Arch Psychiatry. 2026;4(1):47–62.

The Archives of Psychiatry | Systematic Review

Echoes of the Mind: A Systematic Review of Deep Transfer Learning for the Detection and Prediction of Depression and Anxiety

Tariq Saeed Mian[1,*]

[1]Department of Information Systems, College of Computer Science and Engineering, Taibah University, Madinah Almunawwarah, Kingdom of Saudi Arabia

[*]Correspondence should be addressed to Tariq Saeed Mian, tmian@taibahu.edu.sa

Received date: August 10, 2026, **Accepted date:** October 30, 2026

Citation: Mian TS. Echoes of the Mind: A Systematic Review of Deep Transfer Learning for the Detection and Prediction of Depression and Anxiety. Arch Psychiatry. 2026;4(1):47–62.

Copyright: © 2026 Mian TS. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Background: Depression and anxiety are common and often detected late. Deep learning needs large labelled datasets, which are rare in mental health. Transfer learning reuses knowledge from pre-trained models and may reduce this problem.

Objective: To map how deep transfer learning has been used to detect, classify or estimate the severity of depression and anxiety, and to examine how the reported results depend on validation design.

Methods: The review followed PRISMA 2020. PubMed and OpenAlex were searched on 27 October 2026 for studies published from January 2019 to September 2026. The earlier version of this manuscript was also checked for eligible studies. Primary studies were included if they applied a pre-trained deep model, fine-tuning, domain adaptation, self-supervised pre-training or a large language model to individual-level detection or severity estimation. Data were extracted from full abstracts and synthesized narratively.

Results: Of 798 records, 318 studies met the criteria. Most studies (81.1%) were published from 2024 onward. Depression alone was studied in 288 studies, anxiety alone in 10 and both in 20. Text from social media (34.9%) and clinical interviews or narratives (17.3%) were the main data sources. Fine-tuning (143 studies) and feature extraction (86) were most common, and 87 studies used large language or foundation models. Only 45 studies (14.2%) reported participant-independent or external validation, and 98 abstracts gave no numerical result. The median reported accuracy was 91.3%, and external validation gave lower values. Small local language models performed poorly on interview transcripts.

Conclusions: Transfer learning is now the default approach in this field, but the evidence is mostly based on random splits of small or public benchmark datasets. Reported accuracies should not be read as clinical performance. Future work needs participant-independent and external testing, TRIPOD+AI reporting, more work on anxiety and comorbidity, and data from under-represented languages such as Arabic.

Keywords: Depression, Anxiety, Transfer learning, Deep learning, Large language models, Pre-trained models, Speech, Social media, Mental health screening, Systematic review

Introduction

Background

Depression and anxiety are among the leading causes of disability worldwide. The Global Burden of Disease Study 2019 placed depressive and anxiety disorders among the top mental disorders for disability-adjusted life years [1]. The COVID-19 pandemic added an estimated 53.2 million cases of major depressive disorder and 76.2 million cases of anxiety disorders in 2020 [2]. The World Health Organization now estimates that about 322 million people live with depression, and that in high-income countries only about one third of them receive treatment [3]. In Saudi Arabia, the Saudi National Mental Health Survey found that about one in three adults had a lifetime mental disorder, and anxiety disorders were the most common group [4].

Diagnosis still depends on clinical interviews and self-report scales such as the PHQ-9 and GAD-7. These tools are valid, but they need clinical time and honest reporting. Many people do not seek help because of stigma. For this reason, researchers

have tried to detect depression and anxiety from data that people produce in daily life, such as written text, speech, facial behavior, physiological signals and health records [5,6].

Early studies used classical machine learning with hand-crafted features, and later studies trained convolutional and recurrent networks from the start [7,8]. Both approaches face the same limit. Labelled mental health data are small, expensive to collect and hard to share. The DAIC-WOZ interview corpus, which is used in many studies, contains fewer than 200 participants [9]. The MODMA dataset contains (Electroencephalogram) EEG and speech from about 50 participants [10]. Deep networks trained on such data tend to over fit.

Transfer learning and its variants

Transfer learning uses knowledge learned on a large source task to help a smaller target task [11,12]. In this review, five forms are separated, because they differ in cost and in risk.

Feature extraction keeps the pre-trained network frozen and uses its outputs as input to a new classifier. A typical example is a BERT sentence embedding passed to a support vector machine, or an ImageNet ResNet applied to EEG spectrogram images. Fine-tuning updates all or part of the pre-trained weights on the target data [13]. Parameter-efficient fine-tuning updates only a small set of new parameters, such as prefix vectors or Low-Rank Adapters (LoRA). Domain adaptation aligns the feature distributions of a labelled source domain and a target domain, for example across EEG subjects, MRI sites or languages. Self-supervised pre-training learns representations from unlabeled data, as in wav2vec 2.0 for speech [14]. Large Language Models (LLMs) add a further option. They can be used without any training through zero-shot or few-shot prompting, or they can be fine-tuned on mental health text. Domain-specific encoders such as Mental BERT were pre-trained on mental health forum posts for the same purpose [15].

Rationale

Several reviews have covered machine learning and deep learning for depression [5–8], social media data [16], and LLMs in mental health [17,18]. Recent reviews have looked at transformer language models [19], multimodal LLMs [20] and fusion methods [21]. These reviews treat transfer learning as one technique among many, or they focus on one modality. None has compared the transfer strategies across text, speech, video, EEG, imaging and health records, and none has related the reported performance to the validation design. This matters because many studies split data at the level of segments or posts. The same person can then appear in both training and test sets, and the accuracy is inflated [22]. The earlier version of this manuscript was a narrative review and did not report a search method. The present version is a systematic review conducted and reported according to PRISMA 2020 [23].

Objectives

The review addresses four questions:

RQ1. Which transfer learning strategies and pre-trained models have been used to detect, classify or estimate the severity of depression and anxiety, and for which data modalities?

RQ2. What performance do these studies report, and how does it vary with the validation design?

RQ3. How well do the studies report methods that matter for clinical use, such as participant-independent testing, external validation, fairness and privacy?

RQ4. Which gaps remain for future research?

Methods

Protocol and registration

The review was not registered, and no protocol was published. The eligibility criteria, information sources, and search strings were fixed before screening began. The review follows the PRISMA 2020 statement [23]. The completed PRISMA 2020 checklist is given in **Appendix B**.

Eligibility criteria

Table 1 lists the criteria. Primary studies were included if they (a) addressed individual-level detection, classification, risk prediction or severity estimation of depression or anxiety in humans; (b) used at least one deep model with a transfer learning component, as defined in Section 1.2, in the proposed method or in a reported comparison; (c) reported at least one quantitative result; and (d) were peer-reviewed journal articles or full conference papers in English, published between 1 January 2019 and 30 September 2026. Reviews, editorials, protocols, conference abstracts, dataset papers without a detection experiment, and preprints were excluded. Studies were also excluded if the outcome was treatment response, suicide risk alone, stress alone or bipolar disorder alone, or if an LLM was used only as a chatbot, for therapy or to answer patient questions. Studies that used only classical machine learning, networks trained from random initialization, or static word embedding's (for example, GloVe) were excluded.

Information sources

Two sources were searched on 27 October 2026. PubMed covers the clinical and psychiatric literature. Open Alex is an open index of scholarly works that covers journals and conference proceedings from IEEE, ACM, Springer, Elsevier and MDPI, among others. Scopus, Web of Science, and IEEE Xplore were not available to the author through an application

Table 1. Eligibility criteria.

Criterion	Included	Excluded
Population	Humans of any age, clinical or general population	Animal or cellular studies
Condition and task	Individual-level detection, classification, risk prediction, or severity estimation of depression and/or anxiety	Treatment response, suicide risk alone, stress alone, bipolar disorder alone; population-level trend analysis; LLM used only as chatbot, therapist or information source
Method	At least one deep model with a transfer learning component: pre-trained network (feature extraction or fine-tuning), parameter-efficient tuning, domain adaptation, self-supervised pre-training, or a pre-trained LLM/foundation model (prompted or fine-tuned)	Classical machine learning only; deep networks trained from random initialization only; static word embeddings only
Outcome measure	At least one quantitative result (e.g., accuracy, F1, AUC, MAE, RMSE, correlation)	No quantitative result
Report type	Peer-reviewed journal article or full conference paper; primary study	Reviews, editorials, letters, protocols, conference abstracts, dataset papers without an experiment, preprints
Language and period	English; published 1 January 2019 to 30 September 2026	Other languages; outside the period

programming interface at the time of the search. To reduce the effect of this gap, the reference list of the earlier version of this manuscript, which contained 111 references, was screened for eligible studies (citation searching).

