detection accuracy. These findings highlight the potential of leveraging machine learning and natural language processing techniques to gain insights into mental health conditions through social media data.

Furthermore, the studies presented in the table emphasize the importance of considering multimodal data, user characteristics, and sentiment analysis for a comprehensive understanding of depression^{80,81}. They also propose the use of lexicon features and emotional information capture to improve depression detection^{82,83}. The fusion of lexical features and the development of bipolar feature vectors demonstrate promising results in enhancing prediction effectiveness^{84,85}. In addition, the studies suggest the potential of analyzing social signals and user timelines to capture semantic features for depression detection^{86,87}.

It is important to note that while these studies offer valuable insights, there are still ethical considerations that need to be addressed. Responsible data usage, privacy protection, and potential biases are crucial aspects that should be carefully considered when developing AI-based depression detection tools using social media data. Furthermore, future research should focus on the validation and generalization of these findings across diverse populations and cultures.

By synthesizing these findings (see Fig. 3), we have gained a deeper understanding of the potential of social media data in detecting and understanding depression. These insights can inform the development of effective mental health interventions, improve clinical practice, and contribute to the responsible and ethical usage of AI in this domain.

Problematic social media use. Table 4 provides a summary of the main findings from studies examining the association between problematic social media use and depression, including several

Table 4. The impact of problematic social media use on mental health.

Author	Main findings
Cunningham et al. ⁸⁸	Future research should focus on individuals with problematic social media use.
Shensa et al. ⁸⁹	Problematic social media use is strongly associated with depressive symptoms.
Woods et al. ⁹⁰	Adolescents who abused social media have a higher depression risk.
Radovic et al. ⁹¹	Over-sharing and stressed posting is associated with depression.
Ivie et al. ⁹²	Significant positive correlation between social media use and depressive symptoms.
Raudsepp et al. ⁹³	Excessive social media use was correlated with increased symptoms of depression.
Zhong et al. ¹⁶⁴	Breaks from social media use could have mitigated mental health trauma during the pandemic.
Haand et al. ¹⁶⁵	Addiction to social media is positively linked to depression.
Brailovskaia et al. ¹⁶⁶	Depressed individuals often intensively use social media to escape negative moods.
Jeri-Yabar et al. ¹⁶⁷	Excessive use of social media is associated with depressive symptoms.
Kircaburun et al. ¹⁶⁸	Social media addiction indirectly affects depression levels.

Table 5. Analyzing the efficacy of machine-learning models in detecting depression through social media data.

Author	Main findings
Yazdavar et al. ⁷⁴	Achieved 68% accuracy and 72% precision in identifying clinical depressive symptoms using a semi-supervised statistical model.
Zogan et al. ⁷⁵	Proposed a new computational model and achieved a recall of 0.904, precision of 0.909, and F1 score of 0.912.
Aragón et al. ⁵⁷	Using fine-grained emotions to obtain competitive results in comparison to state-of-the-art approaches.
Paul et al. ³²	AdaBoost classifier outperformed other methods for depression likelihood assessment.
Ricard et al. ¹⁶⁹	Leveraging community-generated content from social media can be informative for automated depression assessment.
Peng et al. ¹⁷⁰	Demonstrated that a multi-kernel support vector machine is the most appropriate approach to identifying depression in individuals using social media.
Aldarwish et al. ¹⁷¹	Trained a support vector machine based on term frequency to classify depression levels.
Chiong et al. ³¹	The proposed model effectively determines depression presence via social media posts, even when the training datasets do not contain depression-related words.
Burdisso et al. ¹⁷²	Introduced a general framework for early depression detection with less computational cost and higher interpretability.
Smys et al. ¹⁷³	A machine-learning model consisting of a support vector machine and a naive Bayes model can predict depression in its early stages.
Bucur et al. ¹⁷⁴	Latent semantic analysis shows a significant difference in writing topics depending on users' mental health.
Kayalvizhi et al. ¹⁷⁵	A word2vec pre-trained word embedding and random forest classifier achieved their best performance with a 0.877 F ₁ score.
Mann et al. ¹⁷⁶	Fusion model can detect moderate depression or higher with 0.92 recall and 0.69 precision.
Sadeque et al. ¹⁷⁷	Proposed a system to effectively detect depression using social media content with an accuracy of 88% and F ₁ score of 93%.
Hussain et al. ¹⁵²	Application accurately identifies indicators of depression in Facebook users with 94% accuracy.
Tadesse et al. ⁵⁸	Achieved 91% accuracy and F ₁ score of 93% with a multi-layer perceptron algorithm and combined features.
Fatima et al. ¹⁷⁸	Achieved an accuracy, recall, and precision of 91.7% using a combination of text-based features and machine-learning techniques.
Katchapakirin et al. ¹⁷⁹	Facebook behaviors can be used to predict depression levels with an accuracy of 85% and F ₁ score of 88.9%.
Shen et al. ¹⁸⁰	The model outperformed several baselines by 3% to 10% with an F ₁ score of 85%.
Li et al. ¹⁸¹	Proposed a correlation explanation learning algorithm to detect COVID-19-related stress symptoms.
Lin et al. ¹⁸²	Social media use is significantly associated with increased depression risk.

studies^{88–90} which consistently found a strong correlation. In addition, findings from refs. ^{91–93} indicate that over-sharing and stressed posting on social media is associated with depression. These findings collectively highlight the importance of recognizing problematic social media use as a potential risk factor for depressive symptoms. The results underscore the need for interventions and guidelines to promote healthier social media habits, particularly in vulnerable populations such as adolescents and university students. It is important to acknowledge the limitations of the studies summarized in Table 4, including factors like sample size, study design, and generalizability. Future research should consider longitudinal studies to examine the long-term effects of excessive social media use on depression and explore

the underlying mechanisms of this association.

Overall, the findings contribute to our understanding of the complex relationship between social media use and depression, providing valuable insights for mental health promotion and clinical practice.

Machine-learning models. Automated depression detection algorithms have been a subject of study by various researchers. Table 5 provides a summary of these studies, highlighting machine-learning models and performance metrics employed. Salas et al. conducted a comprehensive review of previous studies that utilized language cues and found that word embedding was the most commonly used linguistic feature extraction method, while

the support vector machine was the most prominent machine-learning model⁹⁴. Liu et al. summarized the findings of studies focused on using machine-learning methods to detect depressive symptoms in social media data, highlighting their potential as complementary tools in public mental health practice⁹⁵. Yazdavar et al. developed a semi-supervised statistical model to assess the alignment between depressive symptoms expressed on Twitter and medical findings reported via the Patient Health Questionnaire (PHQ-9), achieving 68% accuracy and 72% precision in identifying clinical depressive symptoms⁷⁴. Zogan et al. introduced a computational approach for automatically identifying depression using a hybrid extractive summarization technique applied to tweets, achieving high recall, precision, and F_1 score⁷⁵. Aragón et al. proposed a novel representation called bag of sub-emotions (BoSE) that improved the detection of depression using fine-grained emotions, demonstrating competitive results compared to existing approaches⁵⁷. These studies collectively emphasize the significance of linguistic features, word embeddings, and machine-learning models in automated depression detection.

However, some weaknesses and variations in the literature have been raised. McCrae et al. conducted a review of studies examining the relationship between social media use and depression symptoms, highlighting the need for comparative analysis due to variations in methods, sample sizes, and results across studies⁹⁶. They also suggested that future research should incorporate longitudinal analysis, as most studies were cross-sectional. Similarly, Heffer et al. found no predictive association between social media use and depressive symptoms over time, challenging the assumption that social media use leads to depressive symptoms⁹⁷. These contrasting perspectives call for further investigation and highlight the complexity of the relationship between social media use and depression.

In conclusion, while automated depression detection algorithms show promise, the field still faces challenges in terms of standardization, methodological variations, and the need for longitudinal analysis. Future research should address these limitations, conduct a comparative analysis, and explore the intricate mechanisms underlying the relationship between social media use and depression. In addition, ethical considerations and the potential impact of using social media data for mental health assessment should be carefully examined. Advancements in this field can contribute to the development of effective and reliable tools for early detection and intervention in depression.

Privacy issues and other factors. Ford et al. investigated social media users' opinions on providing mental healthcare services for depression-vulnerable individuals by analyzing social media content. Their survey indicated that social media users post negative content during low moods. They could see the benefits of identifying depression using social media content but did not believe that the risks of privacy breaches outweighed these benefits. In this survey, most participants consider identifying depression symptoms using social media content is intrusive and would not grant permission to researchers to conduct linguistic analysis⁹⁸.

Gender also affects text patterns and must be accounted for when developing depression detection models. Hou et al. investigated the gender differences in depression and explored associated factors during the COVID-19 pandemic among Chinese social media users. Their findings showed an increased prevalence of depression and anxiety in the Chinese population, with females more likely to experience more severe symptoms than males⁹⁹.

Several studies have built text corpora from social media, which were used to train baseline datasets, thus allowing other researchers to develop more efficient prediction models. Choudhury et al. built a large collection of tweets from individuals clinically diagnosed with depression. They developed a support

vector machine with a radial basis function kernel to characterize depression levels in populations¹⁰⁰. Narynov et al. presented a dataset collected from social network platforms that are commonly used by the youth of the Commonwealth of Independent countries. They demonstrated that the dataset has high validity and can be used for further research in mental health¹⁰¹. These studies contribute to our understanding of social media as a tool for mental health analysis and the importance of gender differences in depression research.