Search strategy

The search combined three concept blocks: the condition (depression or anxiety in the title), the transfer learning or pre-trained model (in the title or abstract), and the task (detection, prediction, classification, diagnosis, screening, recognition or identification). The PubMed string was:

(depress*[ti] OR anxiety[ti] OR anxious[ti]) AND ("transfer learning"[tiab] OR "pre-trained"[tiab] OR pretrained[tiab] OR "fine-tun*"[tiab] OR finetun*[tiab] OR "domain adaptation"[tiab] OR BERT[tiab] OR RoBERTa[tiab] OR "large language model*"[tiab] OR "foundation model*"[tiab] OR GPT*[tiab]) AND (detect*[tiab] OR predict*[tiab] OR classif*[tiab] OR diagnos*[tiab] OR screen*[tiab] OR recogni*[tiab] OR identif*[tiab]) AND ("2019/01/01"[dp] : "2026/09/30"[dp]) AND english[la]

The OpenAlex search used the same blocks with the title. search and title_and_abstract.search filters, restricted to the document type "article" (which includes conference papers) and to English. The full strings are given in **Appendix A**.

Selection process

Records were exported with their abstracts and merged. Duplicates were removed by DOI, PubMed identifier, and normalized title. Titles and abstracts were then screened against the criteria. Records that passed screening were assessed for eligibility on the full abstract and bibliographic record. Where neither an abstract nor an open full text could be obtained, the record was classed as not retrieved. Screening and eligibility assessment were done by one reviewer (the author). A large language model (Claude, Anthropic) was used as an assistant: it proposed a decision and a reason code for each record, and the author checked these decisions. The decision log is provided as **Supplementary File S2**.

Data collection process and data items

Data were extracted into a structured form (**Supplementary File S1**). The items were: year; condition (depression, anxiety or both); data modality; transfer learning strategy; pre-trained backbone; dataset and population; validation design; and the main reported result. Modality was coded as social media text, clinical or narrative text (interview transcripts, diaries, notes, open responses), speech, video or facial images, EEG, (Magnetic Resonance Imaging) MRI or (Functional Magnetic Resonance Imaging) fMRI, wearable or physiological signals, electronic health records or tabular data, or multimodal. The strategy was coded as feature extraction, fine-tuning, parameter-efficient fine-tuning, domain adaptation, self-supervised pre-training, cross-corpus transfer, prompting of an LLM (zero-shot, few-shot, chain-of-thought or retrieval-augmented), or LLM fine-tuning. A study could have more than one strategy. Validation was coded as random hold-out split, k-fold cross-validation, participant-independent (subject- or user-disjoint) evaluation, external validation on an independent dataset or site, or not reported. Where an abstract gave only qualitative results, the result was recorded as not reported.

Study risk of bias assessment

A full PROBAST assessment [24] needs the full text of each paper, and this was not possible for 318 studies. A simplified appraisal of the PROBAST analysis domain was used instead. Three signaling items were judged for each study: (1) whether the validation was participant-independent or external; (2) whether the abstract reported a quantitative result; and (3) whether the evaluation unit was a segment, post or frame rather than a person. Studies that reported only random hold-out splits of segments or posts were judged to be at high risk of optimistic bias in the analysis domain.

Effect measures and synthesis methods

The studies used different datasets, labels and metrics, so meta-analysis was not appropriate. The results were grouped by modality and by transfer strategy. Counts, proportions and medians were calculated. For studies that reported accuracy, the median accuracy was compared across validation designs. Other metrics (F1 score, area under the curve, mean absolute error and root mean square error) are reported as given by the authors. Representative studies were selected for each modality on the basis of sample size, validation design and direct comparisons between strategies.

Reporting bias and certainty assessment

Reporting bias was examined by looking at the distribution of reported accuracies and at the share of abstracts without numerical results. A formal GRADE assessment was not applied, because GRADE is designed for intervention effects. The certainty of the main findings is described in words in Section 3.8.

Results

Study selection

The database searches returned 798 records: 257 from PubMed and 541 from Open Alex (**Figure 1**). After 212 duplicates were removed, 586 records were screened. Of

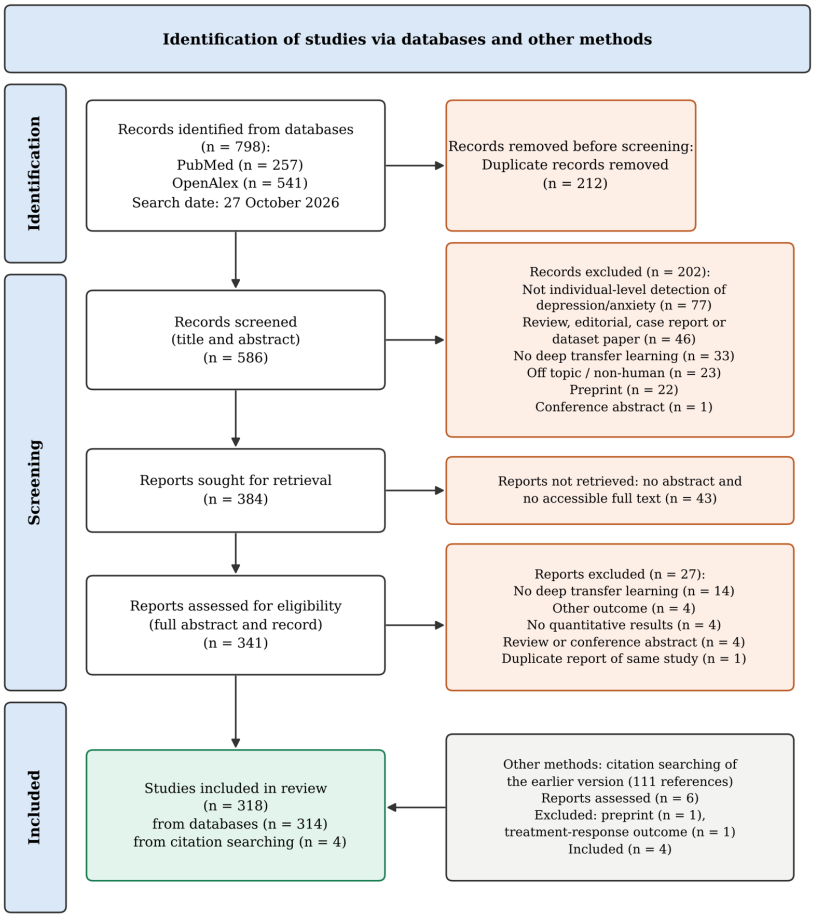


Figure 1. PRISMA 2020 flow diagram of study selection.

these, 202 were excluded: 77 did not address individual-level detection of depression or anxiety, 46 were reviews, editorials, case reports, or dataset papers, 33 did not use a transfer learning component, 23 were outside the topic (for example, respiratory depression or animal studies), 22 were preprints, and one was a conference abstract. Of the 384 reports sought, 43 had no abstract and no accessible full text. The remaining 341 were assessed, and 27 were excluded (14 without a deep transfer learning component, 4 with a different outcome, 4 without quantitative results, 4 that were reviews or conference abstracts, and one duplicate report of the same study). This left 314 studies. Citation searching of the earlier version identified six further reports; two were excluded (one preprint and one treatment-response study), and four were included [25–28]. In total, 318 studies were included.

Study characteristics

Table 2 summarizes the included studies. The field has grown quickly. Three studies were published in 2019, 28 in 2023, 65 in 2024, 106 in 2025, and 87 in the first nine months of 2026. In total, 258 studies (81.1%) were published from 2024 onward (**Figure 2**).

Table 2. Characteristics of the 318 included studies.

Characteristic	Category	n	%
Publication year	2019–2021	14	4.4
Publication year	2022–2023	46	14.5
Publication year	2024	65	20.4
Publication year	2025	106	33.3
Publication year	2026 (January–September)	87	27.4
Condition	Depression only	288	90.6
Condition	Anxiety only	10	3.1
Condition	Depression and anxiety	20	6.3
Data modality	Social media text	111	34.9
Data modality	Multimodal	60	18.9
Data modality	Clinical or narrative text	55	17.3
Data modality	Speech	28	8.8
Data modality	EEG	21	6.6
Data modality	Health records or tabular	15	4.7
Data modality	MRI or fMRI	11	3.5
Data modality	Video or facial images	8	2.5
Data modality	Wearable or physiological	8	2.5
Data modality	Other images (drawings)	1	0.3
Backbone family	Pre-trained text encoder (BERT family)	141	44.3
Backbone family	LLM or other foundation model	87	27.4
Backbone family	Vision CNN or vision transformer	40	12.6
Backbone family	Pre-trained speech or audio model	17	5.3
Backbone family	Other pre-trained or adapted networks	33	10.4
Data source	Used at least one public dataset	99	31.1
Data source	DAIC-WOZ, E-DAIC or AVEC	62	19.5
Data source	MODMA	14	4.4
Data source	Language other than English	45	14.2

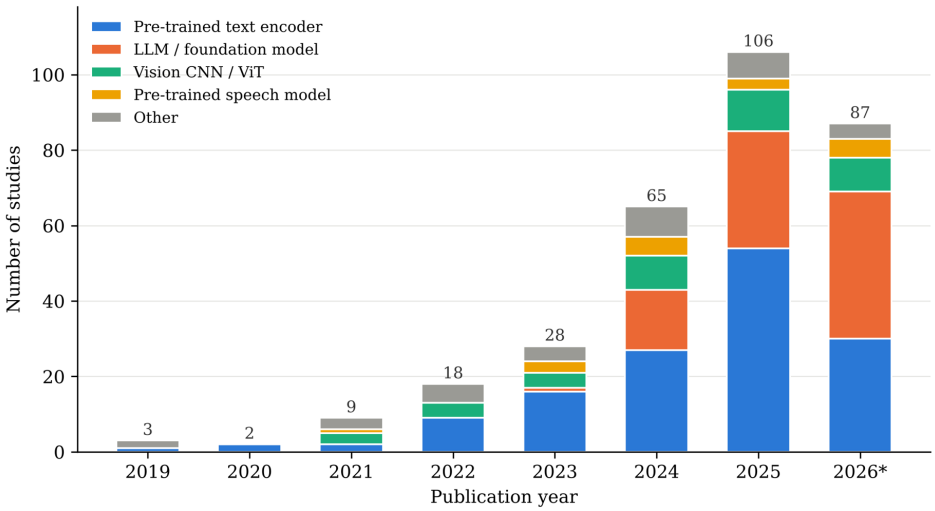


Figure 2. Number of studies included by publication year and pre-trained model family. *2026 covers January to September only.

Depression was the focus of almost all studies. Only 10 studies (3.1%) addressed anxiety alone, and 20 (6.3%) addressed both conditions. Text was the main data source. Social media posts from platforms such as Reddit, X (Twitter) and Weibo were used in 111 studies (34.9%), and clinical or narrative text in 55 (17.3%). Multimodal designs that combined text, audio, and video were used in 60 studies (18.9%). Speech alone was used in 28 studies, EEG in 21, electronic health records or tabular data in 15, MRI or fMRI in 11, video or facial images in 8, and wearable or physiological signals in 8.

Public benchmark datasets were common. At least 99 studies used a public dataset, and 62 used the DAIC-WOZ, E-DAIC or AVEC interview corpora. Fourteen studies used MODMA. At least 45 studies worked with languages other than English, including Chinese, Bengali, Urdu and Roman Urdu, Korean, Thai, Indonesian, Spanish, Turkish, Hebrew, and Arabic. Only two studies on Arabic text were found [29,30].

Fine-tuning was the most common strategy (143 studies), followed by feature extraction (86), prompting of an LLM (59), domain adaptation (18), LLM fine-tuning (18), self-supervised pre-training (10), parameter-efficient fine-tuning (9) and cross-corpus transfer (2). By backbone family, 141 studies used pre-trained transformer encoders such as BERT, RoBERTa or MentalBERT, 87 used LLMs or other foundation models, 40 used ImageNet-type convolutional networks or vision transformers, and 17 used pre-trained speech models such as wav2vec 2.0, HuBERT, WavLM, or Whisper. The use of LLMs rose from one study in 2023 to 16 in 2024, 31 in 2025, and 39 in 2026. **Table 3** gives examples of each strategy.

Table 3. Transfer learning strategies used in the included studies.

Strategy	Studies (n)a)	Typical pre-trained models	Examples
Fine-tuning	143	BERT, RoBERTa, MentalBERT, XLM-RoBERTa, ResNet, VGG, Whisper	[31–34]
Feature extraction (frozen)	86	BERT and SBERT embeddings, wav2vec 2.0, VGGish, ResNet	[35–38]
LLM prompting (zero-, few-shot, chain-of-thought, retrieval)	59	GPT-4/4o, Llama, Qwen, Gemma, Gemini, DeepSeek	[39–42]
LLM fine-tuning	18	Llama, Qwen, ChatGLM, GPT-3.5, DeepSeek-R1	[43–46]
Domain adaptation	18	Adversarial and discrepancy-based networks on EEG, MRI, speech	[47–50]
Self-supervised pre-training	10	Contrastive pre-training on unlabelled EHR, ECG, MRI, speech	[51–54]
Parameter-efficient fine-tuning	9	Prefix tuning, LoRA, adapters, prompt tuning	[55–58]
Cross-corpus transfer	2	Emotion models trained on non-clinical corpora	[59,60]

a) A study can use more than one strategy, so the counts add up to more than 318.

Methodological quality

Table 4 shows the validation designs. A random hold-out split was the most common design (165 studies, 51.9%). Cross-validation was reported in 21 studies (6.6%). Only 16 studies (5.0%) reported participant-independent evaluation and 29 (9.1%) reported external validation. Together, 45 studies (14.2%) met the first signaling item. **Table 5** lists representative studies from this group. In 87 studies (27.4%), the abstract did not describe the validation design.

Table 4. Validation design and reported accuracy.

Validation design	Studies, n (%)	Studies reporting accuracy, n	Median accuracy (%)	Accuracy ≥ 95%, n
Random hold-out split	165 (51.9)	48	90.8	16
k-fold cross-validation	21 (6.6)	11	90.0	3
Participant-independent	16 (5.0)	9	90.8	1
External validation	29 (9.1)	5	88.0	0
Not reported in abstract	87 (27.4)	30	95.0	16
All studies	318 (100)	103	91.3	36

Accuracy values are those reported in the abstracts; studies with segment-level and person-level evaluation are mixed within each group.

Table 5. Representative studies with participant-independent or external validation.

Study	Year	Cond.	Modality	Strategy and backbone	Data	Validation	Main reported result
[65]	2026	D	Speech	Self-supervised pre-training; WavLM, HuBERT	multicentre 1816	External	AUC 0.932, ext 0.879
[33]	2026	D	Speech	Fine-tuning; Whisper	1838 adolescents, 145 external	External	AUROC 0.95, ext 0.88
[43]	2026	D	Multimodal	LLM prompting + LLM fine-tuning; Gemma, Qwen, Llama	in-house 3536, E-DAIC	External	F1 0.865
[40]	2026	D	Clinical/narrative text	LLM prompting; LLM	DAIC-WOZ, E-DAIC	External	MAE 2.91, ext 2.86
[66]	2026	D	Clinical/narrative text	LLM fine-tuning; LLM	in-house 3699 narratives	External	AUC 0.789
[67]	2025	D	Clinical/narrative text	Parameter-efficient fine-tuning; quantised LLM	DAIC-WOZ, PROMPT	External	F1 0.84, ext 0.82
[68]	2025	D+A	Clinical/narrative text	LLM fine-tuning; LLM (EmoScan)	PsyInterview synthetic, external	External	F1 0.747, ext 0.67
[69]	2025	D	Multimodal	Feature extraction; GPT-2 chatbot, pretrained encoders	270 internal, 100 external	External	AUC > 0.95, ext 0.978
[51]	2024	D+A	Health records/ tabular	Self-supervised pre-training; contrastive pretrained encoder	85085 + 17491 T2DM	External	AUC 0.91, ext 0.75
[53]	2026	D	MRI/fMRI	Self-supervised pre-training; 3D CNN pretrained on PGS	ABCD 4741, Korean 108	External	AUC 0.62, ext 0.67
[70]	2026	D	MRI/fMRI	Domain adaptation; GCN pretrained on other sites	REST-meta-MDD 33 sites 3335	External	Acc 70.1%
[34]	2025	D	EEG	Fine-tuning; ResNet18	two public EEG	Participant-independent; External	Acc 88-98%, cross-db 75.1%