Predictive facial features in major depressive disorder

Table 6 provides a summary of previous studies on automated depression detection using facial features. The studies examined various aspects of facial expressions and their relationship to depression. Among the 18 studies that utilized facial landmark features, 13 studies identified depression based on clinician diagnosis, while four studies used self-report depression rating scales. One study did not report the specific criterion used to determine depression levels.

Li et al.¹⁰² proposed a deep residual regression model that evaluated depression levels. Their findings indicated that enhancing techniques can significantly improve prediction performance by reducing the influence of external factors, such as lighting and head pose. Wang et al.⁶² analyzed facial expressions in videos for automated depression diagnosis. Their study demonstrated the effectiveness of facial analysis, achieving an accuracy of 78%, recall of 80%, and an F_1 score of 79%. Hao et al.¹⁰³ investigated optimal methods of depression detection using contextual temporal information. Their proposed bidirectional long-short-term memory network (LSTM) with an attention mechanism achieved an accuracy of 82% and an F_1 score of 81%. Hunter et al.⁶³ evaluated the eye-tracking patterns of individuals with non-clinical depressive symptomatology in processing emotional expressions, revealing distinct differences compared to healthy individuals.

In addition to the studies mentioned in the original paragraph, several more recent studies provide valuable insights into automated depression detection using facial features. For example, Liu et al.¹⁰⁴ proposed the Part-and-Relation Attention Network, which outperformed state-of-the-art models with smaller prediction errors and higher stability. Hamid et al.¹⁰⁵ designed a hybrid model that integrates electroencephalogram (EEG) data and facial features, surpassing existing diagnosis systems. Nasir et al.¹⁰⁶ explored multimodal classification systems using geometrical facial features, indicating potential for robust and knowledge-driven depression classification. Dai et al.¹⁰⁷ proposed a multimodal model that achieved superior performance on multiple datasets, emphasizing the accuracy of the visual model. Shangguan et al.¹⁰⁸ demonstrated that video stimuli and an aggregation method can be effective for automatic depression detection.

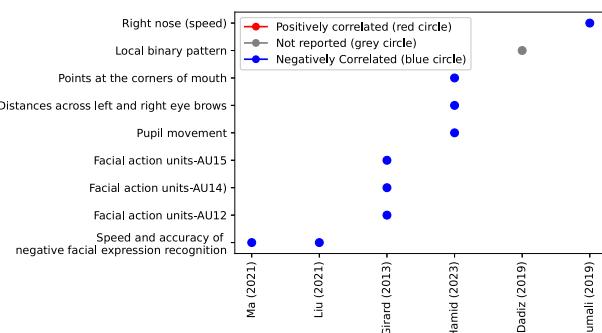
Overall, the summarized studies (see Fig. 4) highlight the significance of facial expressions in automated depression detection. They showcase various approaches, including deep learning models, multimodal techniques, and analysis of specific facial features. These findings contribute to the understanding of how facial expressions can serve as valuable indicators for detecting and diagnosing depression.

DISCUSSION

Most studies in this literature review adopted automated speech feature extraction to assess major depressive disorder. This is probably due to the Audio/Visual Emotion Challenge Workshop (AVEC) competitions, which provide automated extracted audio and video features to predict the severity of these conditions. Many other studies then used the public datasets in competitions like Distress Analysis Interview Corpus Wizard of Oz (DAIC-WOZ).

Table 6. Exploring the predictive relationship between facial expressions and depression.

Author	Main findings
Li et al. ¹⁰²	A deep residual regression model to evaluate depression levels using enhancement techniques can reduce the influence of external factors on the image, significantly improving prediction performance.
Wang et al. ⁶²	Facial analysis is effective in automated depression diagnosis with an accuracy of 78%, recall of 80%, and F1 score of 79%.
Hao et al. ¹⁰³	A bidirectional LSTM network with an attention mechanism achieved an accuracy of 82% and F1 score of 81%.
Hunter et al. ⁶³	Individuals with depressive symptomatology showed a different eye-tracking pattern in processing emotional expressions.
Jan et al. ⁶¹	The linear regression method applied to the AVEC 2014 dataset can predict BDI score using natural facial expressions.
Mohan et al. ¹⁸³	The proposed LSTM had the highest accuracy compared to other baselines.
Lee et al. ¹⁸⁴	An accessible depression diagnosis system using real-time object recognition and facial expressions obtained with a smartphone camera.
Liu et al. ¹⁰⁴	Proposed Part-and-Relation Attention Network for depression recognition, which outperforms state-of-the-art models with smaller prediction errors and higher stability.
Hamid et al. ¹⁰⁵	Designed a model for depression detection using electroencephalogram (EEG) and facial features. A hybrid model is proposed, outperforming existing diagnosis systems.
Nasir et al. ¹⁰⁶	A multimodal classification system for depression detection using geometrical facial features. The proposed visual feature sets show potential for robust and knowledge-driven depression classification.
Dai et al. ¹⁰⁷	A multimodal model with high performance on the AVEC 2013, AVEC 2014, and Emotion-Gait datasets. They concluded that the visual model is accurate.
Shangguan et al. ¹⁰⁸	An aggregation method which achieved comparable performance to 3D models with fewer parameters. The study suggests that video stimuli can be used for automatic depression detection.
Sumali et al. ¹⁸⁵	Significant differences were observed in facial landmark features (e.g., average right nose (speed), median left ear top (speed), and left pupil-right pupil positions) between healthy and depressive volunteers.
Dadiz et al. ¹⁸⁶	The uniformed local binary pattern extracted from videos for depression detection focuses on specific facial areas.

**Fig. 4 Synthesis of visual feature analysis in major depressive disorder.** Features that are significantly higher in a psychiatric group than healthy controls or that correlate positively with the depression level receive a score of 1 (red), features that are lower or correlate negatively receive a score of -1 (blue), and findings without reporting their changes receive a score of 0 (gray). Features not studied in any studies are blank.

Of the 264 studies in this review, 39% used DAIC-WOZ or AVEC datasets.

Most of the studies used some form of cross-validation for evaluating the performance of the trained models. However, only some studies used held-out test sets, which means that most models' reported performance may not generalize well. Without a held-out test set, performance may drop from the development set to the test set, as has been observed in AVEC competitions^{109–111}. In contrast, models that used held-out test sets generally performed better on the test set¹⁷. In addition, Zhang et al. achieved performance close to the state-of-the-art on the AVEC-2019 dataset (mean absolute error = 5.77 on the test set) by using automated extracted features and a random forest classifier¹¹¹. This suggests that the performance of a depression detection model is dependent on the dataset size, preprocessing strategy, feature engineering, and the model itself, which are all

determined by the dataset used for training. Different models applied to different datasets can lead to different complexity-accuracy tradeoffs, and there is no universal best model.

When studying the significance of acoustic features for major depressive disorder, many automated extracted features have been found to be predictive, and their correlations may be useful for making depression predictions^{70,106,112,113}. As a result, we examine the connection between these automated extracted features and the observable symptoms of major depressive disorder.

Associating acoustic features to depression symptoms

Many studies have reported that individuals with depression have lower values of the fundamental frequency F_0 and its range, which indicates that their speech becomes monotonous^{70,112,114}. In addition, acoustic features such as jitter and shimmer^{115,116} have been observed to increase with the severity of depression, which may be a result of slower thought, reaction, and physical movement in people with depression.

Review for data collection

The datasets used in automated depression detection studies vary greatly in size, participants' demographics, depression rating scales, the task used to elicit emotion, and the interview environment. Therefore, the performance of a detection model can be misleading if the dataset used for training is not representative of the studied population. In this review, we discussed the data collection strategies used in these studies to elicit emotions, record videos, and maintain participant privacy while avoiding confounding factors such as clinician questions and patient responses.

Identifying the presence of comorbidity

Many previous studies have not reported comorbidities¹¹⁷ which presents additional challenges when developing automated depression detection models. However, Scherer et al. discovered a strong association (Pearson's $r > 0.8$) between scores for

depression on the PHQ-9 scale and scores for PTSD on the PTSD Checklist-Civilian Version (PCL-C) in the DAIC-WOZ dataset¹¹⁸. Only a few articles stated that they excluded individuals with comorbidities from their study, such as ref. ¹¹⁹. To improve the data quality through consideration of comorbidities, researchers should use multiple mental disorder rating scales when collecting data. Additionally, researchers should develop and compare models trained with and without comorbidities to better understand the impact on the model performance.

Factors to consider in recruiting control groups

When selecting control groups for depression studies, it is important to ensure that individuals in the control group do not match any diagnostic criteria for other pathological conditions. For example, an individual may not be assessed as having depression based on their depression rating scale, but they may be assessed as suffering from PTSD that affects their speech patterns. Age, gender, first language, comorbidities, brain injury, respiratory disorders, and drug abuse can also affect speech and facial landmark patterns. In addition, variables such as education level, race, medication, and gender^{120–122} can also affect speech patterns. Hert et al. have reported that antipsychotic therapies may lead to dyskinesia, an involuntary movement of facial muscles that affects speech and facial landmarks¹²³. Therefore, individuals with a history of antidepressant medication should be excluded or reported in depression detection studies. Other variables such as gender, age, and education level can be adjusted via propensity score matching if they are statistically different between the depressive and healthy control groups.

Self-report depression rating scales: pros and cons in depression diagnosis

The traditional method for diagnosing depression is via a clinical evaluation by a registered psychologist, which is considered the gold standard compared to self-report depression rating scales. However, clinical diagnosis can be costly and subject to the experience and expertise of the clinician, leading to lower inter-rater reliability³⁸. Most studies included in this review relied on self-reporting depression rating scales instead of clinical evaluation, such as AVEC^{121,122} and DAIC-WOZ¹²⁰. When using these self-rating scales, the task becomes predicting the self-report rating scale rather than a clinical diagnosis, which may not align with a clinician's evaluation. On the other hand, using open-source datasets for research can improve reproducibility and objectively compare model performance.

Eliciting emotions in depression diagnosis

Choosing appropriate tasks for eliciting emotions is crucial, as specific features may be linked to certain depression rating scales but not others. In Table 7, we summarize the tasks used in previous articles and their advantages. Kane et al. proposed that sustained vowels are optimal for estimating glottal source features because it can be difficult to identify voiced sections in free speech¹²⁴. Scherer et al. demonstrated that the voices of participants with moderate to severe depression are tenser than those of healthy participants¹¹⁸. Alghowinem et al. proposed that spontaneous speech leads to better results for most features than reading speech and that the first few seconds of speech perform better than the entire recording¹²⁵. Another interesting approach to emotion elicitation is to use virtual agents for interviews, which can reduce data collection costs and can be less stressful for participants when discussing their symptoms. Multiple articles have reported successes in developing virtual interviewers^{126–128}, and the widely used AVEC challenges have also adopted virtual interviewers.

Diarization of speech segments in interview recordings: methods and considerations

It is common practice to separate the speech segments of the participants from interview recordings to train depression prediction models. This process is commonly referred to as diarization. The participant's speech can be extracted using a microphone next to each speaker. If participants have headsets with lapel microphones during the interview, their voiced sections can be easily extracted, which may make some participants uncomfortable. Desk microphones can also be used, but they can introduce confounds because they are not targeted and make it difficult to separate participants' speech in the data processing. We suggest using two desk microphones next to each speaker with a sound barrier between them. It is important to record all metadata after the interview in a separate spreadsheet, such as participant ID, group, task, and other demographic information.

Ensuring privacy in interview recordings

Clinicians should obtain verbal or written consent from the participants before interviewing. Participants must be informed that their interview recordings and demographic information may be distributed, pre-processed, and used for training machine-learning models for academic research purposes. Even if participants grant permission for their data to be further processed, researchers must minimize the risk of data leakage because the interview audio (or video) recordings may contain sensitive information. To address this, researchers can share only the automated extracted speech and facial features rather than the raw interview recordings. If hackers were to gain access to the interview data, it would be impossible to reconstruct the original interview recordings using only the automated extracted features. Additionally, researchers can train the depression prediction model in real-time or use bone conduction microphones, which only record acoustic features without speech content¹²⁹, but this limits researchers to training semantic models. Edge computing can also be a solution to improve privacy by allowing computation to be performed on the participants' devices, with only the trained models being returned to the researchers, not the data.

Review of machine-learning models

We believe that a well-trained depression prediction model can accurately detect the disorder in a randomly selected individual, regardless of the environment in which the individual is being interviewed. The participant may be of a different age and use a different language or accent, but the depression prediction model should still be able to generalize and be robust to these environmental factors. However, most previous studies have only been designed to detect mental disorders in new individuals collected in similar settings. Our suggestions for future studies to improve robustness and generalization are as follows.

Data preprocessing

Automatic speech recognition (ASR) can transcribe speech into transcripts for training semantic-based depression prediction models. ASR can also filter unvoiced sections and noise in the interview audio (or video) recordings. In most in-person interviews, two speakers are present, and the clinicians' segments can be discarded if the ASR system includes automatic diarization. To prevent overfitting, techniques like dimensionality reduction or feature selection should be applied to the training and test sets during preprocessing. This will help ensure that the model is not overly influenced by the specific characteristics of the training data and can be generalized to new data.

Table 7. A comparative study of different speech-eliciting tasks.

Task and examples	Advantages
Constrained	Repeating "PATAKA" ^{187,188}
	Sustained vowel ¹⁸⁹
	Counting ¹⁹⁰
	Reading
	• The "Nordwind" passage ^{191,192}
	• Rainbow passage ¹⁸⁹
Free speech	• Emotion-evoking movie clips ¹⁹³
	Monologue
	• Describing, memory recalling ¹⁹⁴
	Dialogue
	• Semi-structured interviews ¹²⁰
	• Phone conversations ¹⁹⁵

Automated feature extraction

The most commonly used automated feature extraction tools in the studies we reviewed were openSMILE, COVAREP, pyAudioAnalysis, and openEAR. As shown in Table 8, some features were found to be predictive in multiple studies. To ensure the deep learning models converge, it is recommended to standardize or normalize features as they may be in different scales. Before training the model, we recommend performing exploratory data analysis or visualization to better understand how these features characterize mental disorders. This can help inform the selection and preprocessing of features, as well as the design of the model.

Evaluate models with small datasets: bootstrapping and K-fold cross-validation

To avoid overfitting the model on the test set, we typically evaluate the trained model on the held-out test set only once. However, when training a model for predicting depression using a small dataset (e.g., around 100 data points), which is commonly seen in the medical field, a 20% held-out test set or K-fold cross-validation can decrease the number of samples available for training. In addition, a small test set is unlikely to represent the entire population accurately. As a result, we suggest using repeated bootstrapping to evaluate the depression prediction model, which provides a distribution of performance metrics with mean and standard deviation. However, given the computational complexity of deep learning models, the bootstrapping method may not be feasible. When working with a small dataset, K-fold cross-validation can be a viable alternative to bootstrapping, as deep learning models tend to need a large number of data points which reduces the need for bootstrapping.

Evaluating the performance of depression prediction: best practices and considerations

Performing better than chance does not indicate the model learned from the training data, and the resulting metrics must be generalizable and statistically significant for clinical use. Aloshtan et al. demonstrated that their accuracy is always better than chance to a statistically significant extent¹³⁰. However, to further prove generalizability, we suggest performing a permutation test in future works, where models are trained on permuted labels to evaluate the model's performance based on mistaken labels, which is often better than chance. A statistical test can then determine if the difference between the permuted and non-permuted scores is statistically significant. On the other hand, clinical datasets can be imbalanced, with a greater number of healthy cases compared to the population of individuals with depression. In this case, using the accuracy of the classification

model as the sole metric to evaluate its performance may not be objective since it will be biased towards predicting every sample as negative. To evaluate model performance more objectively, metrics such as the F_1 score, precision, recall, and area under the curve (AUC) should be considered. These metrics account for class imbalance and provide a more balanced view of model performance. Saito et al. have shown that the precision-recall curve is more useful than the receiver operating characteristic curve when evaluating binary classifiers on imbalanced datasets¹³¹. In addition to the precision-recall curve, metrics such as root mean square error (RMSE), mean square error (MSE), and the coefficient of determination (r^2) are commonly used to evaluate the performance of regression models for predicting depression scores. In the AVEC competition, the performance of baseline models was evaluated using the concordance correlation coefficient (CCC), which takes into account changes in scale and includes measures of both precision and accuracy¹³². It is generally helpful for other researchers to see a range of metrics when evaluating the performance of a model, as this allows for more objective comparison. A model's performance must be generalizable and statistically significant to be truly useful in a clinical setting.

Explainable depression detection model

Recent evidence suggests that individuals may lack trust in black box models and that these models may cause harm in high-stakes decision-making processes^{133–135}. As a result, researchers are exploring ways to explain the decision-making processes of algorithms better^{136,137} through publications and software packages implement explainable machine-learning models^{138,139}. By providing explanations of a model's feature contributions, clinicians can gain a better understanding of depression and improve the model itself. Lundberg et al. proposed an additive method for evaluating feature contributions, which calculates the difference in the model's output when a given feature is considered versus not considered¹³⁹. This can provide valuable insights into the factors influencing the model's predictions. Once high-impact features have been identified, we can retrain the model using only these features to evaluate their performance. In some of the reviewed articles, we observed that while the studies presented excellent feature engineering for distinguishing between groups, they lacked quantitative analysis to support their findings. In addition, we suggest linking changes in the automated extraction of features to mental disorder symptoms to provide a more comprehensive view of the model's performance. By combining these approaches, we can better understand the factors influencing the model's predictions and improve its

Table 8. An overview of acoustic features. for more details, see the cooperative voice analysis repository (COVAREP).