Study	Year	Cond.	Modality	Strategy and backbone	Data	Validation	Main reported result
[47]	2026	D	EEG	Domain adaptation; CNN	MODMA	Participant-independent	Acc 87.1%
[71]	2022	D	EEG	Domain adaptation; CNN	wearable 3-channel EEG	Participant-independent	Acc 77.0%
[72]	2026	D	Speech	Feature extraction; wav2vec2, Whisper	DAIC-WOZ	Participant-independent	Acc 0.70, Sens 0.36
[73]	2026	D	Social media text	Fine-tuning; RoBERTa	Reddit	Participant-independent; External	Acc 91.2%
[32]	2026	D	Social media text	Fine-tuning; MentalBERT, BERT, RoBERTa	19800 posts three platforms	External	F1 0.918, AUC 0.962
[59]	2022	D	Speech	Cross-corpus transfer; emotion model trained on non-clinical corpora	Danish 82	External	AUC 0.71
[49]	2025	D	Speech	Domain adaptation; CNN autoencoder	DAIC-WOZ to MODMA	External	Acc 88.7%/81.1%
[52]	2026	A	Wearable/ physiological	Self-supervised pre-training; transformer pretrained on VitalDB	POPANE, stroke patients	External	AUC 0.859
[74]	2026	D+A	Social media text	Fine-tuning; RoBERTa, DistilRoBERTa	Reddit	Participant-independent	Acc 93.4%, AUC 0.98
[75]	2026	D	Social media text	Fine-tuning; BERT	eRisk, CLPsych	Participant-independent	Acc 93.1%, AUC 0.944
[76]	2023	D	Speech	Feature extraction; speaker embedding models	DAIC-WOZ, VocalMind	External	BAcc 66%/64%

Cond.: Condition; D: Depression; A: Anxiety. Results are as reported by the authors

The second item was also weak. In 98 studies (30.8%), the abstract gave no numerical result, only statements such as "outperforms state-of-the-art methods". For the 103 studies that reported accuracy, the median was 91.3%, and 36 of them reported 95% or more. Ten studies reported 99% or higher [29,35,61–63]. Several of these very high values came from small EEG datasets, from evaluation on segments rather than people, or from datasets in which one class was generated or curated by the authors. For example, the normal class in one Arabic tweet study was generated with ChatGPT [29]. One speech study stated clearly that its 96.7% accuracy was a segment-level result and should not be read as participant-level performance [64]. Such statements were rare.

Accuracy fell as validation became stricter. The median was 95.0% in studies that did not report their design, 90.8% for hold-out splits, 90.0% for cross-validation, 90.8% for participant-independent testing and 88.0% for external validation. None of the five externally validated studies with an accuracy value reported 95% or more. The numbers in the stricter groups are small, so the pattern is descriptive.

Results by data modality

Social media text

Social media studies were the largest group. Most fine-tuned BERT-family encoders on labelled posts and reported high accuracy on in-domain test sets. Domain-specific pre-training helped with several comparisons. MentalBERT gave the best results among BERT, RoBERTa, and DistilBERT on 19,800 posts from three platforms (F1 0.918, AUC 0.962), but all models lost performance when tested on a different platform [32]. A large comparison of 14,983 tweets labelled with five depression subtypes found that fine-tuned encoders reached a macro-F1 of 0.94–0.96, while few-shot prompting of instruction-tuned LLMs reached 0.73–0.77 [31]. A DSM-5-guided BERT model on the WU3D Weibo dataset (32,570 users) reached an F1 of 0.918, compared with 0.869 for few-shot GPT-4 [77]. Studies that used user-level splits and an external dataset were few [73,74].

Prediction of future diagnosis was harder than classification of posts. When BERT and MentalBERT were used to identify

Reddit users who later disclosed a depression diagnosis, the mean F1 was 0.64 [78]. In low-resource languages, transfer learning did not always win. A support vector machine outperformed BERT on a Roman Urdu and English dataset [79], a stacking ensemble of character n-grams outperformed AlephBERT on Hebrew health forum posts [80], and an LSTM outperformed a fine-tuned BERT in another comparison [81].

Clinical interviews, narratives and notes

Interview transcripts, diaries, sentence completion tests, and clinical notes were used in 55 studies. These data are closer to clinical practice, and the results were more modest. Two studies tested locally deployed open-source LLMs with fewer than 15 billion parameters on DAIC-WOZ. Accuracy ranged from 0.23–0.68 in one study [39] and reached 0.677 at best in the other [82]. The authors concluded that such models were not accurate enough for screening. In contrast, LLMs used as feature extractors performed better. Prompting an LLM to extract interpretable depression factors, followed by linear regression, gave a PHQ-8 mean absolute error of 2.91 on DAIC-WOZ and 2.86 on E-DAIC [40]. Prefix tuning of a pre-trained language model gave a root mean square error of 4.67 [55].

Fine-tuning improved LLM results in direct comparisons. In 3,536 Korean participants, supervised fine-tuning of five open-source LLMs raised the F1 score to between 0.818 and 0.865, compared with a best zero-shot F1 of 0.735, and the models were tested externally on E-DAIC [43]. In another comparison, fine-tuned GPT-3.5 reached an accuracy of 0.902 on 428 diary entries [44]. On psychotherapy transcripts, fine-tuned BERT, RoBERTa, Longformer and Mistral-7B did not improve on a support vector machine with engineered features [83].

Clinical notes gave mixed results. A GPT-4 prompt detected depression in notes of patients with multiple sclerosis with a sensitivity of 97.3% but a specificity of 68.3% [41]. GPT-4o estimates of PHQ-9 scores from 15,000 outpatient notes were only modestly correlated with the true scores ($r^2 = 0.264$), with a positive predictive value of 0.309 [84]. In 299 participants, four of eight LLMs trained on Hamilton interview transcripts outperformed the self-report HADS-D, and the gap was largest in people with alexithymia [85]. For anxiety, a fine-tuned transformer reached an AUROC of 0.64 for generalized anxiety disorder in 2,000 speech transcripts, compared with 0.58 for a LIWC baseline [86]. GPT-4 ratings of social anxiety from 51 interviews correlated with self-report scores at $r = 0.79$ [87].

Speech

Pre-trained speech models gave some of the best externally validated results. In a multicenter cohort of 1,816 participants, self-supervised WavLM and HuBERT representations reached an AUC of 0.932 internally and 0.879 in external validation [65]. In 1,838 Chinese adolescents, a fine-tuned Whisper model reached an AUROC of 0.95 in development and 0.88 in an external cohort, and performance was similar across age, sex, and severity subgroups [33]. In 865 adults from the general population, Whisper features gave AUCs of 0.76–0.78 for depression and 0.77 for anxiety, with fairness and uncertainty analyzed [88]. Earlier, transfer from a deep language model trained on 16,000 recorded interactions gave AUCs of 0.79–0.86 for depression, anxiety, and their co-occurrence [25], and transfer of encoder weights gave relative gains of up to 27% in PHQ-8 prediction [26].

Cross-corpus and cross-language transfer was possible but weaker. A speech emotion model trained on German and English corpora separated Danish patients with depression from controls with an AUC of 0.71 [59]. Transfer from English to Chinese speech reached 81.1% accuracy [49]. Under participant-independent evaluation on DAIC-WOZ, a wav2vec 2.0 model reached 70.2% accuracy but a sensitivity of only 0.36 [72]. This contrasts with studies on the same corpus that reported accuracies above 96% without a participant-independent design [89].

Multimodal data

Multimodal studies combined pre-trained text encoders (mostly BERT or RoBERTa), speech models and visual networks. Most reported gains over single modalities on DAIC-WOZ, E-DAIC, or AVEC. A model that fused auditory, visual, and textual cues from chatbot interviews reached an AUC above 0.95 internally and 0.978 in an external dataset of 100 participants [69]. Other studies were more cautious. On the DAIC-WOZ test set, a late-fusion model reached an F1 of only 0.50 for the depressed class, and visual features added noise [90]. A controlled study of 512 fusion conditions found that affective pre-training could help one modality and harm another, which shows that negative transfer occurs [91]. The AudiBERT model used audio and text representations for screening and reported higher F1 scores than earlier models [28].

EEG and neuroimaging

EEG studies often converted signals into time-frequency images and applied ImageNet networks such as ResNet, VGG, DenseNet, or EfficientNet. Within-dataset accuracies were high, often above 95% [35,61]. Ten of the 21 EEG studies reported cross-subject (participant-independent) evaluation, a higher rate than in any other modality. Cross-subject results were lower: 87.1% on MODMA with multi-source few-shot adaptation [47] and 77.0% with three-channel wearable EEG [71]. One study reported 88%–98% within each database but 75.1% when a model trained on one database was tested on another [34]. For anxiety, a model reached 95.7% on a public dataset and 86.6% after domain adaptation to seven new participants [92].