Acoustic feature	Description
Source features	Features reflecting airflow from the lungs through the glottis (i.e., glottal features) or vocal fold vibrations (i.e., voice quality features), which is the sound source later filtered by the vocal tract following the source-filter theory of speech production.
Jitter (%)	Deviations in the consecutive lengths of the f_0 period, which suggests irregular and uneven vocal fold vibrations.
Shimmer (%)	The variation in the peak amplitudes of consecutive f_0 periods, which implies unevenness in voice loudness.
Tremor (Hz)	The number of occurrences of the most powerful low-frequency fundamental frequency-modulating element within a defined examination range.
Harmonics-to-noise ratio (HNR) (dB)	Ratio between f_0 and noise components, which indirectly correlates with perceived aspiration.
Frequency disturbance ratio (FDR) (%)	The average relative value of the frequency variation over 5 to 5 cycles (calculated using an average of five data points).
Amplitude disturbance ratio (ADR) (%)	Relative mean amplitude value over a set of windows.
Quasi-open quotient	Ratio of the vocal folds opening time. Functional dysphonias often reduce QOQ range.
Normalized amplitude quotient (NAQ)	A measurement that compares the amplitude between the highest and lowest points of the differentiated flow glottogram to the amplitude of the negative peak and normalizing it with respect to the period time. It can be used as an approximation of glottal adduction.
Peak slope	Slope of the regression line that is fit to \log_{10} of the maxima of each frame.
Filter features	The resonant properties of the vocal and nasal tracts filter the sound source from the vocal folds: the filter attenuates certain frequencies and strengthens others by the shape of the vocal and nasal tracts.
F_1 mean (Hz)	First peak in the spectrum that results from a resonance of the human vocal tract.
F_2 mean (Hz)	Second peak in the spectrum that results from a resonance of the human vocal tract.
F_1 variability (Hz)	Measures of dispersion of F_1 (variance, standard deviation).
F_2 variability (Hz)	Measures of dispersion of F_2 (variance, standard deviation).
F_1 range (Hz)	Difference between the lowest and highest F_1 values.
Vowel space	F_1 and F_2 2D space for the vowels.
Linear predictive coding (LPC) coefficients	Coefficients that best predict the values of the next time point of the audio signal using the values from the previous n time points, which is used to reconstruct filter properties.
Spectral features	Features characterizing the frequency distribution of the speech signal at a particular moment in time.
Mel-frequency cepstral coefficients (MFCCs)	The coefficients derived by analyzing the Mel-spectrum of the log-magnitude of an audio segment.
Prosodic features	Changes over longer segments of time, which is perceived in the rhythm, stress, and intonation of speech.
f_0 mean (Hz)	Fundamental frequency: lowest frequency of the speech signal, perceived as pitch (mean, median).
f_0 variability (Hz)	Measures of dispersion of f_0 (variance, standard deviation).
f_0 range (Hz)	Difference between the lowest and highest f_0 .
Intensity (dB)	Defined as the acoustic intensity (i.e., power carried by sound per unit area in a direction perpendicular to that area in decibels relative to a reference value, perceived as loudness).
Intensity variability (dB)	Measures of dispersion of intensity (variance, standard deviation).
Energy velocity	Measured as the mean-squared central difference across frames and may correlate with motor coordination.
Maximum phonation time (s)	The mean of three attempts of the following measure is taken: the maximum time during which phonation of a vowel is sustained as long as possible with an upright position, deep breath, and a comfortable pitch and loudness.
Speech rate	Number of speech utterances per second over the duration of the speech sample (including pauses).
Articulation rate	Number of speech units per second throughout the speech sample (excluding pauses).
Time talking (s)	Sum of the duration of all speech segments.
Utterance duration mean (s)	Mean duration of utterance length.
Pause duration mean (s)	Mean duration of pause length.
Pause variability (s)	Measures of dispersion of pause duration (variance, standard deviation).
Pause rate (s)	Total length of pauses divided by the total length of speech (including pauses).
Pause total (s)	Total duration of pauses.

performance. On the other hand, there are arguments on whether a complex, hard-to-explain model with good performance should be discarded in favor of a simpler, easy-to-interpret but lower-performing model¹⁴⁰. From our perspective, since complex but high-sensitivity examination methods would not surpass the use of less sensitive models, these models would not be abandoned if they are proven effective in clinical trials. Thus,

validating and explaining complex models will be important in this field.

Ensuring reproducibility in automated depression detection

Reproducibility is a critical issue in machine learning, particularly when artificial intelligence is applied to healthcare^{141,142}. One obstacle to reproducing previous studies is that clinical datasets

are not always available for redistribution. As we mentioned in Section Ensuring privacy in interview recordings, automated extracted features can be shared without violating privacy concerns. However, sharing the code used for training and evaluating the model is also important. Even when the code and data are publicly available, other researchers may still have difficulty reproducing the results due to differences in the software environment. To address this issue, we suggest that researchers use containers, such as Docker, which include the data, code, and environment in one package that can be easily redistributed. This will make it easier for other researchers to reproduce the results, ultimately accelerating the advancement of automated depression detection.

Evaluating models from competitions in automated depression diagnosis

About 39% of the studies included in this review were developed during or after the AVEC. Some of these studies introduced innovative approaches to feature engineering and model architecture. However, since teams were allowed to submit their models multiple times, overfitting the test set is a risk. The test set provided in the AVEC competition is relatively small, which means that the state-of-the-art model can outperform the second-best model by chance due to overfitting. Therefore, we cannot simply assume that a model that performs slightly better on the test set will also generalize well to new data. We believe that multiple comparison corrections should be applied to evaluate the performance of the models, and simpler models should be prioritized¹⁴³. This will help to ensure that the results are more reliable and can be more confidently applied in a clinical setting.

Due to the limited number of studies we were able to review and include in this review, we only searched for keywords in the titles of articles rather than in other sections. The screening process of the articles involved reading the title and abstract. Only articles that were relevant to using machine learning to detect depression and had "machine learning" or related terms in the title were included, and the others were excluded. Our study may not have captured all relevant articles on this topic, and other studies not focused specifically on automated depression diagnosis using machine-learning methods may have been missed.

Future work

Cross-cultural generalization. While machine-learning models are optimized to work well on the training data, it is not always clear how well they will generalize to new sample sets where different ages, languages, socioeconomic and education levels are reported. Alghowinem et al. developed a cross-cultural depression recognition model and evaluated it on datasets in English and German¹⁴⁴. Some previous studies have also investigated the effects of different smartphones on the quality of acoustic features, which can impact the accuracy of depression prediction^{145,146}. Mitra et al. examined the effects of noise and reverberation on depression detection using speech and found that spontaneous speech performed better than read speech¹⁴⁷. These studies highlight the importance of considering factors that may affect the performance of machine-learning models for automated depression diagnosis.

Ethical considerations in automated depression detection: addressing risks and ensuring responsible use. Automated depression detection can benefit society by reducing the workload of the healthcare system, preventing suicidal or self-harm behaviors, and enabling law enforcement authorities to track abnormal behaviors. However, the use of automated depression detection also raises some ethical concerns. For example, insurance companies and employers may use the results to evaluate

candidates without their knowledge or consent and reject them if a mental disorder is present or likely to develop in the future. In addition, it can be difficult for individuals to fully understand the implications of consent forms, which can further complicate the ethical considerations surrounding automated depression detection¹⁴⁸. To ensure that automated depression detection is used ethically in clinical settings, researchers should provide clear and understandable explanations of how the collected data will be used. Participants should also have the right to revoke permission to use their data at any time. Like other developing technologies, these systems may be vulnerable to abuse and have unexpected side effects. As researchers, engineers, and clinicians, it is our responsibility to educate the public and policymakers about the potential benefits and harms of automated depression detection to both prevent abuse and further advance these techniques, which have the potential to help many people.

Leveraging machine learning for advancing psychiatry. With this review paper, we aim to demonstrate the potential for psychiatry to benefit from advances in machine learning. Many individuals have difficulty accessing qualified mental healthcare or may be hesitant to seek psychotherapy due to stigmatization¹⁴⁹. Automated depression detection models can provide an accessible and efficient method for early screening, which can help individuals determine that they may need professional healthcare. In addition, psychiatric visits often include interviews that can be recorded in video or audio format, which provides a wealth of data that can be used to associate mental health assessments with acoustic, semantic, and facial features. By following the guidelines outlined in this paper for collecting and analyzing this data, we hope to enable new collaborations between clinicians and machine-learning engineers to advance our understanding of mental health disorders.

CONCLUSION

We reviewed 264 studies that measure acoustic, semantic, and facial landmark features to distinguish between individuals with and without mental health disorders using either null hypothesis testing or predictive machine-learning models. Our synthesis includes significant and nonsignificant features across audio, text, and facial modalities, as well as those correlated with the severity of depression. We also provide guidelines on collecting data, preventing confounding factors, protecting privacy, selecting speech-eliciting tasks, and improving machine-learning model generalizability and reproducibility. We also found a few studies have been conducted on post-traumatic stress disorder, bipolar disorder, and postpartum depression, thanks to open-access research datasets provided by the AVEC and DAIC. Competitions provide a useful framework for comparing innovations under the same conditions, such as using the same dataset and metrics. This approach enables researchers to evaluate the model's performance using a held-out test set and estimate overfitting; however, overfitting is still a concern in such competitions. In addition, these competitions make it possible for future studies to be conducted using the same dataset, which facilitates comparisons and helps advance the field. Based on their proven effectiveness, we encourage the collection of open datasets, particularly distributing datasets through competitions. These are highly productive in advancing research in various fields. While productivity is important, reproducibility is also critical. Since the studies in this review involve building computational models, the associated data and code should be shared, ideally through containers. This allows others to test the claims made by these studies and contribute to the development of these models in a collaborative manner. Moreover, conducting more research on multiple datasets may help enhance the models' generalizability and reconcile conflicting results regarding crucial and predictive

features. This approach could lead to more robust and reliable conclusions about the nature of these disorders and their diagnosis and treatment. Using multimodality features to train machine-learning models holds promise for enhancing mental health evaluations and treatment. This approach aligns with the principles of preventive and personalized diagnosis and treatment and could lead to better outcomes for individuals with mental health conditions.

DATA AVAILABILITY

The data supporting this review paper are openly available and can be accessed at the following URLs: <https://bit.ly/3DBQtZk>; <https://bit.ly/43Q6Yv>; <https://bit.ly/44IKaPv>.

CODE AVAILABILITY

The code associated with this review paper is openly available and can be accessed at the following GitHub repository: <https://github.com/uofabinarylab/ADReview>.

Received: 10 March 2023; Accepted: 27 September 2023;
Published online: 20 November 2023

REFERENCES

- Friedrich, M. J. Depression is the leading cause of disability around the world. *J. Am. Med. Assoc.* **317**, 1517–1517 (2017).
- Evans-Lacko, S. et al. Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: results from the WHO world mental health (WMH) surveys. *Psychol. Med.* **48**, 1560–1571 (2018).
- Cai, H., Sha, X., Han, X., Wei, S. & Hu, B. Pervasive eeg diagnosis of depression using deep belief network with three-electrodes EEG collector. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1239–1246 (IEEE, 2016).
- Hosseiniard, B., Moradi, M. H. & Rostami, R. Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal. *Comput. Methods Programs Biomed.* **109**, 339–345 (2013).
- Sano, A. & Picard, R. W. Stress recognition using wearable sensors and mobile phones. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 671–676 (IEEE, 2013).
- Acharya, U. R. et al. Computer-aided diagnosis of depression using EEG signals. *Eur. Neurol.* **73**, 329–336 (2015).
- Hartmann, R., Schmidt, F. M., Sander, C. & Hegerl, U. Heart rate variability as indicator of clinical state in depression. *Front. Psychiatry* **9**, 735 (2019).
- Kan, D. P. X. & Lee, P. F. Decrease alpha waves in depression: an electroencephalogram (EEG) study. In *2015 International Conference on BioSignal Analysis, Processing and Systems (ICBAPS)*, 156–161 (IEEE, 2015).
- Ay, B. et al. Automated depression detection using deep representation and sequence learning with EEG signals. *J. Medi. Syst.* **43**, 1–12 (2019).
- Acharya, U. R. et al. Automated EEG-based screening of depression using deep convolutional neural network. *Comput. Methods Programs Biomed.* **161**, 103–113 (2018).
- Acharya, U. R. et al. A novel depression diagnosis index using nonlinear features in EEG signals. *Eur. Neurol.* **74**, 79–83 (2015).
- Mohan, Y., Chee, S. S., Xin, D. K. P. & Foong, L. P. Artificial neural network for classification of depressive and normal in EEG. In *2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, 286–290 (IEEE, 2016).
- Mumtaz, W. et al. Electroencephalogram (EEG)-based computer-aided technique to diagnose major depressive disorder (MDD). *Biomed. Signal Process. Control* **31**, 108–115 (2017).
- Akbari, H. et al. Depression recognition based on the reconstruction of phase space of EEG signals and geometrical features. *Appl. Acoust.* **179**, 108078 (2021).
- Kim, A. Y. et al. Skin conductance responses in major depressive disorder (MDD) under mental arithmetic stress. *PLoS ONE* **14**, e0213140 (2019).
- Williamson, J. R. et al. Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 11–18 (Association for Computing Machinery (ACM), 2016).
- Jan, A., Meng, H., Gaus, Y. F. B. A. & Zhang, F. Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. *IEEE Trans. Cogn. Dev. Syst.* **10**, 668–680 (2017).
- Derous, A. R., Zakzanis, K. K. & Ruocco, A. Facial emotion recognition in borderline personality disorder. *Psychol. Med.* **43**, 1953–1963 (2013).
- Zhao, Q. et al. Early perceptual anomaly of negative facial expression in depression: an event-related potential study. *Neurophysiologie Clinique/Clin. Neurophysiol.* **45**, 435–443 (2015).
- Seneviratne, N., Williamson, J. R., Lammert, A. C., Quatieri, T. F. & Espy-Wilson, C. Y. Extended study on the use of vocal tract variables to quantify neuromotor coordination in depression. *INTERSPEECH* 4551–4555 (2020).
- Zhao, Z. et al. Hybrid network feature extraction for depression assessment from speech. *Interspeech* <https://api.semanticscholar.org/CorpusID:226203252> (2020).
- Kiss, G. & Vicsi, K. Mono-and multi-lingual depression prediction based on speech processing. *Int. J. Speech Technol.* **20**, 919–935 (2017).
- Dham, S., Sharma, A. & Dhall, A. Depression scale recognition from audio, visual and text analysis. Preprint at <https://arxiv.org/abs/1709.05865> (2017).
- Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S. & Othmani, A. Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomed. Signal Process. Control* **71**, 103107 (2022).
- Jiang, H. et al. Detecting depression using an ensemble logistic regression model based on multiple speech features. *Comput. Math. Methods Med.* **2018**, (2018).
- Sardari, S., Nakisa, B., Rastgoor, M. N. & Eklund, P. Audio based depression detection using convolutional autoencoder. *Expert Syst. Appl.* **189**, 116076 (2022).
- Pampouchidou, A. et al. Facial geometry and speech analysis for depression detection. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 1433–1436 (IEEE, 2017).
- Maxhuni, A. et al. Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. *Pervasive Mobile Comput.g* **31**, 50–66 (2016).
- Amos, B., Ludwiczuk, B. & Satyanarayanan, M. *Openface: A General-purpose Face Recognition Library With Mobile Applications*. Tech. Rep., CMU-CS-16-118 (CMU School of Computer Science, 2016).
- Rahaman, M. A. et al. Multi-modal deep learning of functional and structural neuroimaging and genomic data to predict mental illness. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 3267–3272 (IEEE, 2021).
- Chiong, R., Budhi, G. S., Dhakal, S. & Chiong, F. A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. *Comput. Biol. Med.* **135**, 104499 (2021).
- Paul, S., Jandhyala, S. K. & Basu, T. Early detection of signs of anorexia and depression over social media using effective machine learning frameworks. In *Working Notes of {CLEF} 2018 - Conference and Labs of the Evaluation Forum, Avignon, France*, Vol. 2125 (eds. Cappellato, L., Ferro, N., Nie, J.-Y. & Soulier, L.) (CEUR-WS.org, 2018). <https://dblp.org/rec/conf/clef/PaulJB18.bib>.
- Mohr, D. C. et al. Perceived barriers to psychological treatments and their relationship to depression. *J. Clin. Psychol.* **66**, 394–409 (2010).
- Docherty, J. P. Barriers to the diagnosis of depression in primary care. *J. Clin. Psychiatry* **58**, 5–10 (1997).
- Byatt, N., Simas, T. A. M., Lundquist, R. S., Johnson, J. V. & Ziedonis, D. M. Strategies for improving perinatal depression treatment in North American outpatient obstetric settings. *J. Psychosom. Obstet. Gynecol.* **33**, 143–161 (2012).
- Chekroud, A. M. et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry* **3**, 243–250 (2016).
- Mira, A. et al. An internet-based program for depressive symptoms using human and automated support: a randomized controlled trial. *Neuropsych. Dis. Treat.* **13**, 987 (2017).
- Freedman, R. et al. The initial field trials of DSM-5: new blooms and old thorns. *Am. J. Psychiatry* **170**, 1–5 (2013).
- Regier, D. A. et al. Dsm-5 field trials in the United States and Canada, part ii: test-retest reliability of selected categorical diagnoses. *Am. J. Psychiatry* **170**, 59–70 (2013).
- Yehuda, R. Post-traumatic stress disorder. *New Engl. J. Med.* **346**, 108–114 (2002).
- Cummins, N. et al. A review of depression and suicide risk assessment using speech analysis. *Speech Commun.* **71**, 10–49 (2015).
- Insel, T. R. The NIMH research domain criteria (RDOC) project: precision medicine for psychiatry. *Am. J. Psychiatry* **171**, 395–397 (2014).
- Bzdok, D. & Meyer-Lindenberg, A. Machine learning for precision psychiatry: opportunities and challenges. *Biol. Psychiatry Cogn. Neurosci. Neuroimag.* **3**, 223–230 (2018).
- Vieira, S., Pinaya, W. H. & Mechelli, A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. *Neurosci. Biobehav. Rev.* **74**, 58–75 (2017).
- Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A. & Meneguzzi, F. Identification of autism spectrum disorder using deep learning and the abide dataset. *NeuroImage Clin.* **17**, 16–23 (2018).