MRI studies used pre-training on large cohorts or other disorders, such as Alzheimer's disease. The largest study used 3,335 participants from 33 sites and reached an average accuracy of 70.1% with site-wise transfer [70]. A model pre-trained on polygenic scores and diffusion MRI in 4,741 adolescents reached an AUC of 0.62, and 0.67 in an independent Korean sample [53]. These values are more modest than those of EEG studies, but they come from much larger and more varied samples.

Health records, wearables and other signals

Health record studies used fine-tuned BERT on clinical text or LLMs on tabular data. Contrastive pre-training on 85,085 unlabeled discharge records improved prediction of depression and anxiety in type 2 diabetes, with an AUC of 0.91 internally and 0.75 externally [51]. In 16,159 patients with cancer, BERT models on notes and emails reached an AUROC of about 0.71, below a LASSO model on structured data (0.74), and the BERT risk estimates were poorly calibrated [93]. A fine-tuned 70-billion-parameter LLM on UK Biobank data reached an AUC of 0.892 [46]. Wearable studies were few. A smartwatch model pre-trained on electrocardiogram data tracked post-stroke anxiety with an AUC of 0.859 [52], and a wrist-worn system used domain adaptation to reach 90.8% cross-subject accuracy [50]. A tabular foundation model (TabPFN) did well on a very small student dataset but not on larger ones [94].

Large language models

The 87 studies that used LLMs or other foundation models fell into three groups. In the first group, a general model such as GPT-4, GPT-4o, Llama, Qwen, or Gemini was prompted without training. These studies were quick to run, but results were variable and often lower than those of fine-tuned encoders [31,77]. The performance of small local models was low [39,82]. In the second group, an LLM was fine-tuned, often with LoRA or quantisation, on depression data. These studies reported clear gains over prompting [43,44,67]. Some also trained the model to give DSM-5-based explanations, with reinforcement learning [95]. In the third group, the LLM served as a feature extractor or data generator. It produced symptom scores, summaries or synthetic samples that a smaller model then used [40,96,97]. One study found that clinician-selected LLM outputs gave better classifiers than outputs selected by the LLM itself [97]. Few LLM studies tested the stability of outputs across repeated runs, and agreement among different LLMs was poor in one study [39].

Comparison of transfer strategies

Direct comparisons within the same study allow some conclusions. Fine-tuning a pre-trained speech model usually outperformed a frozen model with a new head. Full fine-tuning of WavLM outperformed a head-only WavLM model and a CNN-RNN model trained from the start (ROC-AUC 0.90) [98]. LoRA fine-tuning of wav2vec 2.0 outperformed WavLM and HuBERT with far fewer trainable parameters [56]. Parameter-efficient methods were less prone to overfitting on DAIC-WOZ [55]. Domain adaptation was mainly used for EEG and MRI, where differences between subjects and sites are large [48,99–101]. Pre-training on emotion data was a common bridge to depression, but the effect depended on how close the source domain was to the target. Negative transfer was reported when the source audio domain did not match [91].

Anxiety and comorbidity

Anxiety was studied far less than depression. The 30 studies that included anxiety used text [57,86,87,102], speech [25,56,88], EEG [92] and wearable data [52]. Comorbidity was addressed in only a few studies. Multi-label classification of depression, anxiety, and substance use disorder in 1,661 Korean counselling sessions gave F1 scores of 0.90 for depression and 0.85 for anxiety [103]. In 130 Turkish women, comorbid cases were assessed more accurately from speech than depression-only cases [56], and the best results in one large speech corpus were seen for comorbid cases [25]. A stacked transformer model separated anxiety from depression posts with an accuracy of 93.4% under a user-wise split [74].

Reporting biases and certainty of evidence

The distribution of results suggests reporting bias. Over a third of studies that reported accuracy gave values of 95% or higher, while studies with stricter designs gave lower values. Almost one third of abstracts did not report any number. Negative findings were reported, but by a small group of studies [39,82–84,93]. On this basis, the certainty that deep transfer learning can detect depression from text and speech in research datasets is moderate. The certainty that the reported accuracies would hold in clinical populations is low. For anxiety, comorbidity, wearable data, and non-English populations, the certainty is very low, because the studies are few and small.

Discussion

Summary of evidence

This review identified 318 studies of deep transfer learning for depression and anxiety. Transfer learning has become the standard approach in this field. Pre-trained transformer encoders dominate text studies, pre-trained speech models dominate speech studies, and LLMs have spread rapidly since 2024. The reported performance is high, but the evidence base has clear weaknesses. Only one study in seven tested models on new participants or new data sources, and these tests gave lower values. The strongest evidence comes from a small number of large speech and health record studies with external validation [33,51,65].

Interpretation

The results agree with earlier reviews that describe high reported accuracy and limited clinical validation [5,6,17]. This review adds three points. First, the strategy of transfer matters. Fine-tuned domain-specific encoders outperformed prompted general LLMs on most text classification tasks [31,77]. LLMs were most useful when they were fine-tuned or used to extract interpretable features. Second, transfer learning does not remove the need for representative data. In low-resource languages and in clinical text, simple models sometimes performed as well as or better than pre-trained ones [79,80,83]. Third, the validation design is the main source of the variation in results. Segment-level or post-level splits allow information from one person to leak into the test set [22]. A model can then learn speaker identity or writing style rather than signs of depression. One study showed that depression signals in speech overlap with the features used to recognize speaker identity [76], which explains why participant-independent results are lower.

The narrow focus on depression is also a problem. Depression and anxiety often occur together and share symptoms. Few studies modelled them jointly, although the available results suggest that comorbid cases can be detected well [25,56]. Social media labels are a further concern. Many datasets use self-disclosure, subreddit membership or keyword-based labels rather than clinical assessment. Studies that used clinical labels, such as Hamilton or PHQ-9 scores and psychiatrist diagnosis, gave a more realistic picture [33,65,85].

Ethics, privacy, and fairness

Privacy was addressed more often in recent studies. Several groups deployed open-source LLMs on local servers so that patient text did not leave the hospital [39,43,82]. One study used federated learning with audio and EEG data [104], and a recent review describes the wider use of federated learning in mental health [105]. Another study removed speaker identity from speech features to protect privacy [106]. Fairness was tested in only a few studies. BERT models for patient messages in cancer care showed larger performance gaps across sex, race, and ethnicity than a Reddit-pre-trained variant [107]. LLM accuracy varied with age and language in a multilingual corpus [108], whereas a speech model for adolescents showed no significant fairness gaps [33]. Social media screening also raises questions of consent, because posts are often collected without the users' knowledge. These issues need ethical review and clear consent procedures, and data protection should follow national rules.

Implications for research and practice

Five practical steps follow from this review. Studies should split data by participant and should report results per person, not per segment or post. External validation on a different site, population or language should be routine before any clinical claim is made. Reporting should follow TRIPOD+AI [109], including sample size, class balance, calibration and subgroup performance, and risk of bias should be assessed with PROBAST [24]. Anxiety and comorbidity need more attention, preferably with clinically labelled data. Finally, more work is needed in under-represented languages. Only two Arabic text studies were found [29,30], and none used clinical data from Arabic-speaking patients. Given the burden reported by the Saudi National Mental Health Survey [4], Arabic speech and text corpora with clinical labels would be a useful contribution, and they would support the health aims of Saudi Vision 2030 and Sustainable Development Goal 3 on good health and well-being.

For clinicians, the current evidence supports the use of these models as research tools and, at most, as aids to triage that are always followed by clinical assessment. They should not replace validated scales or clinical interviews.

Limitations of the evidence

The included studies used different datasets, labels, and metrics, so the results could not be pooled. Many studies used the same few public datasets, especially DAIC-WOZ, so their results are not independent. Sample sizes were small in most EEG, video, and clinical text studies. Labels in social media studies were often weak proxies for clinical diagnosis. Many abstracts did not report the validation design or any number.

Limitations of the review process

This review has important limitations. Only two databases were searched. Open Alex covers much of the engineering literature, but Scopus, Web of Science, and IEEE Xplore were not searched directly, so some studies may have been missed. Eligibility assessment and data extraction were based on the full abstract and bibliographic record, not on the full text of each paper, because of the size of the evidence base. Some details, such as sample size and the exact split, could therefore not be checked. Forty-three records could not be assessed because no abstract or full text was available. Screening was done by one reviewer with the help of a large language model, and a second independent reviewer was not available. Only English-language publications were included, and preprints were excluded. Finally, the field is moving fast, and many relevant studies will have appeared after the search date.