46. Nahum-Shani, I. et al. Just-in-time adaptive interventions (jitaits) in mobile health: key components and design principles for ongoing health behavior support. *Annal. Behav. Med.* **52**, 446–462 (2018).
47. Andrade, L. et al. The epidemiology of major depressive episodes: results from the international consortium of psychiatric epidemiology (ICPE) surveys. *Int. J. Methods Psychiatric Res.* **12**, 3–21 (2003).
48. Girgus, J. S., Yang, K. & Ferri, C. V. The gender difference in depression: are elderly women at greater risk for depression than elderly men? *Geriatrics* **2**, 35 (2017).
49. Schuch, J. J., Roest, A. M., Nolen, W. A., Penninx, B. W. & De Jonge, P. Gender differences in major depressive disorder: results from the Netherlands study of depression and anxiety. *J. Affect. Disord.* **156**, 156–163 (2014).
50. Gao, W., Ping, S. & Liu, X. Gender differences in depression, anxiety, and stress among college students: a longitudinal study from China. *J. Affect. Disord.* **263**, 292–300 (2020).
51. Albert, P. R. Why is depression more prevalent in women? *J. Psychiatry Neurosci. JPN* **40**, 219 (2015).
52. Kraepelin, E. *Manic-Depressive Insanity and Paranoia* (E. & S. Livingstone, 1921).
53. Kumar, M., Dredze, M., Coppersmith, G. & De Choudhury, M. Detecting changes in suicide content manifested in social media following celebrity suicides. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, 85–94 (ACM, 2015).
54. Pirina, I. & Çöltekin, Ç. Identifying depression on Reddit: the effect of training data. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, 9–12 (Association for Computational Linguistics (ACL), 2018).
55. Yates, A., Cohan, A. & Goharian, N. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2968–2978 (Association for Computational Linguistics, Copenhagen, Denmark, 2017). <https://doi.org/10.18653/v1/D17-1322>.
56. Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H. & Eichstaedt, J. C. Detecting depression and mental illness on social media: an integrative review. *Curr. Opin. Behav. Sci.* **18**, 43–49 (2017).
57. Aragón, M. E., López-Monroy, A. P., González-Gurrola, L. C. & Montes, M. Detecting depression in social media using fine-grained emotions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* 1481–1486 (Association for Computational Linguistics (ACL), 2019).
58. Tadesse, M. M., Lin, H., Xu, B. & Yang, L. Detection of depression-related posts in Reddit social media forum. *IEEE Access* **7**, 44883–44893 (2019).
59. De Choudhury, M. & De, S. Mental health discourse on Reddit: self-disclosure, social support, and anonymity. In *Eighth International AAAI Conference on Weblogs and Social Media* (MIT Press, 2014).
60. Alghamdi, N. S., Mahmoud, H. A. H., Abraham, A., Alanazi, S. A. & García-Hernández, L. Predicting depression symptoms in an Arabic psychological forum. *IEEE Access* **8**, 57317–57334 (2020).
61. Jan, A., Meng, H., Gaus, Y. F. A., Zhang, F. & Turabzadeh, S. Automatic depression scale prediction using facial expression dynamics and regression. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 73–80 (Association for Computing Machinery (ACM), 2014).
62. Wang, Q., Yang, H. & Yu, Y. Facial expression video analysis for depression detection in Chinese patients. *J. Vis. Commun. Image Represent.* **57**, 228–233 (2018).
63. Hunter, L., Roland, L. & Ferozpuri, A. Emotional expression processing and depressive symptomatology: eye-tracking reveals differential importance of lower and middle facial areas of interest. *Depression Res. Treatment* **2020**, (2020).
64. Alghowinem, S., Goecke, R., Wagner, M., Parkerx, G. & Breakspear, M. Head pose and movement analysis as an indicator of depression. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 283–288 (IEEE, 2013).
65. Low, D. M., Bentley, K. H. & Ghosh, S. S. Automated assessment of psychiatric disorders using speech: a systematic review. *Laryngoscope Investig. Otolaryngol.* **5**, 96–116 (2020).
66. Cummins, N., Sethu, V., Epps, J., Schnieder, S. & Krajewski, J. Analysis of acoustic space variability in speech affected by depression. *Speech Commun.* **75**, 27–49 (2015).
67. Cummins, N., Sethu, V., Epps, J. & Krajewski, J. Probabilistic acoustic volume analysis for speech affected by depression. In *Fifteenth Annual Conference of the International Speech Communication Association*. (International Speech Communication Association, 2014).
68. Harati, S., Crowell, A., Mayberg, H. & Nemati, S. Depression severity classification from speech emotion. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 5763–5766 (IEEE, 2018).
69. Cummins, N., Vlasenko, B., Sagha, H. & Schuller, B. Enhancing speech-based depression detection through gender dependent vowel-level formant features. In *Conference on Artificial Intelligence in Medicine in Europe*, 209–214 (Springer, 2017).
70. Morales, M. R. & Levitan, R. Speech vs. text: a comparative analysis of features for depression detection systems. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, 136–143 (IEEE, 2016).
71. Vicsi, K., Szláhó, D. & Kiss, G. Examination of the sensitivity of acoustic-phonetic parameters of speech to depression. In *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*, 511–515 (IEEE, 2012).
72. Kiss, G. & Vicsi, K. Comparison of read and spontaneous speech in case of automatic detection of depression. In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 000213–000218 (IEEE, 2017).
73. Stasak, B., Epps, J. & Goecke, R. Elicitation design for acoustic depression classification: an investigation of articulation effort, linguistic complexity, and word affect. *INTERSPEECH*, 834–838 (2017).
74. Yazdavar, A. H. et al. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 1191–1198 (IEEE/ACM, 2017).
75. Zogan, H., Razzak, I., Jameel, S. & Xu, G. Depressionnet: a novel summarization boosted deep framework for depression detection on social media. Preprint at <https://arxiv.org/abs/2105.10878> (2021).
76. Yang, K., Zhang, T. & Ananiadou, S. A mental state knowledge-aware and contrastive network for early stress and depression detection on social media. *Inf. Process. Manag.* **59**, 102961 (2022).
77. Lara, J. S., Aragón, M. E., González, F. A. & Montes-y Gómez, M. Deep bag-of-sub-emotions for depression detection in social media. In *Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings* 24, 60–72 (Springer International Publishing, 2021).
78. Figueiredo, J. S. L., Maia, A. L. L. & Calumby, R. T. Early depression detection in social media based on deep learning and underlying emotions. *Online Soc. Netw. Media* **31**, 100225 (2022).
79. Stankevich, M., Isakov, V., Devyatkin, D. & Smirnov, I. V. Feature engineering for depression detection in social media. *ICPRAM*, 426–431 (2018).
80. de Jesús Titla-Tlatelpa, J., Ortega-Mendoza, R. M., Montes-y Gómez, M. & Villa-senor-Pineda, L. A profile-based sentiment-aware approach for depression detection in social media. *EPJ Data Sci.* **10**, 54 (2021).
81. Li, Z. et al. Mha: a multimodal hierarchical attention model for depression detection in social media. *Health Inf. Sci. Syst.* **11**, 6 (2023).
82. Cha, J., Kim, S. & Park, E. A lexicon-based approach to examine depression detection in social media: the case of Twitter and university community. *Humanit. Soc. Sci. Commun.* **9**, 1–10 (2022).
83. Cui, B. et al. Emotion-based reinforcement attention network for depression detection on social media: algorithm development and validation. *JMIR Med. Informatics* **10**, e37818 (2022).
84. Guo, Z., Ding, N., Zhai, M., Zhang, Z. & Li, Z. Leveraging domain knowledge to improve depression detection on Chinese social media. In *IEEE Transactions on Computational Social Systems* (IEEE, 2023).
85. Hosseini-Saravani, S. H., Besharati, S., Calvo, H. & Gelbukh, A. Depression detection in social media using a psychoanalytical technique for feature extraction and a cognitive based classifier. In *Advances in Computational Intelligence: 19th Mexican International Conference on Artificial Intelligence, MICAI 2020, Mexico City, Mexico, October 12–17, 2020, Proceedings, Part II*, 282–292 (Springer, 2020).
86. Ramandrisoa, F. & Mothe, J. Early detection of depression and anorexia from social media: a machine learning approach. *Circle* **2621**, 2020 (2020).
87. Zogan, H., Razzak, I., Wang, X., Jameel, S. & Xu, G. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web* **25**, 281–304 (2022).
88. Cunningham, S., Hudson, C. C. & Harkness, K. Social media and depression symptoms: a meta-analysis. *Res. Child Adolesc. Psychopathol.* **49**, 241–253 (2021).
89. Shensa, A. et al. Problematic social media use and depressive symptoms among us young adults: a nationally-representative study. *Soc. Sci. Med.* **182**, 150–157 (2017).
90. Woods, H. C. & Scott, H. # sleepyteens: social media use in adolescence is associated with poor sleep quality, anxiety, depression and low self-esteem. *J. Adolesc.* **51**, 41–49 (2016).
91. Radovic, A., Gmelin, T., Stein, B. D. & Miller, E. Depressed adolescents' positive and negative use of social media. *J. Adolesc.* **55**, 5–15 (2017).
92. Ivie, E. J., Pettitt, A., Moses, L. J. & Allen, N. B. A meta-analysis of the association between adolescent social media use and depressive symptoms. *J. Affect. Disord.* **275**, 165–174 (2020).
93. Raudsepp, L. & Kais, K. Longitudinal associations between problematic social media use and depressive symptoms in adolescent girls. *Prev. Med. Rep.* **15**, 100925 (2019).

94. Salas-Zárate, R. et al. Detecting depression signs on social media: a systematic literature review. In *Healthcare*, Vol. 10, 291 (MDPI, 2022).
95. Liu, D. et al. Detecting and measuring depression on social media using a machine learning approach: systematic review. *JMIR Mental Health* **9**, e27244 (2022).
96. McCrae, N., Gettings, S. & Purssell, E. Social media and depressive symptoms in childhood and adolescence: a systematic review. *Adolesc. Res. Rev.* **2**, 315–330 (2017).
97. Heffer, T., Good, M., Daly, O., MacDonell, E. & Willoughby, T. The longitudinal association between social-media use and depressive symptoms among adolescents and young adults: an empirical reply to Twenge et al. (2018). *Clin. Psychol. Sci.* **7**, 462–470 (2019).
98. Ford, E., Curlewis, K., Wongkoblap, A. & Curnin, V. et al. Public opinions on using social media content to identify users with depression and target mental health care advertising: mixed methods survey. *JMIR Mental Health* **6**, e12942 (2019).
99. Hou, F., Bi, F., Jiao, R., Luo, D. & Song, K. Gender differences of depression and anxiety among social media users during the covid-19 outbreak in China: a cross-sectional study. *BMC Public Health* **20**, 1–11 (2020).
100. De Choudhury, M., Counts, S. & Horvitz, E. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, 47–56 (ACM, 2013).
101. Narynov, S., Mukhtarkhanuly, D. & Omarov, B. Dataset of depressive posts in Russian language collected from social media. *Data Brief* **29**, 105195 (2020).
102. Li, X., Guo, W. & Yang, H. Depression severity prediction from facial expression based on the drr_depressionnet network. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2757–2764 (IEEE, 2020).
103. Hao, Y., Cao, Y., Li, B. & Rahman, M. Depression recognition based on text and facial expression. In *International Symposium on Artificial Intelligence and Robotics 2021*, vol. 11884, 513–522 (SPIE, 2021).
104. Liu, Z. et al. Pra-net: Part-and-relation attention network for depression recognition from facial expression. *Comput. Biol. Med.* **157**, 106589 (2023).
105. Hamid, D. S. B. A., Goyal, S. & Bedi, P. Integration of deep learning for improved diagnosis of depression using eeg and facial features. *Mater. Today Proc.* **80**, 1965–1969 (2023).
106. Nasir, M., Jati, A., Shivakumar, P. G., Nallan Chakravarthula, S. & Georgiou, P. Multimodal and multiresolution depression detection from speech and facial landmark features. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 43–50 (Association for Computing Machinery (ACM), 2016).
107. Dai, Z., Li, Q., Shang, Y. & Wang, X. Depression detection based on facial expression, audio and gait. In *2023 IEEE 6th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, vol. 6, 1568–1573 (IEEE, 2023).
108. Shangguan, Z. et al. Dual-stream multiple instance learning for depression detection with facial expression videos. In *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (IEEE, 2022).
109. Rodrigues Makiuchi, M., Warnita, T., Uto, K. & Shinoda, K. Multimodal fusion of bert-cnn and gated cnn representations for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 55–63 (Association for Computing Machinery (ACM), 2019).
110. Yin, S., Liang, C., Ding, H. & Wang, S. A multi-modal hierarchical recurrent neural network for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 65–71 (Association for Computing Machinery (ACM), 2019).
111. Zhang, L., Driscoll, J., Chen, X. & Hosseini Ghomi, R. Evaluating acoustic and linguistic features of detecting depression sub-challenge dataset. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 47–53 (Association for Computing Machinery (ACM), 2019).
112. Yang, Y., Fairbairn, C. & Cohn, J. F. Detecting depression severity from vocal prosody. *IEEE Trans. Affect. Comput.* **4**, 142–150 (2012).
113. McGinnis, E. W. et al. Giving voice to vulnerable children: machine learning analysis of speech detects anxiety and depression in early childhood. *IEEE J. Biomed. Health Inform.* **23**, 2294–2301 (2019).
114. Sanchez, M. H. et al. Using prosodic and spectral features in detecting depression in elderly males. In *Twelfth Annual Conference of the International Speech Communication Association* (International Speech Communication Association, 2011).
115. Silva, W. J., Lopes, L., Galdino, M. K. C. & Almeida, A. A. Voice acoustic parameters as predictors of depression. *J. Voice* (2021).
116. Smith, M., Dietrich, B. J., Bai, E.-w. & Bockholt, H. J. Vocal pattern detection of depression among older adults. *Int. J. Mental Health Nurs.* **29**, 440–449 (2020).
117. Asgari, M., Shafran, I. & Sheeber, L. B. Inferring clinical depression from speech and spoken utterances. In *2014 IEEE international workshop on Machine Learning for Signal Processing (MLSP)*, 1–5 (IEEE, 2014).
118. Scherer, S., Stratou, G., Gratch, J. & Morency, L.-P. Investigating voice quality as a speaker-independent indicator of depression and PTSD. *Interspeech* 847–851 (2013).
119. Pan, W. et al. Re-examining the robustness of voice features in predicting depression: compared with baseline of confounders. *PLoS ONE* **14**, e0218172 (2019).
120. Gratch, J. et al. *The Distress Analysis Interview Corpus of Human and Computer Interviews*. Tech. Rep. (UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES, 2014).
121. Valstar, M. et al. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 3–10 (Association for Computing Machinery (ACM), 2014).
122. Valstar, M. et al. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, 3–10 (ACM, 2013).
123. De Hert, M., Detraux, J., Van Winkel, R., Yu, W. & Correll, C. U. Metabolic and cardiovascular adverse effects associated with antipsychotic drugs. *Nat. Rev. Endocrinol.* **8**, 114–126 (2012).
124. Kane, J., Aylett, M., Yanushevskaya, I. & Gobl, C. Phonetic feature extraction for context-sensitive glottal source processing. *Speech Commun.* **59**, 10–21 (2014).
125. Alghowinem, S. et al. Detecting depression: a comparison between spontaneous and read speech. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 7547–7551 (IEEE, 2013).
126. DeVault, D. et al. Simsensei kiosk: a virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, 1061–1068 (Association for Computing Machinery (ACM), 2014).
127. Hartholt, A. et al. All together now. In *International Workshop on Intelligent Virtual Agents*, 368–381 (Springer, 2013).
128. Burton, C. et al. Pilot randomised controlled trial of help4mood, an embodied virtual agent-based system to support treatment of depression. *J. Telemed. Telecare* **22**, 348–355 (2016).
129. Nemes, V., Nikolic, D., Barney, A. & Garrard, P. A feasibility study of speech recording using a contact microphone in patients with possible or probable Alzheimer's disease to detect and quantify repetitions in a natural setting. *Alzheimer's Dementia* **8**, P490–P491 (2012).
130. Alosiban, N., Esposito, A. & Vinciarelli, A. What you say or how you say it? depression detection through joint modeling of linguistic and acoustic aspects of speech. *Cogn. Comput.* **14**, 1585–1598 (2022).
131. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **10**, e0118432 (2015).
132. Lawrence, I. & Lin, K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**, 255–268 (1989).
133. Wang, F., Kaushal, R. & Khullar, D. Should health care demand interpretable artificial intelligence or accept "black box" medicine? *Ann. Intern. Med.* **172**, 59–60 (2020).
134. Quinn, T. P., Jacobs, S., Senadeera, M., Le, V. & Coghlann, S. The three ghosts of medical AI: can the black-box present deliver? *Artif. Intell. Med.* **124**, 102158 (2022).
135. Sendak, M. et al. "the human body is a black box" supporting clinical decision-making with deep learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 99–109 (Association for Computing Machinery (ACM), 2020).
136. Lipton, Z. C. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. *Queue* **16**, 31–57 (2018).
137. Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. Preprint at <https://arxiv.org/abs/1702.08608> (2017).
138. Molnar, C. *Interpretable Machine Learning* (Lulu. com, 2020).
139. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4768–4777 (2017).
140. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
141. Beam, A. L., Manrai, A. K. & Ghassemi, M. Challenges to the reproducibility of machine learning models in health care. *J. Am. Med. Assoc.* **323**, 305–306 (2020).
142. McDermott, M. B. et al. Reproducibility in machine learning for health research: still a ways to go. *Sci. Transl. Med.* **13**, eabb1655 (2021).
143. Blum, A. & Hardt, M. The ladder: a reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning*, 1006–1014 (PMLR, 2015).
144. Alghowinem, S., Goecke, R., Epps, J., Wagner, M. & Cohn, J. F. Cross-cultural depression recognition from vocal biomarkers. *Interspeech*, 1943–1947 (2016).