Conclusions

Deep transfer learning is now the main approach to automatic detection of depression and, to a smaller extent, anxiety. Pre-trained language and speech models give the best results, and fine-tuning generally works better than prompting a general

LLM. However, most of the 318 included studies used random splits of small or public datasets, and only 14.2% tested models on new participants or new data sources. The high accuracies reported in many abstracts should therefore be read with caution. The next step for the field is fewer new architectures and more careful evaluation: participant-independent splits, external and multi-site validation, clinically labelled data, joint modelling of depression and anxiety, fairness testing, and data from under-represented languages such as Arabic.

Declarations

Funding

This research received no external funding.

Conflicts of interest

The author declares no conflict of interest.

Data availability

The search strings, the screening decision log and the data extraction sheet are provided as supplementary files.

Author contributions

T.S.M. designed the review, carried out the searches, screened the records, extracted and synthesized the data, and wrote and revised the manuscript.

Use of generative AI

A large language model (Claude, Anthropic) was used to assist with screening, data extraction, and language editing. The author checked all decisions and the final text and takes full responsibility for the content.

Acknowledgements

The author thanks the two anonymous reviewers, whose comments led to this revised version.

Supplementary materials

Supplementary File S1: data extraction sheet for the 318 included studies, with full bibliographic details. Supplementary File S2: record-level screening and eligibility decision log for all 586 screened records and the six reports from citation searching.

References

1. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry. 2022 Feb;9(2):137–50.
2. COVID-19 Mental Disorders Collaborators. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet. 2021 Nov 6;398(10312):1700–12.
3. World Health Organization. Depressive disorder (depression): fact sheet [Internet]. Geneva: WHO; 2026 Sep 11 [cited 2026 Oct 27]. Available from: https://www.who.int/news-room/fact-sheets/detail/depression
4. Altwaijri YA, Al-Subaie AS, Al-Habeeb A, Bilal L, Al-Desouki M, Aradati M, et al. Lifetime prevalence and age-of-onset distributions of mental disorders in the Saudi National Mental Health Survey. International Journal of Methods in Psychiatric Research. 2020 Sep;29(3):e1836.
5. Squires M, Tao X, Elangovan S, Gururajan R, Zhou X, Acharya UR, et al. Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment. Brain Informatics. 2023 Apr 24;10(1):10.
6. Iyortsuun NK, Kim SH, Jhon M, Yang HJ, Pant S. A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis. Healthcare (Basel). 2023 Jan 17;11(3):285.
7. Safayari A, Bolhasani H. Depression diagnosis by deep learning using EEG signals: A systematic review. Medicine in Novel Technology and Devices. 2021 Dec 1;12:100102.
8. Khosla A, Khandnor P, Chand T. Automated diagnosis of depression from EEG signals using traditional and deep learning approaches: A comparative analysis. Biocybernetics and Biomedical Engineering. 2022 Jan 1;42(1):108–42.
9. Gratch J, Artstein R, Lucas GM, Stratou G, Scherer S, Nazarian A, et al. The distress analysis interview corpus of human and computer interviews. InLrec. 2014 May 26;14:3123–8.
10. Cai H, Yuan Z, Gao Y, Sun S, Li N, Tian F, et al. A multi-modal open dataset for mental-disorder analysis. Scientific Data. 2022 Apr 19;9(1):178.
11. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H, He Q. A comprehensive survey on transfer learning. Proceedings of the IEEE. 2020 Jul 7;109(1):43–76.
12. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. Journal of Big data. 2016 May 28;3(1):9.
13. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT 2019. Minneapolis: ACL; 2019. p. 4171–86.
14. Baevski A, Zhou Y, Mohamed A, Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems. 2020;33:12449–60.
15. Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E. Mentalbert: Publicly available pretrained language models for mental healthcare. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022). Marseille: ELRA; 2022. p. 7184-90.

16. Liu D, Feng XL, Ahmed F, Shahid M, Guo J. Detecting and measuring depression on social media using a machine learning approach: systematic review. JMIR Mental Health. 2022 Mar 1;9(3):e27244.

17. Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language models for mental health applications: systematic review. JMIR Mental Health. 2024 Oct 18;11(1):e57400.

18. Omar M, Levkovich I. Exploring the efficacy and potential of large language models for depression: A systematic review. Journal of Affective Disorders. 2025 Feb 15;371:234–44.

19. Mao H, Han Q. Applications of transformer-based language models for depression detection: A scoping review. Journal of Integrated Social Sciences and Humanities. 2025 Jun 20:1–8.

20. Dai A, Shi W, Gu X, Huang L. Large Language Models for Depression Detection: A Review with Prospects of Incomplete Multimodality. Brain Sciences. 2026 May 30;16(6):593.

21. Davoudi Z, Roostaee M, Aghajani K. From ensembles to knowledge-augmented reasoning: Information fusion for depression detection in the generative AI era. Informatics in Medicine Unlocked. 2026 Jul 1;64:101781.

22. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023 Sep 8;4(9):100804.

23. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021 Mar 29;372:n71.

24. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S; PROBAST Group†. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Annals of Internal Medicine. 2019 Jan 1;170(1):51–8.

25. Rutowski T, Shriberg E, Harati A, Lu Y, Chlebek P, Oliveira R. Depression and anxiety prediction using deep language models and transfer learning. In: 2020 7th International Conference on Behavioural and Social Computing (BESC). IEEE; 2020. p. 1-6.

26. Harati A, Shriberg E, Rutowski T, Chlebek P, Lu Y, Oliveira R. Speech-based depression prediction using encoder-weight-only transfer learning and a large corpus. In: ICASSP 2021 - IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE; 2021. p. 7273–7.

27. Lorenc P, Uban AS, Rosso P, Šedivý J. Detecting early signs of depression in the conversational domain: The role of transfer learning in low-resource scenarios. In: Natural Language Processing and Information Systems (NLDB 2022). Cham: Springer; 2022. p. 358–69.

28. Toto E, Tlachac ML, Rundensteiner EA. Audibert: A deep transfer learning multimodal classification framework for depression screening. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM 2021). New York: ACM; 2021. p. 4145–54.

29. Elmajali S, Ahmad I. Toward early detection of depression: Detecting depression symptoms in Arabic tweets using pretrained transformers. IEEE Access. 2024 Jun 24;12:88134–45.

30. Rabie EM, Hashem AF, Alsheref FK. Recognition model for major depressive disorder in Arabic user-generated content. Beni-Suef University Journal of Basic and Applied Sciences. 2025 Jan 24;14(1):7.

31. AlSaad R, Alshakhs S, Thomas R. Depression subtype classification from social media posts: few-shot prompting vs. fine-tuning of large language models. Frontiers in Digital Health. 2026 Mar 23;8:1790533.

32. Abdrazakh Y, Ismailova R, Zhunissov N, Aben A, Amanov A, Baimakhanova A. Comparative Analysis of Neural Network Architectures for Classifying Depressive Content in Social Networks. International Journal of Advanced Computer Science and Applications. 2026;17(2).

33. Yan J, Liyanage BN, Zhang K, Kang S, Zhu Z, Cui Z, et al. An objective smartphone speech diagnostic aid for adolescent major depressive disorder. NPJ Digital Medicine. 2026 Jul 29.

34. Bagherzadeh S, Norouzi MR, Ghasri A, Tolou Kouroshi P, Bahri Hampa S, Farokhshad F, et al. Automated depression detection via cloud based EEG analysis with transfer learning and synchrosqueezed wavelet transform. Scientific Reports. 2025 May 23;15(1):18008.

35. Sun C, Guan M, Duan K, Gao S, Chen Z. A depression detection approach leveraging transfer learning with single-channel EEG. Journal of Neural Engineering. 2025 Jun 1;22(3):036001.

36. Pan R, Sidu F, Sun Y, Xu J, Zhai T, Wu X, et al. CFRAFN: A Cross-Feature Residual Attention Fusion Network for Major Depressive Disorder Prediction Using Clinical Voice Recordings. IEEE ournal of Biomedical and Health Informatics. 2026 Oct;30(10):8645–58.

37. Wang N, Zhang W, Kamil R, Renner I, Abdul Rahman Al-Haddad S, et al. Depression detection through dual-stream modeling with large language models: a fusion-based transfer learning framework integrating BERT and T5 representations. Frontiers in big Data. 2025;8:1651290.

38. Shahid SI, Tayarani Najaran MH, Förster F, Steuber V. Screening anxiety via contrastive autobiographical recall. Frontiers in Digital Health. 2026 May 25;8:1798100.

39. Ling SH, Chorney W. The use of large language models in automated depression detection. Acta Psychologica. 2026 Sep 1;269:107602.

40. Lee JJ, Han J, Woo CW. Interpretable depression assessment using a large language model. PLOS Digital Health. 2026 Feb 9;5(2):e0001205.

41. Julian-Kwong M, Poole S, Henderson K, Wijangco J, Sisodia N, Gelfand JM, et al. Large language models to infer depression in patients with neurological conditions. BMJ Neurology Open. 2026 May 13;8(1):e001502.

42. Xie W, Song X, Lu Z, Fei G, Zhou Y, Ding S, et al. Retrieval-augmented large language models for depression screening and suicide risk stratification. BMC Psychiatry. 2026 Dec;26(1):386.

43. Kwon S, Kim Y, Jhon M, Park JH, Lim B, Jeon E, et al. On-premises open-source large language models for privacy-preserving multimodal depression screening. International Journal of Medical Informatics. 2026 Nov 1;220:106577.

44. Shin D, Kim H, Lee S, Cho Y, Jung W. Using large language models to detect depression from user-generated diary text data as a novel approach in digital mental health screening: instrument validation study. Journal of Medical Internet Research. 2024 Sep 18;26:e54617.

45. Liu J, Ding P, Chen J. DepGLM: Depression degree recognition on social media based on large language models. Digital Health. 2025 Dec;11:20552076251408281.

46. Sha Y, Pan H, Xu W, Meng W, Luo G, Du X, et al. MDD-LLM: Towards accuracy large language models for major depressive disorder diagnosis. Journal of Affective Disorders. 2025 Nov 1;388:119774.

47. Wang K, Zhang Y, Zhang Y, Zhang F, Shen J, Hu B. MSFSNet: Multi-Source Few-Shot Adaptation Network for Cross-Subject Depression Recognition from EEG Signals. IEEE Journal of Biomedical and Health Informatics. 2026 May 7.

48. Gao Y, Wang Q, Sun Y, Wang C, Liang Y, Liu M. Learning from heterogeneous structural MRI via collaborative domain adaptation for late-Life depression assessment. Neural Networks. 2026 Mar 28:108911.

49. Qin S, Zhang Y, Ma Y, Li H, Li X, Lian B, et al. A cross-linguistic depression detection method based on speech data. Journal of Affective Disorders. 2025 Dec 1;390:119739.

50. Zhang Y, Jin S, Kuang W, Zheng Y, Song Q, Fan C, et al. Depguard: Depression recognition and episode monitoring system with a ubiquitous wrist-worn device. IEEE Transactions on Mobile Computing. 2025 Jul 21;25(1):197–214.

51. Feng W, Wu H, Ma H, Tao Z, Xu M, Zhang X, et al. Applying contrastive pre-training for depression and anxiety risk prediction in type 2 diabetes patients based on heterogeneous electronic health records: a primary healthcare case study. Journal of the American Medical Informatics Association. 2024 Feb 1;31(2):445–55.

52. Choi S, Seo WK, Jung JM, Park S, Seo HC. Smartwatch-Based Unobtrusive Continuous Anxiety Tracker for Evaluating Post-Stroke Patients' Quality of Life. IEEE Journal of Biomedical and Health Informatics. 2026 Jul;30(7):6292–305.

53. Min JJ, Wang H, Lee E, Kim BG, Kim G, Choi SY, et al. Polygenic risk-informed white matter integrity improves deep learning-based prediction of youth depression. Communications Medicine. 2026 Aug 4.

54. Han Z, Shang Y, Shao Z, Liu J, Guo G, Liu T, et al. Spatial–temporal feature network for speech-based depression recognition. IEEE Transactions on Cognitive and Developmental Systems. 2023 May 8;16(1):308–18.

55. Lau C, Zhu X, Chan WY. Automatic depression severity assessment with deep learning using parameter-efficient tuning. Frontiers in Psychiatry. 2023 Jun 15;14:1160291.

56. Beşirli A, Karakadıoğlu T, Demiroğlu C. Estimating depression and anxiety scores from conversational speech in females with and without comorbidity. EURASIP Journal on Audio, Speech, and Music Processing. 2025 Dec 13;2026(1):6.

57. Senanayake A, Yapa P, Liyanage SR. SOFIA: a knowledge-infused prompt-based reasoning framework for anxiety screening in social media. Frontiers in Digital Health. 2026 Aug 14;8:1899582.

58. Gerczuk M, Triantafyllopoulos A, Amiriparian S, Kathan A, Bauer J, Berking M, et al. Zero-shot personalization of speech foundation models for depressed mood monitoring. Patterns. 2023 Nov 10;4(11).

59. Hansen L, Zhang YP, Wolf D, Sechidis K, Ladegaard N, Fusaroli R. A generalizable speech emotion recognition model reveals depression and remission. Acta Psychiatrica Scandinavica. 2022 Feb;145(2):186–99.

60. Prabhu S, Mittal H, Varagani R, Jha S, Singh S. Harnessing emotions for depression detection. Pattern Analysis and Applications. 2022 Aug;25(3):537–47.

61. Sharma G, Joshi AM, Gupta R, Cenkeramaddi LR. DepCap: a smart healthcare framework for EEG based depression detection using time-frequency response and deep neural network. IEEE Access. 2023 May 10;11:52327–38.

62. Singh A, Tiwari V, Patel H, Vivekananda GN, Rajput DS. Slitranet: an EEG-based automated diagnosis framework for major depressive disorder monitoring using a novel LGCN and transformer-based hybrid deep learning approach. IEEE Access. 2024 Nov 7;12:173109–26.

63. Mohmand R, Habib U, Usman M, Baili J, Nam Y. A deep learning approach for automated depression assessment using roman urdu. IEEE Access. 2024 Dec 17;12:193387–401.

64. Wang N, Li J, Zhang W, Kamil R, Renner I, Al-Haddad SA, et al. Dual-modality modeling for depression detection using speech signals. Scientific Reports. 2026 Sep 2.

65. Lin Y, Liyanage BN, Shi C, Zhu Z, Sun Y, Sun Y, Zhang ZS, Yang J, Xu C, Liao Y, Tao Y, Li Z, Zhao G, Yue W. Speech as a biomarker for supported diagnosis of major depressive disorder using self-supervised representations. Nature Communications. 2026 Jun 15;17(1):7886.

66. Moon S, Lee A, Kim JE, Kang HJ, Shin IS, Kim SW, et al. Depressllm: interpretable domain-adapted language model for depression detection from real-world narratives. IEEE Journal of Biomedical and Health Informatics. 2026 Jul 9.

67. Ikeuchi K, Kishimoto T, Nakai F, Horigome T, Kitazawa M, Ohtsuki T. Efficient and Effective Fine-tuning Method for Depression Detection from Conversation. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2025 Jul;2025:1–5.

68. Liu JM, Gao M, Sabour S, Chen Z, Huang M, Lee TM. Enhanced large language models for effective screening of depression and anxiety. Communications Medicine. 2025 Nov 5;5(1):457.

69. Xu C, Chen Y, Tao Y, Xie W, Liu X, Lin Y, et al. Deep learning-based detection of depression by fusing auditory, visual and textual clues. Journal of Affective Disorders. 2025 Dec 15;391:119860.

70. Su J, Qin J, Shen H, Zeng LL, Li B, Liu J, et al. Multi-Site Transfer Classification of Major Depressive Disorder: An fMRI Study in 3335 Subjects. Advanced Science. 2026 Apr;13(19):e02817.

71. Wu W, Ma L, Lian B, Cai W, Zhao X. Few-electrode EEG from the wearable devices using domain adaptation for depression detection. Biosensors. 2022 Nov 28;12(12):1087.

72. Chen Z, Wu M. Speech Depression Screening via Multi-Scale Feature Enhancement and Emotion-Aware Contrastive Learning. Bioengineering. 2026 Jun 25;13(7):739.

73. Karthiga M, Joseph ER, Patil S, Saranya K. Neuro-semantic graph fusion for explainable depression risk trajectory mapping using a graph-enhanced RoBERTa framework. Frontiers in Artificial Intelligence. 2026 Aug 5;9:1854179.

74. Hossain M, Khaleque T, Islam MJ, Das SR, Al Noman A, Morol MK, et al. ProStack-MH: An Explainable Stacked Transformer Approach to Anxiety–Depression Differentiation in Real-World Mental-Health Discussions. IEEE Access. 2026 Jun 23;14:112865–90.