145. Stasak, B. & Epps, J. Differential performance of automatic speech-based depression classification across smartphones. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 171–175 (IEEE, 2017).
146. Gideon, J., Provost, E. M. & McInnis, M. Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2359–2363 (IEEE, 2016).
147. Mitra, V. & Shriberg, E. Effects of feature type, learning algorithm and speaking style for depression detection from speech. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4774–4778 (IEEE, 2015).
148. Custers, B. Click here to consent forever: Expiry dates for informed consent. *Big Data Soc.* **3**, 2053951715624935 (2016).
149. Rahman, A., Malik, A., Sikander, S., Roberts, C. & Creed, F. Cognitive behaviour therapy-based intervention by community health workers for mothers with depression and their infants in rural Pakistan: a cluster-randomised controlled trial. *Lancet* **372**, 902–909 (2008).
150. Marmor, S., Horvath, K. J., Lim, K. O. & Misono, S. Voice problems and depression among adults in the United States. *Laryngoscope* **126**, 1859–1864 (2016).
151. Hartanto, A., Quek, F. Y., Tng, G. Y. & Yong, J. C. Does social media use increase depressive symptoms? a reverse causation perspective. *Front. Psychiatry* **12**, 641934 (2021).
152. Hussain, J. et al. Exploring the dominant features of social media for depression detection. *J. Inf. Sci.* **46**, 739–759 (2020).
153. Liaw, A. S. & Chua, H. N. Depression detection on social media with user network and engagement features using machine learning methods. In *2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, 1–6 (IEEE, 2022).
154. Primack, B. A. et al. Use of multiple social media platforms and symptoms of depression and anxiety: a nationally-representative study among us young adults. *Comput. Hum. Behav.* **69**, 1–9 (2017).
155. Primack, B. A., Shensa, A., Sidani, J. E., Escobar-Viera, C. G. & Fine, M. J. Temporal associations between social media use and depression. *Am. J. Prev. Med.* **60**, 179–188 (2021).
156. Vedula, N. & Parthasarathy, S. Emotional and linguistic cues of depression from social media. In *Proceedings of the 2017 International Conference on Digital Health*, 127–136 (Association for Computing Machinery (ACM), 2017).
157. Nesi, J. et al. Emotional responses to social media experiences among adolescents: longitudinal associations with depressive symptoms. *J. Clin. Child Adolesc. Psychol.* **51**, 907–922 (2021).
158. Thorisdottir, I. E., Sigurvinssdottir, R., Asgeirsdottir, B. B., Allegrante, J. P. & Sigurdottir, I. D. Active and passive social media use and symptoms of anxiety and depressed mood among Icelandic adolescents. *Cyberpsychol. Behav. Soc. Netw.* **22**, 535–542 (2019).
159. Ghosh, S. & Anwar, T. Depression intensity estimation via social media: a deep learning approach. *IEEE Trans. Comput. Soc. Syst.* **8**, 1465–1474 (2021).
160. Aragon, M. E., Lopez-Monroy, A. P., Gonzalez-Gurrula, L.-C. G. & Montes, M. Detecting mental disorders in social media through emotional patterns-the case of anorexia and depression. In *IEEE Transactions on Affective Computing* (IEEE, 2021).
161. Puukko, K., Hietajärvi, L., Maksniemi, E., Alho, K. & Salmela-Aro, K. Social media use and depressive symptoms-a longitudinal study from early to late adolescence. *Int. J. Environ. Res. Public Health* **17**, 5921 (2020).
162. Robinson, A. et al. Social comparisons, social media addiction, and social interaction: an examination of specific social media behaviors related to major depressive disorder in a millennial population. *J. Appl. Biobehav. Res.* **24**, e12158 (2019).
163. De Choudhury, M., Gamon, M., Counts, S. & Horvitz, E. Predicting depression via social media. In *Seventh International AAAI Conference on Weblogs and Social Media (AAAI, 2013)*.
164. Zhong, B., Huang, Y. & Liu, Q. Mental health toll from the coronavirus: social media usage reveals Wuhan residents' depression and secondary trauma in the covid-19 outbreak. *Comput. Hum. Behav.* **114**, 106524 (2021).
165. Haand, R. & Shuwang, Z. The relationship between social media addiction and depression: a quantitative study among university students in khost, afghanistan. *Int. J. Adolesc. Youth* **25**, 780–786 (2020).
166. Brailowskaia, J. & Margraf, J. Relationship between depression symptoms, physical activity, and addictive social media use. *Cyberpsychol. Behav. Soc. Netw.* **23**, 818–822 (2020).
167. Jeri-Yabar, A. et al. Association between social media use (Twitter, Instagram, Facebook) and depressive symptoms: are Twitter users at higher risk? *Int. J. Soc. Psychiatry* **65**, 14–19 (2019).
168. Kircaburun, K. Self-esteem, daily internet use and social media addiction as predictors of depression among Turkish adolescents. *J. Educ. Practice* **7**, 64–72 (2016).
169. Ricard, B. J., Marsch, L. A., Crosier, B. & Hassanpour, S. Exploring the utility of community-generated social media content for detecting depression: an analytical study on Instagram. *J. Med. Internet Res.* **20**, e11817 (2018).
170. Peng, Z., Hu, Q. & Dang, J. Multi-kernel svm based depression recognition using social media data. *Int. J. Mach. Learn. Cybernet.* **10**, 43–57 (2019).
171. Aldarwish, M. M. & Ahmad, H. F. Predicting depression levels using social media posts. In *2017 IEEE 13th international Symposium on Autonomous Decentralized System (ISADS)*, 277–280 (IEEE, 2017).
172. Burdiasso, S. G., Errecalde, M. & Montes-y Gómez, M. A text classification framework for simple and effective early depression detection over social media streams. *Expert Syst. Appl.* **133**, 182–197 (2019).
173. Smyns, S. & Raj, J. S. Analysis of deep learning techniques for early detection of depression on social media network-a comparative study. *J. Trends Comput. Sci. Smart Technol. (TCSST)* **3**, 24–39 (2021).
174. Bucur, A.-M. & Dinu, L. P. Detecting early onset of depression from social media text using learned confidence scores. Preprint at <https://arxiv.org/abs/2011.01695> (2020).
175. Sampath, K. & Durairaj, T. Data set creation and empirical analysis for detecting signs of depression from social media postings. In *International Conference on Computational Intelligence in Data Science*, 136–151 (Springer, 2022).
176. Mann, P., Paes, A. & Matsushima, E. H. See and read: detecting depression symptoms in higher education students using multimodal social media data. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 440–451 (AAAI, 2020).
177. Sadeque, F., Xu, D. & Bethard, S. Measuring the latency of depression detection in social media. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 495–503 (ACM, 2018).
178. Fatima, I. et al. Prediction of postpartum depression using machine learning techniques from social media text. *Expert Syst.* **36**, e12409 (2019).
179. Katchapakirin, K., Wongpatikaseree, K., Yomaboot, P. & Kaewpitakkun, Y. Facebook social media for depression detection in the Thai community. In *2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 1–6 (IEEE, 2018).
180. Shen, G. et al. Depression detection via harvesting social media: a multimodal dictionary learning solution. *IJCAI*, 3838–3844 (2017).
181. Li, D., Chaudhary, H. & Zhang, Z. Modeling spatiotemporal pattern of depressive symptoms caused by covid-19 using social media data mining. *Int. J. Environ. Res. Public Health* **17**, 4988 (2020).
182. Lin, L. Y. et al. Association between social media use and depression among us young adults. *Depress. Anxiety* **33**, 323–331 (2016).
183. Mohan, M., Abhinav, A., Ashok, A., Akhil, A. & Achinth, P. Depression detection using facial expression and sentiment analysis. In *2021 Asian Conference on Innovation in Technology (ASIANCON)*, 1–6 (IEEE, 2021).
184. Lee, Y.-S. & Park, W.-H. Diagnosis of depressive disorder model on facial expression based on fast r-cnn. *Diagnostics* **12**, 317 (2022).
185. Sumali, B., Mitsukura, Y., Tazawa, Y. & Kishimoto, T. Facial landmark activity features for depression screening. In *2019 58th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, 1376–1381 (IEEE, 2019).
186. Dadiz, B. G. & Ruiz, C. R. Detecting depression in videos using uniformed local binary pattern on facial features. In *Computational Science and Technology: 5th ICCST 2018, Kota Kinabalu, Malaysia, 29-30 August 2018*, 413–422 (Springer, 2019).
187. Stasak, B., Huang, Z., Joachim, D. & Epps, J. Automatic elicitation compliance for short-duration speech based depression detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7283–7287 (IEEE, 2021).
188. Huang, Z., Epps, J. & Joachim, D. Speech landmark bigrams for depression detection from naturalistic smartphone speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5856–5860 (IEEE, 2019).
189. Huang, Z., Epps, J., Joachim, D. & Chen, M. Depression detection from short utterances via diverse smartphones in natural environmental conditions. *INTERSPEECH* 3393–3397 (2018).
190. Szabadi, E., Bradshaw, C. & Besson, J. Elongation of pause-time in speech: a simple, objective measure of motor retardation in depression. *Br. J. Psychiatry* **129**, 592–597 (1976).
191. He, L., Jiang, D. & Sahli, H. Multimodal depression recognition with dynamic visual and audio cues. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 260–266 (IEEE, 2015).
192. Pérez Espinosa, H. et al. Fusing affective dimensions and audio-visual features from segmented video for depression recognition: Inaoe-buap's participation at avec'14 challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 49–55 (Association for Computing Machinery (ACM), 2014).

193. Malandrakis, N., Potamianos, A., Evangelopoulos, G. & Zlatintsi, A. A supervised approach to movie emotion tracking. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2376–2379 (IEEE, 2011).
194. Semkovska, M., Noone, M., Carton, M. & McLoughlin, D. M. Measuring consistency of autobiographical memory recall in depression. *Psychiatry Res.* **197**, 41–48 (2012).
195. Saeb, S., Lattie, E. G., Kording, K. P. & Mohr, D. C. et al. Mobile phone detection of semantic location and its relationship to depression and anxiety. *JMIR mHealth uHealth* **5**, e7297 (2017).

ACKNOWLEDGEMENTS

We would like to thank the funding support of MITACS for this research. This research is also supported by the China Scholarship Council (CSC) No. 202000810031 and No. 202308180002.

AUTHOR CONTRIBUTIONS

K.M. was a major contributor to writing the manuscript. K.M. and Y.W. performed the literature review. J.C. supervised the project and provided funds to support the project. All authors have given approval for the final version of the manuscript. All authors reviewed the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s44184-023-00040-z>.

Correspondence and requests for materials should be addressed to Jie Chen.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023