75. Rehman AU, Iqbal A, Sayyed A, Aslam Z, Mohmand MI, Husnain G. Explainable Text-Based Depression and Suicide Risk Prediction from Social Media Using Deep Learning and Graph Neural Networks. Healthcare (Basel). 2026 May 22;14(11):1440.

76. Dumpala SH, Dikaios K, Rodriguez S, Langley R, Rempel S, Uher R, et al. Manifestation of depression in speech overlaps with characteristics used to represent and recognize speaker identity. Scientific Reports. 2023 Jul 10;13(1):11155.

77. Jin H, Zhang L. Depression Detection on Social Media Using Multi-Task Learning with BERT and Hierarchical Attention: A DSM-5-Guided Approach. Electronics. 2026 Jan 29;15(3):598.

78. Owen D, Antypas D, Hassoulas A, Pardiñas AF, Espinosa-Anke L, Collados JC. Enabling early health care intervention by detecting depression in users of web-based forums using language models: longitudinal analysis and evaluation. JMIR AI. 2023 Mar 24;2(1):e41205.

79. Rehmani F, Shaheen Q, Anwar M, Faheem M, Bhatti SS. Depression detection with machine learning of structural and non-structural dual languages. Healthcare Technology Letters. 2024 Aug;11(4):218–26.

80. Keinan R, Margalit E, Bouhnik D. Depression Detection With Machine Learning in A Low-Resource Language: A Case Study in Hebrew Digital Health Communities. Array. 2026 Aug 22:101157.

81. Raj A, Sharma A. Revealing Hidden Pain: A Comparative Analysis of Traditional vs New Deep Learning Approaches for Detecting Depression on Social Media. IEEE Access. 2026 Jan 28; 14:17942–59.

82. Ling SH, Chorney W. Large Language Model in Automatic Depression Detection. European Psychiatry. 2026 Jul 31;69(Suppl 1):S619.

83. Sun J, Ma S, Fan Y, Washington P. LLMs Underperform on Classifying Anxiety and Depression Using Therapy Conversations: A First-Step Benchmark. Applied Sciences. 2026 Mar 31;16(7):3388.

84. McCoy TH, Castro VM, Perlis RH. Estimating depression severity in narrative clinical notes using large language models. Journal of Affective Disorders. 2025 Jul 15;381:270–4.

85. Lam C, Xian L, Huang R, Chen J, Chan KY, Chan JW, et al. Deep learning for detecting depression in individuals with and without alexithymia. Communications Medicine. 2026 Jan 16;6(1):123.

86. Teferra BG, Rose J. Predicting generalized anxiety disorder from impromptu speech transcripts using context-aware transformer-based neural networks: model evaluation study. JMIR mental health. 2023 Mar 28;10(1):e44325.

87. Ohse J, Hadžić B, Mohammed P, Peperkorn N, Fox J, Krutzki J, et al. GPT-4 shows potential for identifying social anxiety from clinical interview data. Scientific Reports. 2024 Dec 16;14(1):30498.

88. Riad R, Denais M, de Gennes M, Lesage A, Oustric V, Cao XN, et al. Automated speech analysis for risk detection of depression, anxiety, insomnia, and fatigue: algorithm development and validation study. Journal of Medical Internet Research. 2024 Oct 31;26:e58572.

89. Huang X, Wang F, Gao Y, Liao Y, Zhang W, Zhang L, et al. Depression recognition using voice-based pre-training model. Scientific Reports. 2024 Jun 3;14(1):12734.

90. Betrand C, Orish C, Ekwealor O, Onukwugha C, Benson-Emenike M, Oragba N, et al. Development of Predictive Analytics Model for Early Detection of Depression in Women. F1000Research. 2026 Jun 24;15:1002.

91. Matsumoto K, Kiuchi K, Umehara H, Nakataki M, Numata S. Interaction-Driven Dynamic Fusion for Multimodal Depression Detection: A Controlled Analysis of Gating and Cross-Attention Under Class Imbalance. Brain Sciences. 2026 Mar 28;16(4):366.

92. Aldayel M, Al-Nafjan A. Automated Anxiety Detection System Integrating a Brain–Computer Interface for Neurofeedback Applications. Sensors. 2026 Jun 24;26(13):4004.

93. de Hond A, van Buchem M, Fanconi C, Roy M, Blayney D, Kant I, et al. Predicting depression risk in patients with cancer using multimodal data: algorithm development study. JMIR Medical Informatics. 2024 Jan 18;12(1):e51925.

94. Arini FY, Syah MK, Setiawan FD, Indarto RM, Aminuddin MD, Hilmy FT, et al. Performance Evaluation of TabPFN for Student Depression Prediction Across Varying Sample Sizes. Jurnal Ilmu Komputer dan Informasi. 2026 Jul 22;19(2):169–77.

95. Bao E, Perez A, Parapar J. Improving clinical reliability of LLM reasoning for depression assessment via structured generation and GRPO. Science Progress. 2026 Jul;109(3):00368504261467483.

96. García-Ordás MT, Rubio-Martín S, Merayo A, Crespo-Álvaro A, García-Olalla O, García-Rodríguez I. Explainable depression prediction using SHAP and large language models to extract features from text transcriptions. Applied Soft Computing. 2026 Jul 31:116129.

97. Kim JW, Yoon H, Jung D, Oh W, Choi D, Yoon SH, Kim DJ, Lee SY, Kim BN, Yang CM. DSM-Guided Large Language Model Reasoning for Depression and PTSD Assessment From Psychiatric Interview Transcripts. IEEE Access. 2026 Aug 27;14:139882–901.

98. Kim AY, Jeon M, Cho CH, Shin MS, Byun S. Screening for depression risk via smartphone narratives with fully fine-tuned WavLM. Science Reports. 2026 May 19;16(1):22741.

99. Huang H, Li X, Wu M, Ouyang R, Li J, Lv Z. D 2AC-Dep: dynamic adaptive feature fusion and domain adaptation collaboration for cross-subject depression detection. Cognitive Neurodynamics. 2026 Jun 2;20(1):104.

100. Gao Y, Wu M, Wang L, Wang L, Stephens DC, Potter GG, et al. Improving Late-Life Depression Analysis with Collaborative Domain Adaptation: Learning from Heterogeneous Structural MRI. Reconstr Imaging Motion Estim Graph Biomed Image Anal (2025). 2026;16150:151–61.

101. Xu CY, Fan FY, Zhao LX, Jin LC, Zhang YH, Meng QH. Major depressive disorder detection via temporal-frequency-spatial transformer with sub-domain knowledge alignment using EEG. Neural Networks. 2026 Jan 1;193:107965.

102. Zhu J, Zhang Z, Li Z, Hu B. Integrating clinical anxiety scales with pre-trained language models for anxiety recognition on social media. Health Information Science and Systems. 2025 Sep 6;13(1):56.

103. Choi M, Jo J, Jeong J. An Artificial Intelligence-Based Detection of Comorbid Depression, Anxiety, and Substance Use Disorder in Korean Counseling Dialogues Using an Explainable Hierarchical Attention Network with Shapley Additive Explanations. Diagnostics. 2026 Jun 12;16(12):1817.

104. Gupta C, Khullar V, Goyal N, Saini K, Baniwal R, Kumar S, et al. Cross-silo, privacy-preserving, and lightweight federated multimodal system for the identification of major depressive disorder using audio and electroencephalogram. Diagnostics. 2023 Dec 25;14(1):43.

105. Khalil SS, Tawfik NS, Spruit M. Exploring the potential of federated learning in mental health research: a systematic literature review: SS Khalil et al. Applied Intelligence. 2024 Jan;54(2):1619–36.

106. Ravi V, Wang J, Flint J, Alwan A. A Privacy-Preserving Unsupervised Speaker Disentanglement Method for Depression Detection from Speech. CEUR Workshop proceedings. 2024 Feb;3649:57–63.

107. van Buchem MM, de Hond AA, Fanconi C, Shah V, Schuessler M, Kant IM, et al. Applying natural language processing to patient messages to identify depression concerns in cancer patients. Journal of the American Medical Informatics Association. 2024 Oct;31(10):2255–62.

108. Perez-Toro PA, Dineley J, Iniesta R, Zhang Y, Matcham F, Siddi S et al. Exploring biases related to the use of large language models in a multilingual depression corpus: PA Perez-Toro et al. Scientific reports. 2025 Oct 16;15(1):36197.

109. Collins GS, Moons KG, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, Van Smeden M, Boulesteix AL. TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024 Apr 16;385:e078378.